Интеллектуальный фонд «Социотехника»
Институт перспективных технологий

НЕЙРОМОРФНЫЕ ФОТОННЫЕ СИСТЕМЫ

Ростов-на-Дону
2022

УДК 524.8
ББК 22.68
М55

Нейроморфные фотонные структуры : сборник научных работ/Ред.-составитель Г.С.Мельников. – Ростов-на-Дону, 2022 − 280 с.

Сборник включает в себя переводные научные работы по одному из новых и перспективных направлений − по нейроморфной фотонике. На это новое направление возлагаются надежды на дальнейшее продвижение, в частности, в разработке искусственного интеллекта.

Работы, представленные в сборнике, предназначены, прежде всего, специалистам, а также преподавателям ВУЗов, студентам, магистрантам и аспирантов соответствующих специальностей.

Все переводные статьи взяты редактором из свободных публикаций в интернете. Редактор сборника получил разрешение на перепечатку материалов от авторов в т.ч. через сайт ResearchGate.

УДК 524.8
ББК 22.68

Содержание

ПРЕДИСЛОВИЕ

Сборник научных работ «Нейроморфные фотонные структуры» включает в себя переводные статьи научные по одному из новых и перспективных направлений – по нейроморфной фотонике.

Необходимо отметить, что работы, включенные в сборник, дают представление о новом научном направлении и пополнят библиотеку научных знаний на русском языке, станут более доступны для русскоязычного читателя.

Под нейроморфными системами понимаются модели искусственных нейронных сетей, архитектура и дизайн которых основаны на особенностях структуры и принципах работы реальных нейробиологических систем. Успехи в области создания нейроморфных фотонных систем с глубоким обучением за последние три года поразительные. В последнее время появился ряд предложений по реализации функции нелинейной активации (NLAF) для определения порога, включая использование модуляторов электропоглощения и модуляторов микрокольца, а также ряда полностью оптических реализаций. Глубокое обучение за эти годы проникло в новые сферы: в голографию, квантовые вычисления, а также глубокого обучения для создания фантомных изображений, в создание дифракционной визуализация с субволновым разрешением, в адаптивное управление свойствами метаматериалов, в частности, в создание самонастраивающейся микроволновой маскировки и в создании адаптивных акустических систем. Уже были сделаны первые шаги по макетированию элементной базы и программируемого на кристалле нелинейно-оптического процессора, намечены пути дальнейшего развития.

На все работы настоящего сборника получены разрешения Геннадием Семеновичем Мельниковым от самих авторов, либо с разрешённым доступом – через ResearhGate. Переводчиком всех статей сборника является Мельников Г.С. и Google Translate с Яндекс Переводчиком (https://translate.yandex.ru/?). Геннадий Семенович считает, что это надо особо подчеркнуть, как наглядный пример применения нейроморфных структур в действии

НАСТОЯЩЕЕ И БУДУЩЕЕ НЕЙРОМОРФНЫХ, ФОТОННЫХ СИСТЕМ С ГЛУБОКИМ ОБУЧЕНИЕМ

Г.С. Мельников, Э.И. Мельникова,
В.М. Самков, В.А. Бондаренко, В.С. Чураков

Аннотация

Под нейроморфными системами понимаются модели искусственных нейронных сетей, архитектура и дизайн которых основаны на особенностях структуры и принципах работы реальных нейробиологических систем. С момента написания нашего обзора [11] прошло менее 3х лет, но успехи в области создания нейроморфных фотонных систем с глубоким обучением за эти годы поразительные. В последнее время появился ряд предложений по реализации функции нелинейной активации (NLAF) для определения порога, включая использование модуляторов электропоглощения [7] и модуляторов микрокольца [8], а также ряда полностью оптических реализаций [12–14]. Особое место в обзоре уделено обучению искусственных нейронных сетей [2-6, 15-18]. Глубокое обучение за эти годы проникло в новые сферы: в голографию [19], квантовые вычисления [20-23], глубокого обучения для создания фантомных изображений. [24], в создание дифракционной визуализация с субволновым разрешением [25], в адаптивное управление свойствами метаматериалов [26], в частности, в создание самонастраивающейся микроволновой маскировки [27] и в создании адаптивных акустических систем [28-29]. Ужу были сделаны первые шаги по макетированию элементной базы и программируемого на кристалле нелинейно-оптического процессора [30-32], и намечены пути развития [33-34].

Annotation

Neuromorphic systems are understood as models of artificial neural networks, the architecture and design of which are based on the structural features and principles of operation of real neurobiological systems. Less than 3 years have passed since the writing of our review [11], but the successes in the field of creating neuromorphic photonic systems with deep learning over the years are amazing. Recently, a number of proposals have appeared on the implementation of the nonlinear activation function (NLAF) for determining the threshold, including the use of electroabsorption modulators [7] and microring modulators [8], as well as a number of all-optical implementations [12-14]. A special place in the review is given to the training of artificial neural networks [2-5, 15-18]. Deep learning has penetrated new areas over the years: holography [19], quantum computing [20-23], deep learning for creating phantom images. [24], in the creation of diffraction im-

aging with subwavelength resolution [25], in the adaptive control of the properties of metamaterials [26], in particular, in the creation of self-adjusting microwave cloaking [27] and in the creation of adaptive acoustic systems [28-29]. The first steps have already been taken to prototype the element base and a nonlinear optical processor programmable on a chip [30-32], and the development paths have been outlined [33-34].

Введение

Нейроморфный чип это — процессор, работа которого основана на принципах действия человеческого мозга. Такие устройства моделируют работу нейронов и их отростков — аксонов и дендритов — отвечающих за передачу и восприятие данных. Связи между нейронами образуются за счет синапсов — специальных контактов, по которым транслируются электрические сигналы

Обучаемая нейронная сеть является вероятностным представлением импульсной нейронной сети. Синаптические связи представляются как вероятности ij ~ c, такие, что P(cij=1) = ij ~ c, синаптический вес представляется как sij так же, как и в сети нейро-чипа. Предполагается, что переменные sij определены на ограниченном диапазоне значений и на них действует дополнительное ограничение, которое заключается в том, что существуют «блоки», в которых синапсы имеют одинаковые значения весов,

После обучения нейронная сеть с полученными в процессе обучения значениями вероятностей синаптических связей в ядре xbgf преобразуется в импульсную сеть для размещения в чипе.

При этом для каждого распознаваемого класса создается ансамбль из нескольких нейронных сетей, в которых случайно формируется матрица синаптических связей cij = (0, 1) в ядре так, чтобы значения 0 и 1 в этой матрице по всем сетям из ансамбля соответствовали полученным в процессеобучения вероятностям связей. Эта матрица задает наличие связей между -м синапсом и j-м нейроном. Суммарный сигнал на j-м нейроне вычисляется по формуле:

$$Ij = \sum j (Xi\ Cij\ Sij + bj) \quad (1)$$

где xi – входной сигнал на i-м синапсе, sij – вес (сила) i-го синапса применительно к j-му нейрону, bj – смещение j-го нейрона. Веса sij задаются в процессе решения оптимизационной задачи минимизации избыточности нейронов. [6]. Наиболее простое решение по созданию нейроморфной сети решающей эти задачи предложено в патенте РФ.[53] Этот патент позволил строить нейроморфные системы распознавания изображений на отечественной элементной базе с очень высокой вероятностью принятия решения.

Для разрешения указанной задачи предлагается способ нейроморфной обработки данных, включающий операции нейросетевой обработки, в котором при подаче питания в программируемую логическую интегральную схему (ПЛИС), содержащую блок нейронной сети, реализующей выбор нейронной сети для решения текущей задачи, удовлетворяющей критериям выбора, включающим в себя наибольшую ожидаемую точность решения текущей задачи и наименьшие вычислительные ресурсы, затрачиваемые на решение этой задачи, из конфигурационной памяти загружается конфигурация указанного блока нейронной сети.

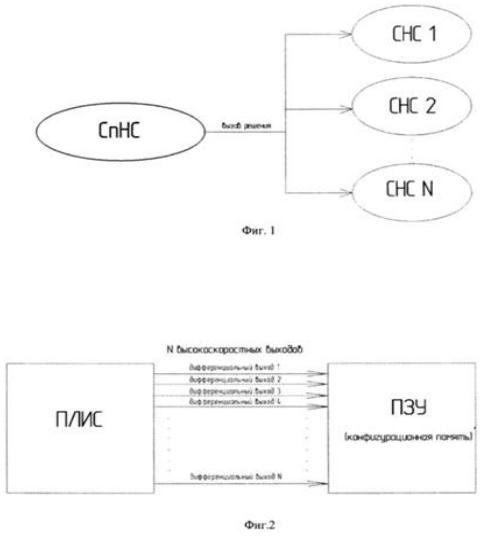


Рисунок 1 Схемы патента [53]. По сравнению с прототипом использование специальных нейронных сетей для выбора нейронных сетей и переход на динамически реконфигурируемые нейронные сети позволяет создавать компактные универсальные нейроморфные устройства обработки, приближенные по своим возможностям к человеческому мозгу, способные к выполнению широкого класса задач с нечетким описанием.

Далее указанный блок нейронной сети на основе анализа входных данных текущей решаемой задачи и в соответствии с выше определенными критериями выбирает из конфигурационной памяти соответствующую нейронную сеть и загружает ее в память ПЛИС, обеспечивая таким образом динамическую реконфигурацию ПЛИС в процессе работы, затем загруженная нейронная сеть выполняет обработку данных для решения текущей задачи, и, в том случае, если решение текущей задачи, получаемое загруженной нейронной сетью, не удовлетворяет предопределенным критериям, то вышеуказанный блок нейронной сети выполняет динамическую реконфигурацию ПЛИС, загружая из конфигурационной памяти образ другой нейронной сети, удовлетворяющей текущим условиям работы, при этом обеспечивая полную динамическую реконфигурацию или частичную динамическую реконфигурацию, обновляя конфигурацию только части логики ПЛИС в том случае, когда основная

часть конфигурационного образа вновь загружаемой нейронной сети и предыдущей загруженной нейронной сети совпадают. [53]

В препринте arXiv preprint [76] Роман Малашин (F eb 2 0 2 1) рассмотрел задачу обучения разреженному ансамблю CNN, когда агента учат использовать знания нескольких предварительно обученных классификаторов с учетом их вычислительной сложности. Редкое ансамблевое обучение позволяет плавно итеративно увеличивать сложность без переобучения с нуля, потому что сам агент можно рассматриватькак «инструмент»; это может помочь в создании систем, которые постепенно усложняются. В [77] Роман Малашин применяет для нейронных сетей обучение систем анализа изображения с динамически конфигурируемой структурой вычислений. Предлагаемые системы, обучаемые в соответствии с принципом наименьшего действия, полезны для повышения скорости обработки больших объемов данных и могут помочь преодолеть другие недостатки глубоких архитектур.

Основная цель этого обзора состоит в том, чтобы классифицировать перспективные технологии и сопоставить потенциальные развертывания фотонных нейронов с текущими и ближайшими технологическими возможностями, сделав вывод об эффективности использования энергии и занимаемой площади, которые можно рассматривать в рамках прагматической досягаемости сегодняшних технологических перспектив.

Нейроморфные кремниевые фотонные сети

Фотонные нейронные сети появились относительно недавно в связи с развитием кремниевой волоконной фотоники. Так, в России известна работа [55]. Е. Б. Соловьева, А. В. Зубарев. Нейронная модель компенсатора нелинейных искажений сигналов для цифрового канала связи Известия вузов России. Радиоэлектроника. 2013. И в 2015 А.В. Зубарев защитил кандидатскую диссертацию [58]. В США с 2016 года известен препринт группы исследователей из Принстонского университета (Princeton University) [54]. Группа представила один из возможных ответов на этот вопрос: они построили первый в мире интегральный кремний-фотонный нейроморфный чип и показали, что он способен работать и вычислять со сверхвысокой скоростью.

Экспериментальная установка и изображение весовой сети MRR показаны на рис. 2. Образцы были изготовлены на пластинах кремний-на-изоляторе (КНИ) на заводе нанотехнологий в Вашингтоне с помощью группы быстрого прототипирования SiEPIC Ebeam [59]. Толщина кремния составляет 220 нм, а толщина скрытого оксида (BOX) составляет 3 мкм. WG шириной 500 нм были сформированы литографией Ebeam и полностью протравлены до BOX.

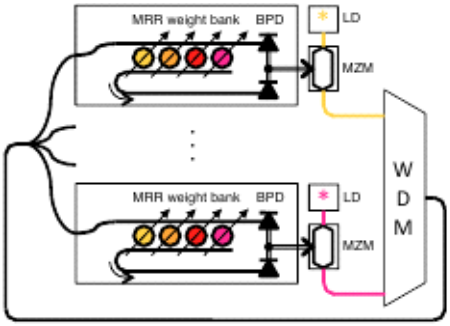


Рисунок 2. Структурная схема интегрального кремний-фотонного нейроморфного чипа[54]. Концепция сети вещания и веса STAR с модуляторами, используемыми в качестве нейронов. MRR: микрокольцевый резонатор, BPD: симметричный фотодиод, LD: лазерный диод, MZM: модулятор Маха-Цендера, WDM: мультиплексор с разделением по длине волны. (3 мкм),

Кремниевая фотоника стала многообещающей КМОП-совместимой альтернативой для реализации нового поколения ускорителей глубокого обучения, которые могут использовать свет как для связи, так и для вычислений. Она была в центре многих беспрецедентных успехов в последние годы в решении очень сложных проблем в областях компьютерного зрения, обработки естественного языка, прогнозирования временных рядов и понимания больших данных. Многие новые приложения, такие как беспилотные автомобили [61], автономная робототехника [62], обнаружение фейковых новостей [63], прогнозирование роста пандемии и тенденций [64], обнаружение сетевых аномалий [65] и языковой перевод в реальном времени [66] опираются на все более сложные модели глубокого –обучения [60].

Процессоры поддерживают все более продвинутые векторные инструкции [67], оба из которых предназначены для ускорения общих матричных и векторных операций. в обработке глубокого обучения. Помимо решений в цифровой области, были также предложены ускорители, которые работают в аналоговой области [68] – [70] или в области аналогово-цифровых смешанных сигналов [71] – [73].

Такие ускорители глубокого обучения на основе кремниевой фотоники могут обеспечить беспрецедентный уровень энергоэффективности и параллелизма. Например, с помощью операций умножения и накопления (MAC), которые доминируют в вычислениях с глубоким обучением, ускорители на основе фотоники могут достичь эффективности использования энергии (определяемой как (MAC / s / mm2) / (joules / MAC)), которая почти в 1000 раз лучше по сравнению с до самых энергоэффективных электронных ускорителей на сегодняшний день [74].

В работе [75] приведены характеристики широкополосного реконфигурируемого линейного оптического процессора 4 × 4 на основе MZI

с низкими потерями. Также исследуется влияние коэффициента затухания MZI на эффективное количество бит (ENOB) на выходе устройства.

В технологиях плазмоника произвела революцию в обычных фотонных модуляторах, резко уменьшив их размеры, предлагая в то же время скорость оптической модуляции до 120 Гбит / с с энергопотреблением в аттоджоулей на бит [36]. Также сообщалось о компактных модуляторах на основе кольцевого резонатора с плазмонной поддержкой для форматов модуляции без возврата к нулю (NRZ) 72 Гбит / с [64].

В работе [38]. Bin Shi, Николой Калабретта и Рипалтаой Стабиле из Технологического университета Эйндховена, (Нидерланды) предложена новая архитектура фотонного ускорителя, основанная на весовом подходе для глубокой нейронной сети через фотонный интегрированный крест. Численно моделируется работа отдельного нейрона и всей нейронной сети. Согласно измерениям, прогнозируется динамический диапазон более 25 дБ. Операция взвешенного сложения также моделируется и анализируется в зависимости от оптических перекрестных помех и количества задействованных входных цветов. В частности, хотя увеличение оптических перекрестных помех отрицательно влияет на смоделированную ошибку, большее количество каналов приводит к лучшей производительности.

В качестве некоторых примеров нейроморфных чипов, IBM TrueNorth [39], Neurogrid [40], SpiNNaker [41] и BrainDrop [42] предназначены для нейронных сетей, в то время как FPGA [43], EIE [44] и Google TPU [45] предназначены для глубоких нейронных сетей. Сообщается, что уровень энергоэффективности составляет порядка нескольких пДж / операцию. Однако скорость вычислений ограничена полосой пропускания электрических соединений. В обэоре [11] рассмотрены так же отечественные нейроморфные чипы.

Исследовательскя группа WinPhos из Университета Аристотеля в Салониках, (Греция) в работе [46] доложила о новых результатах исследований полностью оптических рекуррентных нейронных сети WDM со стробированием. На основе предсказанных в [47 и 48] встроенных систем, которые смогут распознавать объекты, определять, кто говорит, и даже реагировать на раздражители окружающей среды и последующих аппаратных реализациях [49 – 51], они пошли дальше и смогли достичь результатов как в не стробируемом, так и в стробируемом режимах.

В предложенных схемах использовалась сигмоидальная активация на основе полупроводникового оптического усилителя (SOA) в волоконной петле, и они были проверены с использованием асинхронных сигналов с мультиплексированием по длине волны (WDM) с оптическими импульсами 100 пс. В версии Gated-RNN использовался вентиль SOA-Mach-ZehnderInterferometer (SOA-MZI), при этом выход RNN

определял долю входного сигнала, которая требуется для входа в RNN. Наконец, сложная архитектура NN была обучена с использованием набора финансовых данных FI-2010 с использованием предложенных не-стробируемых и стробируемых-RNN, продемонстрировав выдающиеся результаты F1 в 41,68% и 41,85% соответственно, превосходя Multi-Layer Perceptron (MLP). базовых моделей в среднем на 6,49%[46].

2. Успехи глубокого обучения нейроморфных фотонных систем.

Глубокое обучение является апробированной выборкой из широкого семейства методов машинного обучения для представлений данных, наиболее соответствующих характеру задачи. Изображение, например, может быть представлено многими способами, такими как вектор интенсивности значений на пиксель, или (в более абстрактной форме) как множество примитивов, областей определённой формы, и т. д. Удачные представления данных облегчают решение конкретных задач — например, распознавания лиц и выражений лица [78,86]). В системах глубокого обучения автоматизирует сам процесс выбора и настройки признаков, проводя обучение признаков без учителя или с частичным привлечением учителя, используя для этого эффективные алгоритмы и иерархическое извлечение признаков [79].

Такие глубинные нейронные сети (ГНС) могут работать со сложными изображениями в режиме реального времени. К примеру, обученная многослойная нейросеть может распознавать самолёт в необычном ракурсе и на любом фоне. Нейросеть идентифицирует объект как самолёт, даже если это игрушка и он, например, с глазами и в одежде. [81].

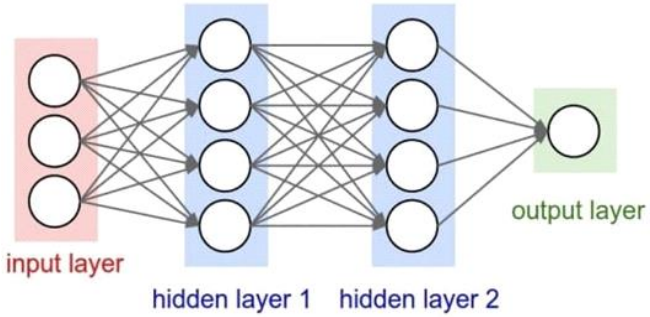


Рисунок 3. Многослойная нейронная сеть [80]

Здесь входные данные отправляются разным слоям нейросетей одновременно, и каждый рассматривает изображения со своего ракурса. Нейроны здесь сформированы в три разных типа слоёв:

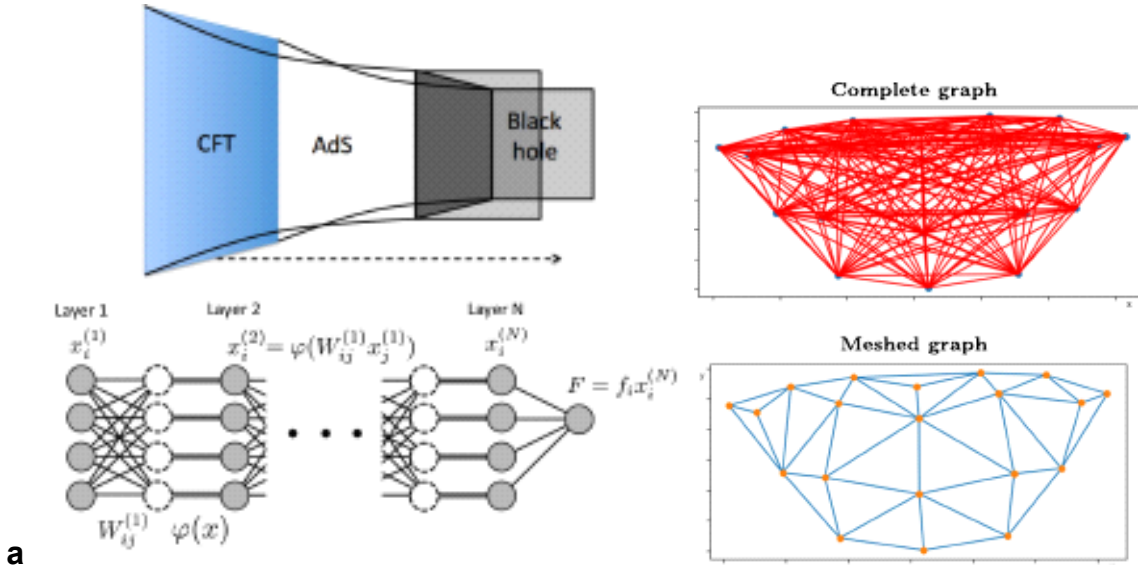
- входной слой;
- скрытые слои;
- выходной слой.

Вычислениями занимаются скрытые слои.

Системы глубокого обучения нашли применение в таких областях, как компьютерное зрение, распознавание речи, обработка естественного языка, аудиораспознавание, биоинформатика. [82]. Глубокое обучение за эти годы проникло в новые сферы: в голографию [19], квантовые вычисления [20-23], глубокого обучения для создания фантомных изображений. [24], в дифракционной визуализация с субволновым разрешением [25], в адаптивное управление свойствами метаматериалов [26], в частности, в создание самонастраивающейся микроволновой маскировки [27] и в создании адаптивных акустических систем [28-29].

2.1. Глубокое обучение в голографии и распознавании выражений лица.

Используя данные решеточной КХД, группа японских исследователей предложила в работе [19] процедуру глубокого обучения голографических данных объемную метрику как веса нейронной сети.



а ***б***

Рисунок 4. а К работе [19]., б К статье [86].Иллюстрации принципов глубокого обучения по сети КХД (а) и по графам обработки лица в статье (б)

2.2. Глубокое обучение в квантовых вычислениях

В последнее время возросшая вычислительная мощность и доступность данных, а также достижения в области алгоритмов привели к впечатляющим результатам в методах машинного обучения в задачах регрессии, классификации, генерации данных и обучения с подкреплением. Несмотря на эти впечатляющие результаты, близость к физическим пределам производства микросхем наряду с увеличивающимся размером наборов данных побуждает все большее число исследователей ис-

следовать возможность использования мощности квантовых вычислений для ускорения классических алгоритмов машинного обучения.

Целый ряд публикаций последних лет связывает квантовые вычисления с глубоким обучением. [20-23], Обучение в присутствии шума и определенных вычислительно сложных задач в машинном обучении определены как многообещающие направления в этой области.

2.3. Глубокое обучение в микроволновой маскировке [85] [27]

В статье [85], написанной в 2010 году Ежи Михальски, управляющим SpaceForest был представлен новый метод настройки фильтра резонатора с использованием искусственной нейронной сети (ИНС). Предлагаемый метод не требовал информации о топологии фильтра, и фильтр рассматривался как черный ящик. Чтобы проиллюстрировать концепцию, применяется многослойная нелинейная искусственная нейронная сеть с прямой связью и обратным распространением.

Был предложен метод подготовки, обучения и тестирования векторов, состоящих из дискретных отстроенных характеристик рассеяния и соответствующих отклонений настроечного винта. Для сбора обучающих векторов была создана машина – интеллектуальный инструмент автоматической настройки фильтров, интегрированный с анализатором векторных сетей. ИНС обучалась на основе выборок, полученных от правильно настроенного фильтра. Тем самым было показано, что после процесса обучения ИНС был проведен эксперимент по настройке фильтров с 6 и 11 резонаторами, что доказало очень высокую эффективность представленного метода и открывало возможность для мультиспектральных систем применять методы глубокого обучения ИИ.

Группа исследователей из Китая в июне 2020 в журнале NaturE PHotoNiCS | VOL 14 | JUNE 2020 | 383–390 |поместила результаты экспериментов «Самонастраивающаяся микроволновая маскировка с поддержкой глубокого обучения без вмешательства человека» [27].

Здесь, предлагая концепцию интеллектуальной (то есть самоадаптивной) маскировки, основанной на глубоком обучении, и представляя маскировку метаповерхностью в качестве примера реализации. В эксперименте маскировка метаповерхностью демонстрирует время отклика в миллисекундах на постоянно меняющуюся падающую волну и окружающую среду без какого-либо вмешательства человека, работа приближает доступные стратегии маскировки к широкому спектру приложений реального времени на месте, таких как движущиеся малозаметные машины. Этот подход открывает путь к упрощению других интеллекту-

альных метаустройств в микроволновом режиме и в более широком электромагнитном спектре и, в более общем плане, позволяет автоматически решать электромагнитные обратные задачи проектирования.

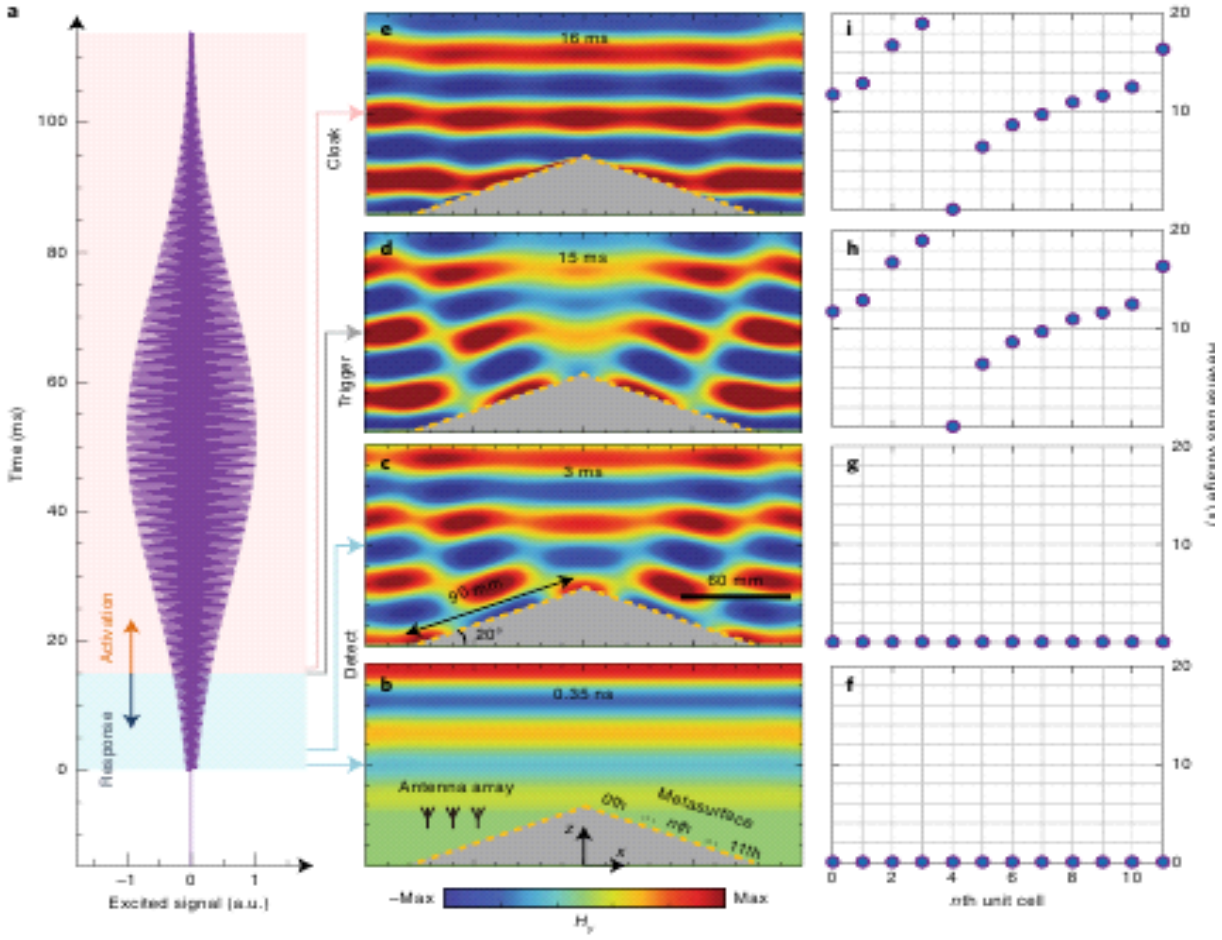


Рисунок 5. Переходный отклик самоадаптивной маскировки при моделировании FDtD. [27]

a, гауссов импульс, излучаемый с z = 120 мм, падает на треугольный выступ tPEC, покрытый метаповерхностью b – e: временная эволюция (0,35 нс (b), 3 мс (c), 15 мс (d) и 16 мс (e)) магнитного поля Ну в области наблюдения (120 мм × 240 мм). Во время этого процесса информация об инциденте обнаруживается антенной решеткой и затем подается в предварительно обученную ИНС вместе с гипотетически известным фоном. При t = 15 мс (d) срабатывает маскировка метаповерхности, которая впоследствии делает выпуклость невидимой. f – i, распределения обратных напряжений смещения, соответствующие b – e, соответственно. Благодаря геометрической симметрии напряжения смещения вдоль двух скошенных кромок идентичны при нормальном падении.

Авторы предложили концепцию интеллектуальной маскировки, основанной на глубоком обучении, и реализовали маскировку метаповерхности в качестве примера. Плащ с метаповерхностью, интегрированный с предварительно обученной ИНС, демонстрирует эффективную и надежную самонастраивающуюся способность реагировать на постоянно меняющу-

юся волну и фон без вмешательства человекаЭта концепция была подтверждена симуляциями FDTD и экспериментальным микроволновым экспериментом, который переводит исследования маскировки на следующий этап – интеллектуальные маскировки. Предлагаемая концепция не только прокладывает путь к радикально новым метаустройствам, но и к другим областям больших данных, таким как оптимизация проектирования наноструктур по требованию. В свою очередь, классическая электромагнетизм и оптика будут способствовать развитию глубокого обучения.

Заключение

1. Одним из основных направлений современных когнитивных исследований является искусственный интеллект (ИИ).

2. Под нейроморфными системами понимаются модели искусственных нейронных сетей, архитектура и дизайн которых основаны на особенностях структуры и принципах работы реальных нейробиологических систем.

3. Нейроморфное моделирование находится на пересечении нескольких областей исследований, в том числе нейробиологии, теории нейронных сетей, математического моделирования, электронной техники, кремниевой фотоники, создании адаптивных метаматериалов

4. Связь между фотоникой и вычислениями является столпом современной оптики и предметом передовых исследований. [1]- [4]

5. Основополагающая работа Хинтона и др. в 2006 году, они ввели название техники «Глубокое обучение». (Точностью распознавания цифр (> 98%) [5].) Примеры такого же подхода включают в себя нейроморфную сеть Цурикова А.Н. [6], см. патент РФ, и обучение импульсных нейронных сетей При этом были получены хорошие результаты для разного количества нейронных сетей в ансамбле: 1, 4, 16 и 64. Получены вероятности правильного распознавания от 92.7% до 99.42% [56]. а так же математическую модель спайковых нейронов Ижикевича (Izhikevich) [9] и крупномасштабное моделирование Элиасмита (Eliasmith) [10]

6. С момента написания нашего обзора [11] прошло менее 3х лет, но успехи в области создания нейроморфных фотонных систем с глубоким обучением за эти годы поразительные.

7. В последнее время появился ряд предложений по реализации функции нелинейной активации (NLAF) для определения порога, включая использование модуляторов электропоглощения [7] и модуляторов микрокольца [8], а также ряда полностью оптических реализаций [12–14]. Особое место в обзоре уделено обучению искусственных нейронных сетей [2-5, 15-18]

8. Приведены блок-схемы алгоритма обработки массива обучающих пар и блок-схемаы алгоритма обучения ИНС, а также структурные схе-

мы обучения. Глубокое обучение за эти годы проникло в новые сферы: в голографию [19], квантовые вычисления [20-23], глубокого обучения для создания фантомных изображений. [24], в создание дифракционной визуализация с субволновым разрешением [25], в адаптивное управление свойствами метаматериалов [26], в частности, в создание самонастраивающейся микроволновой маскировки [27] и в создании адаптивных акустических систем [28-29].

9. Ужу были сделаны первые шаги по макетированию элементной базы и программируемого на кристалле нелинейно-оптического процессора [30-32], и намечены пути развития [33-34].

Список литературы:

[1]. T. Ferreira de Lima, A. N. Tait, A. Mehrabian et al., "Primer on silicon neuromorphic photonic processors: architecture and compiler," Nanophotonics, vol. 9, no. 13, pp. 4055–4073, 2020.

[2]. P. Stark, F. Horst, R. Dangel, J. Weiss, and B. J. Offrein,"Opportunities for integrated photonic neural networks," Nanophotonics, vol. 9, no. 13, pp. 4221–4232, 2020.

[3] Y. Shen, N. C. Harris, S. Skirlo et al., "Deep learning with coherent nanophotonic circuits," Nat. Photon., vol. 11, pp. 441–446, June 2017.

[4]. Abdollahramezani, O. Hemmatyar, and A. Adibi, "Meta-optics enabled optical analog computing," Nanophotonics, vol. 9, no. 13, pp. 4075–4095, 2020.

[5]. G.E. Hinton, S. Osindero, Yu.-V. Tech, "Fast Learning Algorithm for Deep Trust Networks," Neural Computing, vol. 18, no. 7, pp. 1527-1554, 2006

[6]. Цуриков А.Н. Программно-алгоритмическое и структурное обеспечение систем поддержки принятия решений в чрезвычайных ситуациях на железнодорожном транспорте: автореф. дисс. … канд. техн. наук. – Ростов-на-Дону, 2014. Цуриков А.Н. Способ обучения искусственной нейронной сети // Патент на изобретение РФ, RU 2504006 C1, опубликовано 10.01.2014 г.

[7]. R. Amin, J. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, T. El-Ghazawi, and V. J. Sorger, APL Mater. 7, 081112 (2019).

[8]. A. N. Tait, T. F. De Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Phys. Rev. Appl. 11, 064043 (2019).

[9]. Izhikevich EM (2003) Simple model of spiking neurons. IEEE Transactions on Neural Networks 14: 1569–1572.

[10]. Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, et al. (2012) A large scale model of the functioning brain. Science 338: 1202–1205

[11]. Мельников Г.С., Мельникова Э.И., Самков В.М., Бондаренко В.А. Нейроморфные системы распознавания изображений (обзор)//Труды Международного научно-технического конгресса «Интеллектуальные системы и информационные технологии-2020» («ИС& ИТ-2020», «IS&IT'20»). Научное издание в 2-х т.Т.2. –Таганрог: Изд-во Ступина С.А., 2020. – 475с. – (С.120-148).

[12]. Thorsten S. Rasmussen, Yi Yu, Jesper Mork All-optical non-linear activation function for neuromorphic photonic computing using semiconductor Fano lasers. Vol. 45, No. 14 / 15 July 2020 / Optics Letters
[13]. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, Opt. Mater. Express 8, 3851 (2018).
[14]. G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, Opt. Express 27, 9620 (2019).
[15]. FEBIN P SUNNY, EBADOLLAH TAHERI, et al. A Survey on Silicon Photonics for Deep Learning, LicenseCC BY-NC-ND 4.0, COLORADO STATE UNIVERSITY, January 2021
[16]. Xuecong SunXuecong SunHan JiaHan JiaYuzhen YangYuzhen Yang, Acoustic Structure Inverse Design and Optimization Using Deep Learning., LicenseCC BY 4.0., DOI: 10.21203/rs.3.rs-255615/v1., January 2021
[17]. Bowen BaiHaowen ShuHaowen ShuXingjun WangXingjun WangWeiwen ZouWeiwen Zou., Towards silicon photonic neural networks for artificial intelligence., Sciece China. Information Sciences 63(6), June 2020
[18]. Thorsten S. RasmussenThorsten S. RasmussenYi YuJesper Mork., All-optical non-linear activation function for neuromorphic photonic computing using semiconductor Fano lasers., Optics Letters 45(14). June 2020,
[19]. Koji Hashimoto, Koji Hashimoto, Sotaro Sugishita, Akinori Tanaka, Akio Tomiya. Deep learning and holographic QCD. PHYSICAL REVIEW D 98, 106014 (2018)
[20]. Avinash Chalumuri, Raghavendra. Kuhne, Manoj B.S. Training an Artificial Neural Network Using Qubits as Artificial Neurons: A Quantum Computing Approach, Computer Science Procedures 171: 568-575., January 2020
[21]. Maxwell Philip, Henderson Maxwell, Philip Henderson Samriddhi Shakya Shashindra Quantum Neural Networks: Image Recognition Using Quantum Circuits, Quantum Machine Intelligence 2 (1): 1-9., February 2020 DOI: 10.1007 / s42484-020-00012-y
[22]. Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L. Fang‡. Quantum Long Short-Term Memory Computational Science Initiative, Brookhaven National Laboratory.arXIV:2009.01783v1/ (Dated: September 4, 2020)
[23]. Sau Lan WuJay ChanWen GuanShow all 15 authors. Application of Quantum Machine Learning using the Quantum Variational Classifier Method to High Energy Physics Analysis at the LHC on IBM Quantum Computer Simulator and Hardware with 10 qubits. arXiv:2012.11560v1 [quant-ph] 21 Dec 2020
[24]. Tong Bian, Tong Bian, Yuxuan Yi, Jiale Huyu. A residual deep learning approach for generating ghost images. Scientific reports 10 (1). DOI: 10.1038 / s41598-020-69187-5. December 2020
[25]. Abantika Ghosh,Abantika Ghosh, Diane J Rothб Luke H. Nicholls Show all 6 authors Viktor A. Podolskiy. Machine l earning -- based diffractive imaging with sub-wavelength resolu. Preprintion. May 2020
[26]. Jingkai Weng, Yujiang Ding, Chengbo Hu, Xue-Feng Zhu, Bin Liang✉, Jing Yang & Jianchun Cheng Meta-neural-network for real-time and passive deep-learning-based object recognition. NATURE CO[26].

[27]. Chao Qian, Bin Zheng, Yichen Shen, Li Jing, Erping Li, Lian Shen, and Hongsheng Chen ✉ Self-tuning microwave masking with deep learning support without human intervention. NaturE PHotoNiCS | VOL 14 | JUNE 2020 | 383–390
[28]. Xuecong Sun, Han Jia, Yuzhen Yang, Han Zhao. Acoustic Structure Inverse Design and Optimization Using Deep Learning
[29]. Zhaoyong Sun, Zhaoyong Sun, Xuecong Sun, Xuecong Sun, Han Jia, Han Jia. Quasi-isotropic underwater acoustic carpet cloak based on latticed pentamode metafluid. *Applied Physics Letters 114(9):094101. March 2019*
[30]. Chaoran Huang, Aashu Jha, Thomas Ferreira de Lima, Alexander N. Tait, Bhavin J. Shastri,
and Paul R. Prucnal, On-Chip Programmable Nonlinear Optical Signal Processor and Its Applications.
IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS, VOL. 27, NO. 2, MARCH/APRIL 2021
[31]. H. Zhang, M.Gu, X. D. Jiang, J. Thompson An optical neural chip for implementing complex-valued neural networkю NATURE COMMUNICATIONS | (2021) 12:457
[32]. Mitchell A. Nahmias, Hsuan-Tung Peng, Thomas Ferreira de Lima, Chaoran Huang, Alexander N. Tait, Bhavin J. Shastri and Paul R. Prucnal. A TeraMAC Neuromorphic Photonic Processor. Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544 USA
[33]. Paul R. PrucnalA.N. TaitA.N. TaitMitchell A. NahmiasMitchell A. NahmiasShow all 6 authors Bhavin J. Shastri. Photonics for Neuromorphic Computing. Conference: 2018 European Conference on Optical Communication (ECOC). September 2018
[34]. Bowen Bai, Hauen Shu, Hauen Shu, Xingjun Wang, Xingjun Wang, Weiwen Zou, Weiwen Zou. Towards Silicon Photonic Neural Networks for Artificial Intelligence. Sciece China. Information Sciences 63 (6), June 2020
[35]. Angelina Totović, George Dabos, Nikolaos Passalis, Anastasios Tefas and Nikos Pleros: Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap
DOI 10.1109 / JSTQE.2020.2975579, IEEE Journal of Selected Topics in Quantum Electronics JSTQE-INV-PP2020-08305-2019
[36]. B. Baeuerle et al., “120 GBd plasmonic Mach-Zehnder modulator with a novel differential electrode design operated at a peak-to-peak drive voltage of 178 mV,” Opt. Express, vol. 27, no. 12, p. 16823, 2019.
[37] C. Haffner et al., “Low-loss plasmon-assisted electro-optic modulator,” Nature ;556(7702):483-486. doi: 10.1038/s41586-018-0031-4. Epub 2018 Apr 25.018.
[38]. Bin Shi *, Nicola Calabretta and Ripalta Stabile. Numerical Simulation of an InP Photonic Integrated Cross-Connect for Deep Neural Networks on Chip Appl. Sci. 2020, 10, 474; doi:10.3390/app10020474
[39]. Akopyan, F.; Sawada, J.; Cassidy, A.; Alvarez-Icaza, R.; Arthur, J.; Merolla, P.; Imam, N.; Nakamura, Y. Datta, P.; Nam, G.J.; et al. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. IEEE Trans. Comput. Des. Integr. Circuits Syst. 2015, 34, 1537–1557. [CrossRef]

[40]. Benjamin, B.V.; Gao, P.; McQuinn, E.; Choudhary, S.; Chandrasekaran, A.R.; Bussat, J.; Alvarez-Icaza R.;Arthur, J.V.; Merolla, P.A.; Boahen, K. Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations. Proc. IEEE 2014, 102, 699–716. [CrossRef]
[41]. Furber, S.B.; Galluppi, F.; Temple, S.; Plana, L.A. The SpiNNaker Project. Proc. IEEE 2014, 102, 652–665.[CrossRef]
[42]. Neckar, A.; Fok, S.; Benjamin, B.V.; Stewart, T.C.; Oza, N.N.; Voelker, A.R.; Eliasmith, C.; Manohar, R.;Boahen, K. Braindrop: A Mixed-Signal Neuromorphic Architecture with a Dynamical Systems-Based Programming Model. Proc. IEEE 2019, 107, 144–164. [CrossRef]
[43]. Zhang, C.; Li, P.; Sun, G.; Guan, Y.; Xiao, B.; Cong, J. Optimizing FPGA-Based Accelerator Design for Deep Convolutional Neural Networks. In Proceedings of the FPGA 2015–2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, USA, 22–24 February 2015; ACM Press New York, NY, USA, 2015; pp. 161–170.
[44]. Han, S.; Liu, X.; Mao, H.; Pu, J.; Pedram, A.; Horowitz, M.A.; Dally, W.J. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In Proceedings of the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, Korea, 18–22 June 2016; pp. 243–254.
[45]. Jouppi, N.P.; Borchers, A.; Boyle, R.; Cantin, P.; Chao, C.; Clark, C.; Coriell, J.; Daley, M.; Dau, M.; Dean, J.; et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. In Proceedings of the Proceedings-International Symposium on Computer Architecture, Toronto, ON, Canada, 24–28 June 2017; ACM Press: New York, NY, USA, 2017; pp. 1–12.
[46]. George Mourgias-Alexandris, George Dabos, Nikolaos Passalis, Angelina Totović, Anastasios Tefas and Nikos Pleros. ; All-optical WDM Recurrent Neural Networks with Gating.;
DOI 10.1109/JSTQE.2020.2995830, IEEE Journal of Selected Topics in Quantum Electronics
[47]. Park et al., “Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications,” 2018.
[48]. N. P. Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit,” ACM SIGARCH Comput. Archit. News, vol. 45, no. 2, pp. 1–12, 2017.
[49]. M. Davies et al., “Loihi: A Neuromorphic Manycore Processor with OnChip Learning,” IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.
[50]. S. B. Furber et al., “Overview of the SpiNNaker system architecture,” IEEE Trans. Comput., vol. 62, no. 12, pp. 2454–2467, 2013.
[51]. F. Akopyan et al., “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip,” IEEE Trans. Comput. Des. Integr. Circuits Syst., vol. 34, no. 10, pp. 1537–1557, 2015.
[52]. Choudhary S, Sloan S, Fok S, Neckar A, Trautmann E, et al. (2012) Silicon neurons that compute. In: Artificial Neural Networks and Machine Learning – ICANN 2012, Springer Berlin Heidelberg, volume 7552 of Lecture Notes in Computer Science. pp. 121-128

[53]. Малашин Дмитрий Олегович, Малашин Роман Олегович; Способ нейроморфной обработки данных и устройство для его осуществления; Описание патента на изобретение RU2739340C1
[54].. Alexander N. Tait,∗ Ellen Zhou, Thomas Ferreira de Lima, Allie X. Wu, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal; Neuromorphic Silicon Photonics; Princeton University, Princeton, NJ 08544, USA rXiv:1611.02272v1 [q-bio.NC] 5 Nov 2016
[55]. Е. Б. Соловьева, А. В. Зубарев. Нейронная модель компенсатора нелинейных искажений сигналов для цифрового канала связи Известия вузов России. Радиоэлектроника. 2013. Вып. 4,стр 30-34
[56]. Андрей В. Гаврилов, Константин О. Панченко. Методы Обучения Импульсных Нейронных Сетей; 2016 13th International Scientific-Technical Conference APEIE – 39281
[57]. С Соловьева Е.Б., СИНТЕЗ КОМПЕНСАТОРОВ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ ГАММЕРШТЕЙНА Цифровая Обработка Сигналов №1/2012
[58]. Зубарев А. В. РАЗРАБОТКА МЕТОДОВ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НЕЛИНЕЙНЫХ КОМПЕНСАТОРОВ И ФИЛЬТРОВ ИМПУЛЬСНОГО ШУМА НА ОСНОВЕ МНОГОМЕРНЫХ ПОЛИНОМОВ И НЕЙРОННЫХ СЕТЕЙ. АВТОРЕФЕРАТ диссертации на соискание ученой степени кандидата технических наук,
https://etu.ru/assets/files/nauka/dissertacii/2015/Zubarev/Avtoreferat_a.10.pdf
[59]. L. Chrostowski and M. Hochberg, Silicon Photonics Design: From Devices to Systems (Cambridge University Press, 2015).
[60]. F. Sunny, E. Taheri, M. Nikdast, and S. Pasricha. A Survey on Silicon Photonics for Deep Learning. Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373;
[61]. [2] V. K. Kukkala, J. Tunnell, S. Pasricha, and T. Bradley, “Advanced driver-assistance systems: A path toward autonomous vehicles,” IEEE Consumer Electronics Magazine, vol. 7, no. 5, pp. 18–25, 2018
[62] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” The International Journal of Robotics Research, vol. 37, no. 4-5, pp. 421–436, 2018
[63] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media using geometric deep learning,” arXiv:1902.06673 [cs], Feb. 2019
[64] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for covid-19 (sars-cov2) pandemic: A review,” Chaos, Solitons & Fractals, p. 110059, 2020
[65] K. Kukkala, S. V. Thiruloga, and S. Pasricha, “Indra: Intrusion detection using recurrent autoencoders in automotive embedded systems,” arXiv preprint arXiv:2007.08795, 2020
[66]. J. Gu, G. Neubig, K. Cho, and V. O. Li, “Learning to translate in real-time with neural machine translation,” arXiv preprint arXiv:1610.00388, 2016

[67]. A. Sodani, R. Gramunt, J. Corbal, H.-S. Kim, K. Vinod, S. Chinthamani, S. Hutsell, R. Agarwal, and Y.-C. Liu, “Knights landing: Second-generation intel xeon phi product,” IEEE micro, vol. 36, no. 2, pp. 34–46, 2016
[68] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P.Strachan, M. Hu, R. S. Williams, and V. Srikumar, “Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars,” ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016
[69] L. Song, X. Qian, H. Li, and Y. Chen, “Pipelayer: A pipelined reram-based accelerator for deep learning,” 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 541–552, 2017
[70] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, “Recent progress in analog memory-based accelerators for deep learning,” Journal of Physics D: Applied Physics, vol. 51, no. 28, p. 283001, 2018
[71] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, “A 55nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots,” pp. 124–126, 2018
[72] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, “A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement,” pp. 141–142,2018
[73] A. Amaravati, S. B. Nasir, J. Ting, I. Yoon, and A. Raychowdhury, “A55-nm, 1.0–0.4 v, 1.25-pj/mac time-domain mixed-signal neuromorphic accelerator with stochastic synapses for reinforcement learning in autonomous mobile robots,” IEEE Journal of SolidState Circuits, vol. 54,no. 1, pp. 75–87, 2018
[74]. A. R. Totovic, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, “Femtojoule per mac neuromorphic photonics: An energy and technology roadmap,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 5,pp. 1–15, 2020
[75]. Lorenzo De Marinis, Odile Liboiron-Ladouceur, Nicola Andriolli. Characterization and ENOB Analysis of a Reconfigurable Linear Optical Processor. Conference: Photonics in Switching and Computing. DOI: 10.1364/PSC.2020.PsW1F.4. January 2020
[76]. Roman O. Malashin. SPARSELY ENSEMBLED CONVOLUTIONAL NEURAL NETWORK CLASSIFIERS VIA REINFORCEMENT LEARNING. arXiv preprint, a r X i v: 2 1 0 2. 0 3 9 2 1 v 1 [c s. LG] 7 F eb 2 0 2 1
[77]. Р. О. Малашин. Принцип наименьшего действия в динамически конфигурируемых системах анализа изображений. Оптический журнал. / Том 86 № 11 /Ноябрь 2019.
[78]. Glauner, P. (2015). Deep Convolutional Neural Networks for Smile Recognition (MSc Thesis). Imperial College London, Department of Computing. arXiv:1508.06535.
[79]. Song, Lee, Neural Information Processing, 2013
[80]. Интернет источник: https://gb.ru/posts/glubokoe-obuchenie-nejronnyh-setej-i-mashinnoe-obuchenie

[81]. Ле Мань Ха. Свёрточная нейронная сеть для решения задачи классификации; ТРУДЫ МФТИ. 2https://gb.ru/posts/glubokoe-obuchenie-nejronnyh-setej-i-mashinnoe-obuchenie016. Том 8, № 3. стр. 91-97
[82]. Кустикова Валентина, Образовательный курс «Введение в глубокое обучение с использованием Intel® neon™ Framework» Введение в глубокое обучение/ http://hpc-education.unn.ru/files/courses/intel-neon-course/Rus/Lectures/Presentations/1_Deep%20learning%20intro.pdf
[83]. Carpenter G., Grossberg S. Pattern Recognition by Self-Organizing Neural Networks. – Cambridge, MA, MIT Press, 1991. P.601
[84]. Тэцуя Акутагава Кодзи, Хашимото Кодзи Хашимото, Сумимото Такаяки. Глубокое обучение и AdS / QCD. DOI: 10.1103 / PhysRev D.102.026020. Июль 2020 г.
[85].Jerzy Michals, kiJerzy Michalski. Artificial neural networks approach in microwave filter tuning. Progress In Electromagnetics Research M 13:173-188\DOI: 10.2528/PIERM10053105 January 2010
[86]. Riccardo Mengoni, Massimiliano Incudini, Alessandra Di Pierro.. Facial expression recognition on a quantum computer. Quantum Machine Intelligence. (2021)https://doi.org/10.1007/s42484-020-00035-5

ПЕРЕВОДЫ СТАТЕЙ

КРЕМНИЕВАЯ ФОТОНИКА ДЛЯ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Маркес Б.А., Филипович М.Дж., Ховард Э.Р., Бангари В., Гуо Ч., Морисон Х.Д., Феррейра де Лима Т., Тейт А.Н., Прунал П.Р., Шастри Б.Дж.

CO 80305, USA *bama@queensu.ca

https://doi.org/10.1051/photon/202010440

Искусственный интеллект, поддерживаемый нейронными сетями, позволил применять его во многих областях (например, медицина, финансы, автономные транспортные средства). Программные реализации нейронных сетей на обычных компьютерах ограничены по скорости и энергоэффективности. Нейроморфная инженерия нацелена на создание процессоров, в которых оборудование имитирует нейроны и синапсы мозга для распределенной и параллельной обработки. Нейроморфная инженерия с помощью кремниевой фотоники может обеспечить субнаносекундные задержки и может расширить область применения искусственного интеллекта до высокопроизводительных вычислений и сверхбыстрого обучения. Мы обсуждаем текущий прогресс и проблемы, связанные с этими демонстрациями, для масштабирования до практических систем обучения и вывода.

Налоговые вычисления недавно были признаны потенциальным средством снижения затрат энергии и времени на выполнение таких алгоритмов, как глубокие нейронные сети. Аналоговое специализированное оборудование требует производства машин для физического моделирования каждого отдельного компонента таких сетей. Это оказывается серьезной проблемой, поскольку современные глубокие сети масштабируются до тысяч или даже миллиардов нейронов для решения сложных задач искусственного интеллекта (ИИ). Чтобы использовать аналоговые машины для отображения схем мозга, необходимо смоделировать функции биологических нейронов. Наиболее распространенные нейронные модели – это искусственные нейроны и перцептроны. В то время как пикирование искусственных нейронов биологически реалистично, область ИИ в настоящее время основана на перцептронах [1]. Персептроны реализуют операции умножения-накопления (MAC). Операции MAC служат для количественной оценки количества умножений и сложений, необходимых для работы глубоких сетей. Персептрон из M

входов может выполнять M MAC-операций за временной шаг. Несколько операций MAC могут выполняться параллельно для реализации любого типа искусственной нейронной сети (ИНС). MAC в настоящее время являются наиболее обременительным узким местом в оборудовании ANN; например, для глубокой сети AlexNet требуется 724 миллиона MAC-адресов для решения ImageNet [2]. Фотонная платформа в настоящее время является одной из самых многообещающих технологий для решения дорогостоящих вычислений, выполняемых глубокими сетями. Кремниевая фотоника предлагает высокую масштабируемость, широкую полосу пропускания, компактность и низкое потребление энергии [3]. Широкополосный и многоволновый параллельный свет позволяет обрабатывать оптическую информацию с высокой скоростью передачи данных. Способность нейроморфных фотонных систем обеспечивать существенное улучшение наших вычислительных возможностей становится все ближе, возможно, со скоростью обработки петамак в секунду / мм2. В этой статье мы описываем фотонную схему, которая может выполнять параллельные операции MAC на кристалле, и представляем две фотонные платформы, которые позволяют

аппаратное ускорение FOCUS AI: i) специализированное фотонная архитектура для выполнения алгоритма прямого согласования с обратной связью (DFA) для обучения нейронной сети [6], и ii) реализация нейронной сети с долгосрочной памятью (LSTM) [7]. Оба предложенных дизайна предлагают фундаментальные преимущества в скорости и пропускной способности по сравнению с цифровыми электронными реализациями.

ИСТОРИЯ ВОПРОСА: НЕЙРОНАУКА И ВЫЧИСЛЕНИЯ

Цифровые компьютеры – это обычно вычислительные системы, которые выполняют логические и математические операции с высокой точностью. В настоящее время такие сложные системы значительно превосходят человеческие возможности для вычислений и памяти. Тем не менее, если бы мы сравнили человека-агента с цифровой машиной, мы бы увидели, что существует множество абстракций, которые необходимо сделать для проведения однозначных сравнений. Такие абстракции предполагают, что когнитивные процессы человека полностью процедурны и следуют стандартной логике. Однако большинство когнитивных действий человека не следует четко определенным инструкциям. Следовательно, сопоставление "один к одному" между человеком и цифровыми компьютерами может не подходить. Аналоговые нейроморфные вычислительные подходы могут больше подходить для имитации процессов человеческого мозга. Цель состоит в том, чтобы создать взаимно однозначное отображение между нейронной системой и анало-

говой машиной, где каждая биологическая величина моделируется эквивалентной аналоговой искусственной моделью. Для такой архитектуры, как человеческий мозг, это может быть серьезным требованием. Человеческий мозг содержит около 100 миллиардов нейронов и 100 триллионов синаптических соединений, которые должны быть представлены в искусственной машине. Тем не менее, часть схемы мозга все еще может быть представлена в искусственной машине для моделирования некоторых когнитивных процессов человека. В последнее время наиболее значительный прогресс в области ИИ был достигнут с использованием персептрона, показанного на рис. 1, в качестве искусственной модели нейрона. Выходной сигнал y нейрона представляет собой сигналы, посылаемые аксоном биологического нейрона, и математически описывается выражением

$$y = f(W \cdot x + b).$$

Входы xi передают информацию нейрону через веса Wi, которые соответствуют силе синапсов. Суммирование всех взвешенных входных сигналов и их преобразование с помощью функции активации f связаны с физиологической ролью тела клетки нейрона. Смещение b представляет собой дополнительную переменную, которая остается в системе, даже если другие входные данные отсутствуют. ИНС построены с использованием перцептронов в качестве нейронных примитивов, так что синаптические связи либо положительные, либо отрицательные, чтобы имитировать возбуждающее и тормозящее нейронное поведение. Функция нелинейной активации может использоваться для определения активированного и деактивированного поведения в искусственных нейронах. ИНС можно разделить на категории с прямой связью (где соединения между нейронами не образуют цикла) или рекуррентные нейронные сети (где существуют циклы). Попытки построить быстрые и эффективные ИНС на основе персептрона представлены на рис. 1

В последние годы появились сообщения о фотонике и искусственном интеллекте. Интересный метод ускорения вычислений заключается в использовании аппаратных модулей для выполнения операций MAC на высоких скоростях. Блок MAC выполняет процессы умножения и накопления: (a + w.x). Несколько операций MAC могут выполняться параллельно для выполнения сложных операций, таких как свертки и цифровые фильтры. MAC обычно используются в реализациях ИНС в цифровой электронике [4].

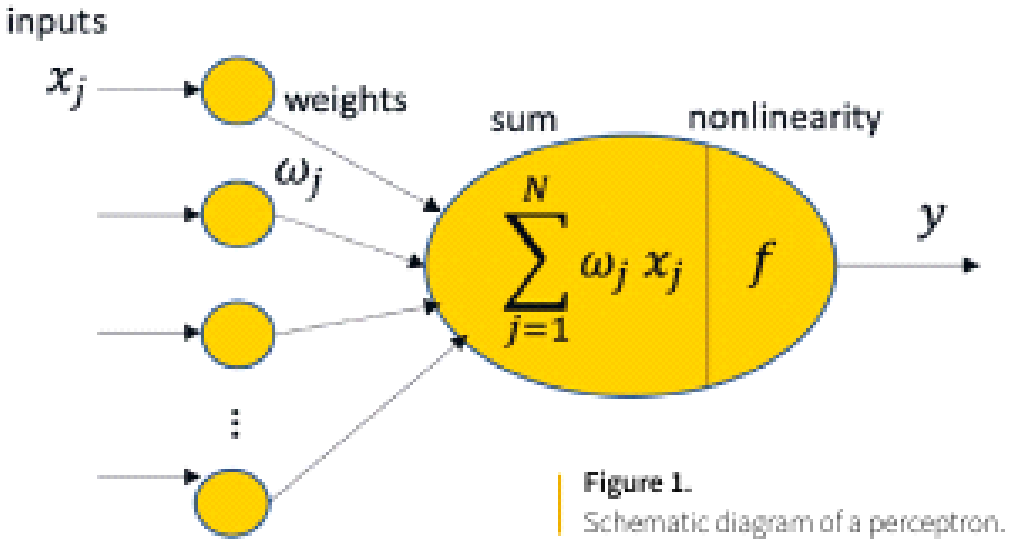


Рис. 1. Принципиальная схема персептрона.

Тем не менее, сериализация слагаемых для выполнения взвешенного сложения делает этот процесс неэффективным; следовательно, разработчики микросхем ищут альтернативные решения, такие как полный параллелизм. Одна из наиболее перспективных технологий для этой цели основана на фотонной платформе.

ФОТОНИЧЕСКИЙ ПЕРЦЕПТРОН И ОПЕРАЦИИ MAC

Масштабируемая фотонная архитектура, реализующая параллельные MAC, может быть достигнута с использованием внутрикристальных технологий мультиплексирования с разделением по длине волны (WDM) [8]. В этой конструкции используются микрокольцевые резонаторы (MRR) [9], то есть фотонные синапсы, для кодирования входных значений и весов в сигналы с несколькими длинами волн. Включение и выключение резонанса данного MRR изменяет передачу каждого сигнала через соответствующий фильтр, эффективно умножая сигнал на желаемый вес. Преимущество использования MRR – это возможность настраивать значения веса с использованием множества различных методов: термического, электрооптического или посредством поглощения света, такого как фазовый переход или графеновые материалы. В данной работе настройка выполняется путем термической модификации показателя преломления MRR-волновода.

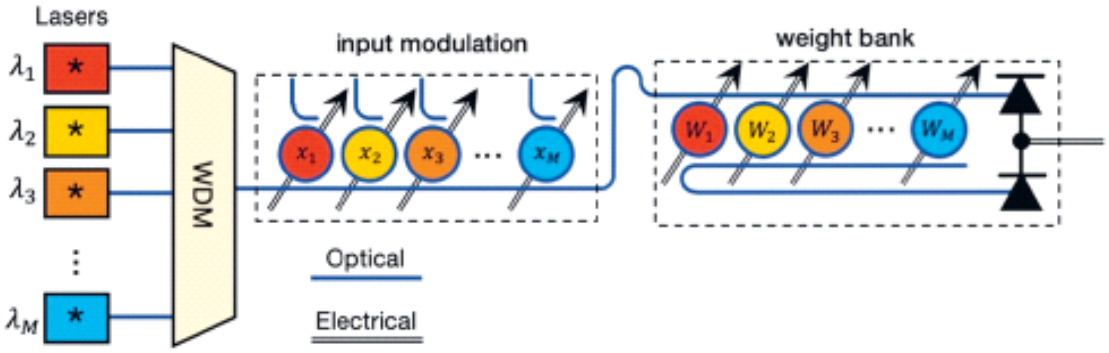


Рисунок 2. Весовой банк MRR с добавлением-сбросом со сбалансированным фотодетектором, реализующим M поэлементных умножителей для выполнения N операций MAC параллельно

Приложение значений напряжения к нагревателю позволяет нам отображать действительные числа на устройстве.

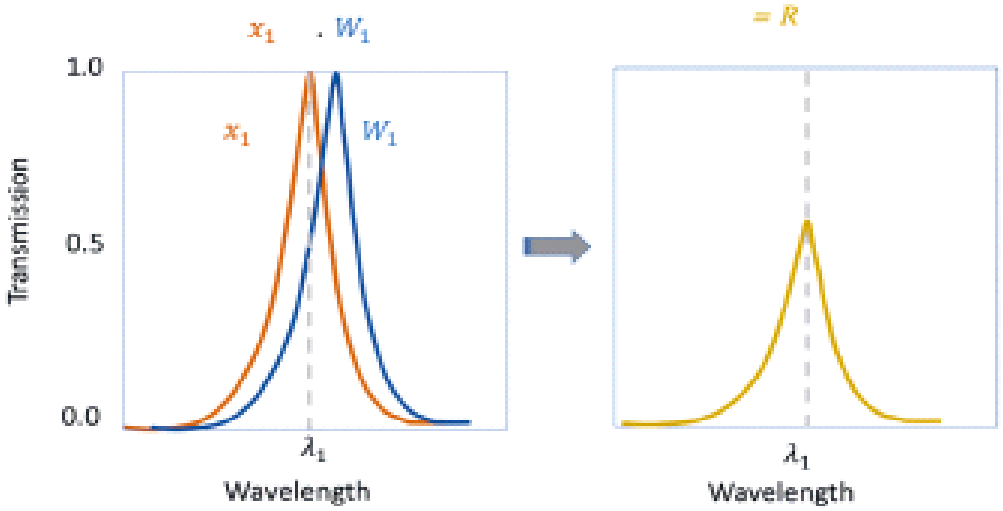


Рис. 3. (a) Кривые зависимости пропускания от длины волны для двух разных MRR (MRR (x1), MRR (W1), выполняющих поэлементное оптическое умножение, и (b) продукт такого умножения.

Массив M MRR может имитировать взвешенное добавление одного нейрона, если в модель включены MRR и сбалансированный фотодетектор, как показано на рис. 2. На этой иллюстрации мы показываем, как параллельно выполнять операции M MAC в фотонике. Входные значения нейрона могут быть сопоставлены со значениями напряжения Vi, которые настраивают каждый отдельный MRR (xi). Каждое значение напряжения имеет однозначное соответствие с профилем передачи MRR Ti, и тот же принцип применяется для значений веса. Экспериментальная реализация этого метода требует использования M лазеров с разными длинами волн λi (с i = 1,…, M), которые представляют каналы M. Два

MRR с разными конфигурациями включения и выключения резонанса на одной и той же длине волны λ1 будут поэтому выполнять поэлементное умножение, как показано на рис.. Здесь представлена w иллюстрация умножения между двумя передающими элементами x1 и W1, что дает результирующее значение R. На рис. 3 (a) элемент x1 настроен на максимальное оптическое пропускание, тогда как W1 настроен на половину максимума. Для реализации x1 MRR (x1) устанавливается в резонанс с λ1, а MRR (W1) настраивается на половину вне резонанса с той же длиной волны. Они представляют собой действительные числа 1 и 0,5 соответственно. Результат такого умножения, показанный на рис. 3 (б), равен R = 0,5. Аналогичный процесс выполняется с остальными наборами (MRR (xi), MRR (Wi)) для i> 1. После того, как взвешенное сложение выполнено с использованием сбалансированного фотодетектора, можно добавить нелинейную функцию на кристалле с помощью микрокольца. модулятор. На основе этой схемы мы можем проектировать системы для решения множества сложных задач ИИ. В следующих разделах мы опишем, как эффективно реализовать обучение и логический вывод ИНС на фотонных чипах.

ПРИМЕНЕНИЕ

Для реализации ИНС на фотонных микросхемах мы складываем N поэлементных умножителей, которые выполняют взвешенное сложение, как показано на рис. 4

Входные значения N × M, полученные от цифро-аналоговых преобразователей (ЦАП), модулируют интенсивность группы. M лазеров с одинаковой мощностью, но с уникальными длинами волн. Эти модулированные входные данные отправляются в массив фотонных весовых банков размером N × M (загруженных из ЦАП), которые затем выполняют умножение для каждого канала. Эта архитектура является общим представлением многоволновой платформы, поскольку ее можно использовать для вывода, как показано в [8], а также для обучения на месте.

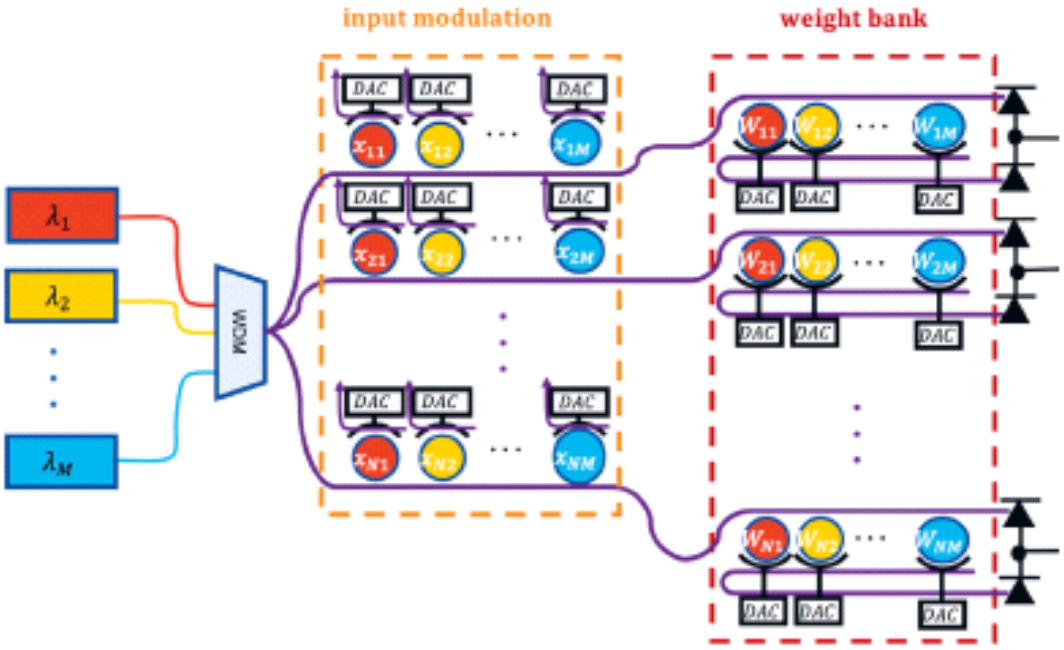


Рисунок 4. Входные значения и значения ядра модулируют MRR посредством электрических токов, пропорциональных этим значениям.

ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ НА ЧИПЕ

Используя преимущества скорости и энергии фотоники по сравнению с традиционными цифровыми компьютерами, алгоритм обучения DFA может быть реализован на кремниевом фотонном оборудовании [6]. Алгоритм DFA – это алгоритм обучения с учителем для обучения ИНС, в котором ошибка распространяется через фиксированные случайные обратные связи непосредственно от выходного слоя к скрытому. Алгоритм DFA использовался для обучения ИНС с использованием наборов данных MNIST, Cifar-10 и Cifar-100 и дает сопоставимую производительность с популярным обучением обратного распространения информации. Благодаря преимуществам фотоники по скорости и энергии по сравнению с традиционными цифровыми компьютерами, алгоритм обучения DFA может быть реализован на кремниевом фотонном оборудовании Алгоритм фотоники и искусственного интеллекта [10]. Фотонная интегральная схема DFA может быть спроектирована с двумя соединенными блоками с M = 10 и N = 100. Эта конструкция может выполнять 2000 MAC за проход, обеспечивая обновление веса между двумя слоями по 1000 нейронов за 1000 проходов.

ДОЛГОВРЕМЕННАЯ НЕЙРОННАЯ СЕТЬ ПАМЯТИ

Подобно схеме DFA, сети LSTM [11] также могут быть реализованы с использованием многоволновой фотонной архитектуры [7]. Сеть LSTM – это повторяющаяся архитектура, которая предлагает преимуще-

ства для обработки временных рядов. Нейроморфные фотонные LSTM предлагают решение для растущего спроса на высокоскоростные нейронные сети с высокой пропускной способностью в приложениях временных рядов, включая обработку видео, автономное вождение и оптическую связь. Производительность фотонного LSTM для задач вывода была протестирована путем применения сети к простой одномерной задаче данных временных рядов в моделировании. Моделирование этой задачи демонстрирует, что даже очень маленькие фотонные сети LSTM, выполняющие до 64 MAC за проход, могут быть очень эффективными при выполнении задач вывода данных временных рядов.

ЗАКЛЮЧЕНИЕ

Нейроморфная фотоника обещает захватывающие разработки для будущего ИИ. В попытке расширить возможности цифровых компьютеров для приложений ИИ, широкополосная работа и полная программируемость аналоговых фотонных интегральных схем могут облегчить сверхбыстрое обучение и логический вывод ИНС. Современные реализации фотонных машин сталкиваются со сложными техническими проблемами, которые многие исследовательские группы и компании начали решать, включая управление процессором и эффективный доступ к памяти. Успешные решения этих проблем могут позволить широкое внедрение фотонных процессоров для решения практических приложений ИИ.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning (The MIT Press, 2016)
[2] V. Sze, Y.-H. Chen, T.-J. Yang et al., Proc. IEEE 105, 2295 (2017)
[3] P.R. Prucnal, B.J. Shastri, Neuromorphic Photonics, (CRC Press, 2017)
Electron. 26, 1 (2020)
[5] J. Feldmann, N. Youngblood, C.D. Wright et al., Nature 569, 208 (2019)
[6] M. J. Filipovich, Z. Guo, B. A. Marquez, et al., in Proc. IEEE Photon. Conf. (IPC), paper TuA3.2 (2020)
[7] E. R. Howard, B. A. Marquez, B. J. Shastri, in Proc. IEEE Photon. Conf. (IPC), paper TuA3.2 (2020)
[8] V. Bangari et al., IEEE J. Sel. Top. Quant. Electron. 26, 7701213 (2020)
[9] A.N. Tait, T. Ferreira de Lima, E. Zhou et al., Sci. Rep. 7, 7430 (2017)
[10] A. Nøkland, Adv. Neural Inf. Process Syst. 29, 1037 (2016)
[11] S. Hochreiter, J. Schmidhuber, Neural Comput. 9, 173 (1997)11] С. Хохрайтер, Дж. Шмидхубер, Neural Comput. 9, 173 (1997).

МЕТА-НЕЙРОННАЯ СЕТЬ ДЛЯ РАСПОЗНАВАНИЯ ОБЪЕКТОВ В РЕАЛЬНОМ ВРЕМЕНИ И ПАССИВНОГО ГЛУБОКОГО ОБУЧЕНИЯ

Вэн Ц., Дин Ю., Ху Ч., Чжу С.-Ф., Лян Б., Ян Ц., Чэн Ц.

https://doi.org/10.1038/s41467-020-19693-x OPEN

Анализ рассеянной волны распознавание объекта имеет фундаментальное значение в волновой физике. Недавно появившаяся техника глубокого обучения достигла большого успеха в интерпретации волнового поля, например, в ультразвуковом неразрушающем контроле и диагностике заболеваний, но обычно требует длительной компьютерной постобработки или дифракционных элементов большого размера. Здесь мы теоретически предлагаем и экспериментально демонстрируем чисто пассивную и компактную мета-нейронную сеть для распознавания сложных объектов в реальном времени путем анализа акустического рассеяния. Мы доказали, что мета-нейронная сеть имитирует стандартную нейронную сеть, несмотря на ее компактность, благодаря уникальной способности ее метаматериальных элементарных ячеек (названных метанейронами) производить сдвиг фазы в глубину субволновой длины в качестве параметров обучения. Получившееся в результате устройство демонстрирует «интеллект» для выполнения желаемых задач с потенциалом преодоления текущих ограничений, продемонстрированный двумя отличительными примерами распознавания рукописных цифр и распознавания смещенных вихрей орбитального углового момента. Наш механизм открывает путь к новым парадигмам глубокого обучения на основе метаматериалов и позволяет концептуальным устройствам автоматически анализировать сигналы с далеко идущими последствиями для акустики и связанных с ней областей.

Из множества важных приложений. Типичные примеры в акустике варьируются от медицинского ультразвукового исследования1 до промышленной неразрушающей оценки2 и подводного обнаружения3. В отличие от традиционных механизмов, которые полагаются на людей-экспертов, таких как врачи, интерпретирующие медицинские ультразвуковые изображения в клинике, которые неизбежно страдали бы от низкой эффективности, потенциальной усталости и большого разнообразия патологий4–6, недавнее появление компьютерных технологий методы глубокого обучения7 достигли высочайшего уровня в решении важной проблемы идентификации и классификации медицинских изображений рассеянных акустических полей, таких как обнаружение анатомических структур, диагностика заболеваний и т. д.6,8,9, среди прочего другие

увлекательные приложения для распознавания речи10–12, анализа эмоций13–16 и т. д., однако, несмотря на заметное улучшение производительности и упрощение процесса, такой перенос бремени с человека на компьютеры все же вызовет проблему вычислительная сложность, энергоснабжение, размер и стоимость устройства из-за их зависимости от точных акустических изображений, которые необходимо измерять с помощью сканирования датчика и компьютера постобработка на основе. Поэтому важно постоянно развивать новые механизмы, основанные на глубоком обучении, с более простой конструкцией, меньшей занимаемой площадью, более высокой скоростью, меньшим потреблением энергии и меньшим количеством датчиков, что было бы жизненно важно для реальных приложений во многих различных сценариях, таких как получение медицинских изображений, где очень желательна быстрая и легкая оценка тканей. В этой статье мы преодолеем такие фундаментальные барьеры, представив физический механизм использования пассивной метанейронной сети, содержащей трехмерную матрицу элементарных ячеек из метаматериала, каждая из которых служит метанейроном, для имитации аналогичной нейронной сети. для классических волн с компактностью, простотой и возможностью аппаратного решения задач. Недавнее быстрое расширение областей исследований фотонных / фононных кристаллов17–22 и метаматериалов23,24 позволяет нетрадиционным манипуляциям волновых полей, таким как аномальное преломление / отражение25,26, невидимость27,28, выпрямление29,30 и т. д., В детерминированном виде. образом, опираясь на рациональный дизайн, основанный на человеческих знаниях. В последние несколько лет мы стали свидетелями значительных усилий, направленных на применение машинного обучения в этих искусственных структурах, но просто нацеленных на разработку активных устройств формирования изображений с уменьшенной сложностью31 или метаматериалов для создания определенных волновых полей32–35. Недавно было доказано, что использование пассивных нейронных сетей возможно благодаря использованию дифракционных слоев с локально модулированной толщиной в соответствии с результатами обучения машинного обучения36, которое генерирует квазинепрерывные фазовые профили и приводит к значительному изменению фазы только на расстоянии в масштабе длины волны37. Кроме того, теоретически предлагается нейронная сеть на основе оптического метаматериала с метаповерхностями38 или наноструктурированной средой39. Напротив, здесь мы представляем теоретические и экспериментальные работы по наделению пассивных акустических метаматериалов «интеллектом» для выполнения сложных задач машинного обучения. Мы доказываем, что исключительная способность метаматериалов обеспечивать резкий фазовый сдвиг в глубинных субволновых

масштабах во всех трех измерениях имеет решающее значение для эквивалентности между обычной и предлагаемой нейронными сетями, и используем компьютер для обучения разработанной метанейронной сети путем итеративной настройки полный фазовый профиль каждого слоя метанейронов. Результирующая метанейронная сеть имеет планарный профиль, высокую пространственную плотность метанейронов и субволновую толщину каждого метанейронного слоя, что особенно важно для акустических волн, которые обычно имеют макроскопическую длину волны. Что еще более важно, мы экспериментально демонстрируем компактную пассивную нейронную сеть на основе метаматериалов, способную непосредственно распознавать сложные объекты в реальном трехмерном пространстве полностью пассивным, в режиме реального времени, без сканирования датчиков и постобработки, как будет продемонстрировано ниже. Результаты Теория мета-нейронной сети. На рисунке 1 схематически показан предлагаемый нами механизм построения акустической метанейронной сети, содержащей несколько параллельных слоев субволновых метанейронов для пассивного распознавания и распознавания в реальном времени и классификации объектов по геометрической форме. Обследуемый объект обычно освещается плоской монохроматической волной, а метанейронная сеть расположена на передающей стороне и принимает рассеянную акустическую волну, создаваемую объектом. Ключевая роль мета-нейронной сети состоит в том, чтобы взаимодействовать с падающей волной после того, как она отражается от объекта, и тем самым сводит акустическую энергию, которая в ее отсутствие рассеивается во всех различных направлениях, к желаемой области на плоскость обнаружения за последним слоем, как показано на рис. 1a. Чтобы объяснить критерий распознавания метанейронной сети, мы приводим пример плоскости обнаружения для типичного случая, когда 10 рукописных цифр от 0 до 9 выбираются в качестве объекта для распознавания. Плоскость обнаружения включает 10 одинаковых квадратных областей, назначенных соответственно этим 10 объектам. Для конкретного объекта только тогда, когда выходной сигнал, в конечном итоге выданный метанейронной сетью, точно перераспределяется на плоскости обнаружения, так что общая интенсивность в ожидаемой области, соответствующей этой цифре, выше, чем в остальных областях, может распознавание и классификации можно считать успешной. Для лучшего имитации реальных приложений здесь мы не транслируем напрямую механизм распознавания изображений для видимого света в акустику, просто используя изображение цифр в качестве входного шаблона или векторизуя входные изображения для облегчения 2D-приложений на кристалле и вместо этого попытайтесь реализовать распознавание объекта в реальном времени с высокой точностью, соответ-

ствующим образом анализируя его рассеянное волновое поле. Сначала рассмотрим распространение рассеянной волны в такой многослойной метаматериальной системе. В качестве основного строительного блока нашей спроектированной мета-нейронной сети каждый метанейрон модулирует амплитуду и фазу падающей волны, затем исходящая волна на переданной стороне служит вторыми источниками и становится входным сигналом для следующего слоя, как регулируется принципом Гюйгенса 40. Очевидно, диаграмма направленности каждого метанейрона зависит от размера элементарной ячейки и расстояния, связанного с длиной волны. Когда каждый метанейрон можно аппроксимировать как источник монополя, соотношение между волновыми полями на двух соседних слоях в нашей метанейронной сети можно записать как Plþ1 ¼ Gl ÁðPl WlÞ; ð1Þ где вектор Pl + 1 обозначает входную волну (l + 1) -го слоя метанейронов, Gl – матрица распространения волны (см. дополнительные примечания 1 и 2), Wl ¼ tlexpðjφlÞ – модуляция, вносимая мета-нейроны на l-м слое, где tl и φl относятся к амплитудной и фазовой модуляции соответственно, «» обозначает поэлементное умножение. В то время как обычная нейронная сеть может быть записана как Ylþ1 ¼ f ðwl Á Yl þ BlÞ; ð2Þ где f – нелинейная функция активации, wl – вес, а Bl – смещение. Сравнение уравнений. Пункты (1) и (2) ясно показывают эквивалентность и различия между мета-нейронной сетью и обычной нейронной сетью. В отличие от характеристики веса в качестве обучаемых параметров в обычных нейронных сетях, функция распространения волны фиксируется после изготовления метанейронной сети, которая определяет расстояние по оси между соседними слоями. Это говорит о том, что функция распространения волны, который формирует связи между соседними слоями, больше похож на гиперпараметр, чем на обучаемый параметр, и нет необходимости оптимизировать расстояние между осями во время процесса обучения при проектировании метанейронной сети.

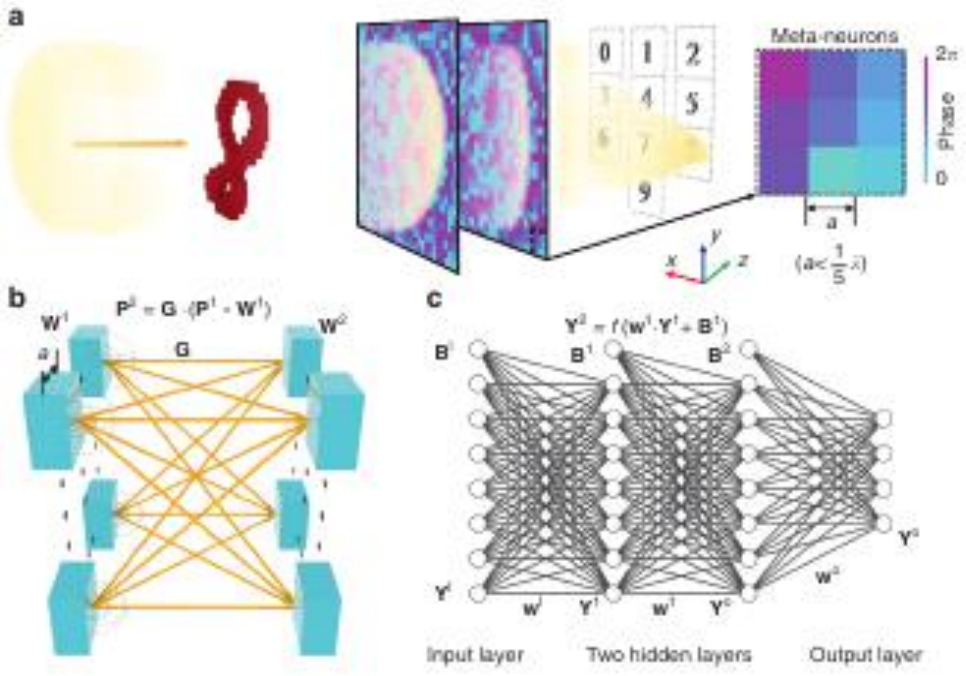


Рис. 1. Распознавание пассивных объектов акустической метанейронной сетью. a Предлагаемая мета-нейронная сеть с параметрами сети, задаваемыми в процессе автоматизированного обучения, способна преобразовывать рассеянную энергию от объекта (выбранную здесь как рукописную цифру «8») в соответствующую область на плоскости обнаружения (отмеченную значком пунктирные рамки за последним слоем). b Схематично иллюстрирует взаимодействие между двумя соседними 2D-слоями метанейронов, чей размер в глубине субволны физически обеспечивает распространение волн от каждого метанейрона 1-го слоя ко всем метанейронам 2-го (после прохождения фазово-амплитудной модуляции с помощью 1-й слой и дифракция в свободном пространстве между ними, описанные W1 и G соответственно). c Обычная нейронная сеть может быть точно воспроизведена с помощью практической физической модели, показанной на b, даже для компактного устройства и / или сложного объекта.

Функция распространения волн также предотвращает вырождение многослойной метанейронной сети в однослойную метанейронную сеть в физических системах вместо того, чтобы просто формировать соединение между соседними уровнями (подробности см. В дополнительных примечаниях 3). В обычной нейронной сети «веса» представляют собой силу соединения между двумя нейронами в соседних слоях, а входные данные последнего слоя определяются выходными значениями первого слоя и «весами» между ними. При настройке весов потери на выходе непрерывно уменьшаются, и, наконец, нейронная сеть сможет выполнять определенные задачи. Точно так же обучаемые параметры в нашей метанейронной сети – это фазовая модуляция, обеспечиваемая метанейронами. Вход метанейронов во втором слое – это интерференция исходящей волны, испускаемой целыми нейронами в первом слое. А регулировка фазовой модуляции перераспределяет энергию волны в выходной

плоскости, что приводит к непрерывному уменьшению потерь и функциональности результирующей метанейронной сети для выполнения задач таким же образом, как и в обычной нейронной сети. (Подробности см. В дополнительном примечании 2). Однако очевидно, что такая эквивалентность математической модели и практической физической системы требует эффективной связи между каждым метанейроном и всеми метанейронами на соседнем слое, что было бы затруднительно для громоздких дифракционных компонентов, непрерывно модулирующих фазу, когда система имеет компактные размеры или предмет имеет сложный узор. Напротив, уникальная способность метанейронов к метаматериалам предлагать произвольный и резкий фазовый сдвиг (41–46) подтверждает монопольное приближение, требуемое формулой. (1), который является стержнем физической аналогии стандартной нейронной сети (подробности см. В дополнительном примечании 1). Учитывая, что потери передачи метанейронов тривиальны, фазовая модуляция, по существу, играет ту же роль, что и вес в традиционной глубинной нейронной сети, и поэтому мы выбираем фазовые сдвиги метанейронов в качестве обучаемых параметров для обучения как будет показано позже. Обратите внимание, что предлагаемая стратегия не требует ни измерения исходного рассеянного поля, ни реконструкции точного акустического изображения, что освобождает от бремени затрат и времени в традиционных парадигмах компьютерного глубокого обучения, которые будут еще больше увеличиваться, когда сложность объекта увеличивается или область обнаружения увеличена. Ограниченное нынешней технологией, это приведет к возникновению множества проблем, включая реализацию крупномасштабных фазированных решеток47, изготовление субволнового датчика (например, пьезоэлектрического преобразователя) и ускорение измерений и анализа огромного количества данных звукового поля. Напротив, мета-нейронная сеть выполняет обнаружение и вычисление одновременно из-за параллельного взаимодействия между волной и метанейронами без сканирования датчиков или постобработки, которая выполняется после прохождения падающей волны, независимо от разрешения или количества метанейронов., а выходное поле нужно измерять только на принимающей стороне с фиксированным числом датчиков (например, рис. 1a), как можно меньшим количеством возможных типов классификации объектов, независимо от того, насколько сложна цель.

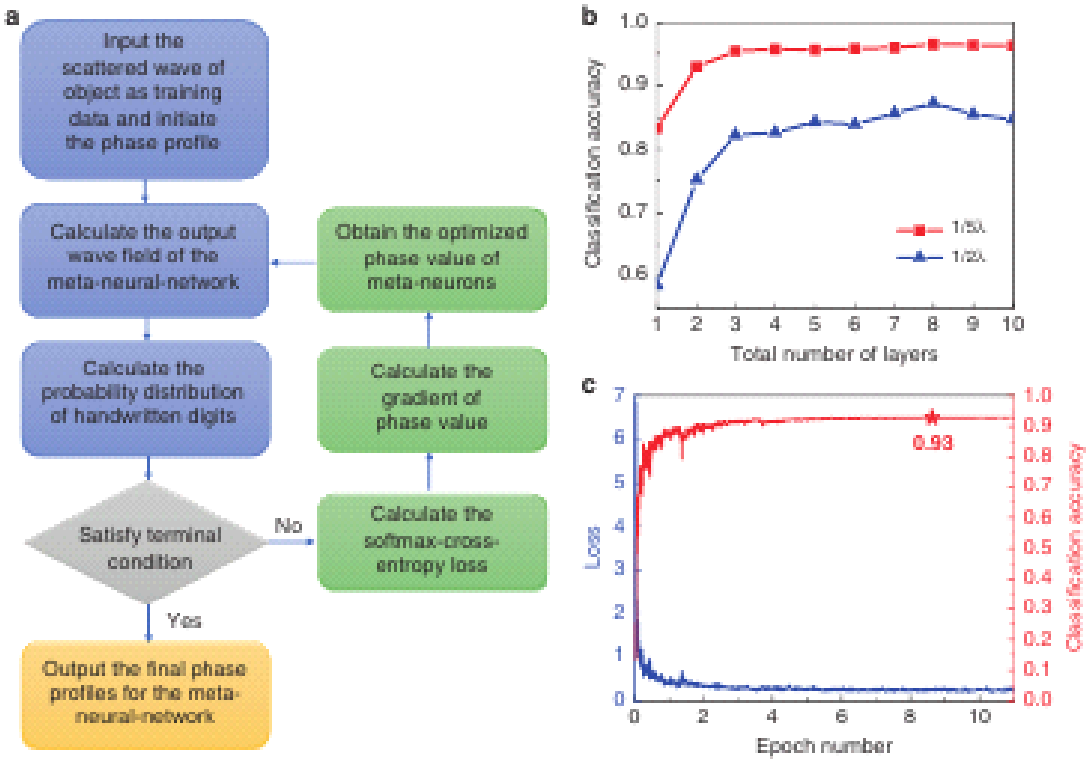


Рис. 2 Результаты моделирования для мета-нейронной сети. a График потока процесса обучения, который использует рассеянную волну, создаваемую различными объектами в качестве обучающих данных, и вычисляет потерю метанейронной работы для итеративной настройки значения фазы каждого метанейрона, пока не будет достигнута максимальная вероятность схождения рассеянная энергия, произведенная определенным классом объектов, в заранее заданную область. b Показывает сравнение точности моделируемой классификации как функции общего количества слоев для метанейронных сетей с различным размером метанейрона. c Изображает смоделированную зависимость величины потерь и точности классификации от номера эпохи, показывая, что точность возрастает с номером эпохи и в конечном итоге достигает максимума (93%) в процессе обучения нашей спроектированной метанейронной сети

В дополнение к этим преимуществам пассивных элементов с точки зрения скорости и простоты, предлагаемая нами метанейронная сеть с компактной плоской геометрией и ультратонким фазовым разрешением позволяет уменьшить размеры устройства до масштабов, недостижимых с помощью дифракционных компонентов, и распознавать объекты, слишком сложные для дифракции. нейронные сети, как мы продемонстрируем ниже (подробности см. в дополнительном примечании 4). Экспериментальная реализация классификации рукописных цифр. Чтобы продемонстрировать уникальные преимущества предложенных нами метанейронных сетей с точки зрения компактности и эффективности, мы сначала выберем продемонстрировать с помощью моделирования и экспериментов распознавание рукописных цифр MNIST (Модифицированный национальный институт

стандартов и технологий) в масштабе, примерно на порядок меньшем, чем достижимый с помощью дифракционных слоев на основе глубокого обучения. База данных содержит 55000 обучающих изображений, 5000 проверочных изображений и 10000 тестовых изображений. Для упрощения проектирования и изготовления образца метанейронной сети в следующих экспериментах мы избегаем одновременной регулировки амплитуды и фазы для передаваемой волны и используем только фазовую модуляцию с коэффициентом передачи, равным 1, что не оказывает заметного влияния на точность полученного устройства, которую мы демонстрируем с помощью численного моделирования (см. дополнительные примечания 5 и 6). Каждый объект реализуется на основе двоичного изображения, сформированного путем округления значения шкалы серого каждого пикселя в соответствующем изображении MNIST в большую сторону (см. Дополнительное примечание 7). Детали тренировочного процесса показаны на рис. 2a. Вводится функция потерь softmax-cross-энтропии 48, которая обычно используется в задачах классификации (см. Подробное обсуждение в дополнительном примечании 2), и градиент значения фазы вычисляется с помощью алгоритма обратного распространения ошибки49. Мы корректируем значения фазы метанейронов в поисках минимального значения потерь, соответствующего максимальной вероятности того, что общая акустическая интенсивность в целевой области будет выше, чем другие, для максимально возможного количества цифр в базе данных MNIST. За счет итеративной подачи обучающих данных точность классификации тестовых данных продолжает расти и в конечном итоге становится стабильной в течение 6 периодов. В нашем моделировании рабочая частота установлена равной 3 кГц (соответствует длине волны ~ 11,4 см в воздухе), так что экспериментальный образец мета-нейронной сети имеет умеренный размер, что облегчает как 3D-печать, так и изготовление субволновых – длина мета-нейронов и измерение звукового поля в безэховой камере. В соответствии с конкретным дизайном, каждый слой выбран так, чтобы он состоял из 28 × 28 (всего 784) метанейронов, что равно количеству пикселей в изображении рукописных цифр в базе данных MNIST. Предполагается, что каждый отдельный мета-нейрон имеет субволновой размер в каждом измерении, соответствующий фактическому размеру практического метаматериала, который мы реализуем при измерении. В частности, поперечный размер метанейрона составляет 2 см (менее 1/5 длины волны), что помогает обеспечить глубинное субволновое разрешение мета-нейронной сети, что жизненно важно для высокоточного распознавания в более сложных случаях. Осевое расстояние между двумя соседними слоями составляет 17,5 см. После обучения дизайн нашего цифрового классификатора метанейронов численно тестируется с помощью 10 000 изображений из набора тестовых данных MNIST. Здесь мы выбираем дизайн метаней-

ронной сети, состоящей из двух слоев метаматериала, только для баланса между точностью и эффективностью классификации, основываясь на нашем численном анализе зависимости точности от номера слоя, как показано на рис. 2b, который указывает, что скорость увеличения точности по отношению к номеру слоя становится намного медленнее для конструкций, содержащих более двух слоев. Точность распознавания такой простой двухслойной структурой может достигать 93%, что является довольно высоким показателем с учетом значительного ускорения процесса обучения, уменьшения количества метанейронов и уменьшения масштаба полученного устройства, и может быть дополнительно улучшено за счет увеличения общее количество мета-нейронов и повышение точности изготовления элементарных ячеек, как следует из наблюдения рис. 2b. Для сравнения мы также вычисляем точность распознавания, когда каждый базовый строительный блок становится шириной в половину длины волны, а расстояние между слоями выбирается таким образом, чтобы эквивалентность в формуле. (1) содержит численные результаты на рис. 2b, которые ясно показывают, что увеличение размера единицы приводит к заметному ухудшению производительности метанейронной сети.

Затем мы проводим экспериментальные измерения, чтобы проверить предложенный нами механизм. В качестве практической реализации в текущем исследовании мы предлагаем разработать элементарную ячейку из метаматериала, состоящую из четырех локальных резонаторов и прямой трубы50, как показано на дополнительном рис. 6. Такая конкретная конструкция позволяет свободно контролировать фазу распространения в полном объеме. Диапазон от 0 до 2π при сохранении высокой эффективности передачи за счет настройки одного структурного параметра h, как показано на дополнительном рисунке 6. Следовательно, слой метанейронов имеет планарный профиль, толщину субволновой длины и, в частности, хорошее фазовое разрешение (~ 1/5 длины волны) имеет решающее значение для обеспечения эквивалентности между стандартом и нашей нейронной сетью на основе метаматериалов (подробности см. В дополнительных примечаниях 1, 2 и 5). На основе зависимости фазового сдвига от параметра, полученной с помощью численного моделирования, мы определили точный геометрический параметр для каждого метанейрона и изготовили метанейронную сеть, состоящую из двух слоев с поперечным размером 56 × 56 см2. С нашей разработанной мета-нейронной сетью рукописные цифры в наборе тестовых данных были хорошо классифицированы, что соответствует соответствующему перераспределению акустической энергии в целевой области, как показано на рис. 3a, b. В ходе эксперимента мы изготовили 2 набора стальных пластин с формами рукописных цифр (а именно, всего 20 объектов, и результат моделирования показан на рис. 3c), которые были выбраны из тестовых изображений, которые были численно

подтверждены способностью быть правильно классифицированным нашей разработанной метанейронной сетью, где каждый метанейрон наделен идеальным значением фазы, заданным в процессе обучения с помощью компьютера. Наблюдается хорошее согласие между теоретическими и экспериментальными результатами, как показано на рис. 3d, на котором в качестве примера показана цифра «8» (более подробная информация и результаты приведены в дополнительном примечании 8), причем оба показывают, что разработанная нами двухслойная метанейронная система. сеть точно перераспределяет входную энергию в область обнаружения, назначенную объекту, за исключением плохой работы мета-нейронной сети при распознавании цифры «4», что в первую очередь связано с экспериментальной ошибкой (см. рис. 3e и дополнительное примечание. 9). Распознавание мультиплексированных лучей OAM. Для дальнейшей демонстрации потенциала нашей мета-нейронной сети для распознавания очень сложных объектов в режиме реального времени с компактными размерами, мы демонстрируем отличительный пример, в котором необходимо точно различать различные пространственные структуры волнового поля, которые кодируются информацией и гораздо более сложный, чем разбросанные узоры, создаваемые простыми объектами в форме цифр.

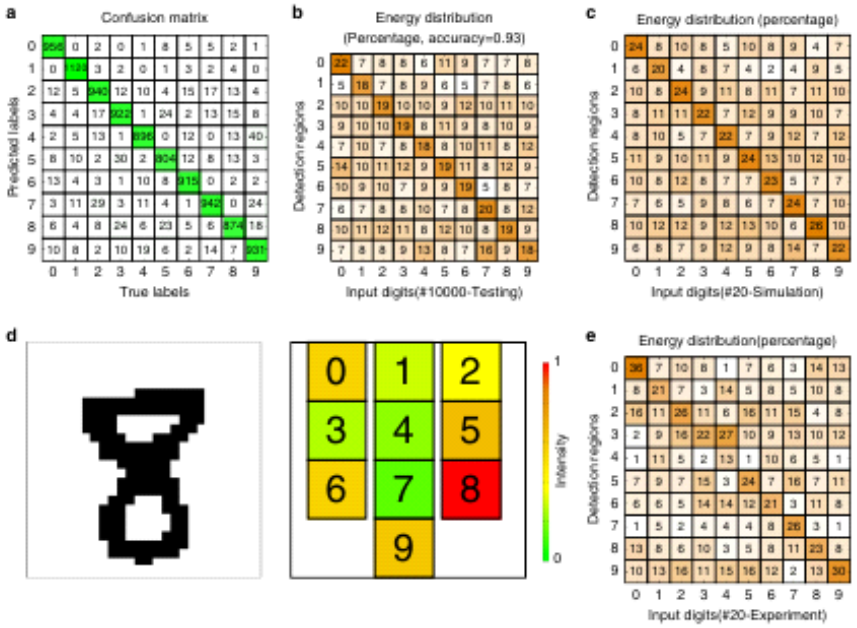


Рис. 3 Экспериментальная проверка акустической метанейронной сети. a Показывает матрицу неточностей для численных результатов двухуровневой мета-нейронной сети с 10 000 рукописных цифр. b Процент распределения энергии 10000 рукописных цифр. c Показывает распределение энергии 20 выбранных цифр при моделировании. d Общая акустическая интенсивность, измеренная в каждой зоне обнаружения, соответствующей цифрам «8». e То же, что и c, но для экспериментов.

В качестве типичного примера, введение орбитального углового момента (OAM) открывает новую степень свободы для кодирования ин-

формации и значительно улучшает способность волн как носителей информации, что имеет решающее значение, особенно для акустических волн, которые доминируют в подводной среде. связи, но по своей природе не имеют вращения53–55. Такой механизм пространственного мультиплексирования использует несколько скрученных лучей с разными топологическими зарядами (ТС) для переноса мультиплексированной информации, которая, однако, должна быть точно считана из сложной пространственной структуры этого синтезированного луча.

Но существующие стратегии, основанные на ортогональности ОАМ для пассивного декодирования, страдают от неконтролируемого пространственного расположения различных выходных лучей и, в частности, строгого выравнивания между лучом и приемным устройством, что жизненно важно для точности декодирования, но является проблемой на практике. Здесь мы предлагаем преодолеть эти фундаментальные ограничения на основе принципиально другого механизма, используя акустическую метанейронную сеть, обученную распознаватьсложные пространственные паттерны, связанные с разными порядками ОАМ. Что еще более важно, путем прямого обучения метаневронной сети как с центрированными, так и с нецентрированными лучами ОАМ, система способна распознавать пространственный образец каждого порядка ОАМ независимо от того, полностью ли перекрываются центры луча и устройства. Четырехслойная мета-нейронная сеть, содержащая 101 × 101 × 4 (всего 40 804) мета-нейронов, предназначена для распознавания максимальной комбинации из 8 порядков ОАМ (± 1, ± 2, ± 3, ± 4, 255 комбинаций в сумме). В текущем проекте мы продемонстрируем реализацию метанейронной сети, способной распознавать несколько лучей ОАМ, центры которых поперечно смещены в произвольных направлениях на максимальное расстояние 6λ, которое достигает 1/3 длины стороны каждого метанейрона. слой и будет довольно сложной задачей для существующих механизмов, использующих устройства одинакового размера. Диапазон значений r и θ составляет [0,6λ] и [0,2π) соответственно, причем (r, θ) является местоположением центра вихря относительно полярной координаты. На рисунке 4a схематично показано, как разработанная метанейронная сеть реализует точное распознавание каждого луча ОАМ в реальном времени посредством тщательно продуманного перераспределения падающей энергии на плоскости обнаружения (которое иллюстрирует распознавание лучей ОАМ, состоящих из +3 и 4 порядков с рассогласование (6λ, 0) в качестве примера). Теперь плоскость обнаружения разделена на 8 областей, каждая из которых содержит две области (отмеченные буквами «Y» и «N», соответствующие наличию и отсутствию определенного состояния ОАМ соответственно), как показано на рис. 4a (подробнее в дополнительном примечании 10).

Распределение интенсивности звука в плоскости обнаружения также показано на рис. 4a, что ясно указывает на то, что звуковая энергия перераспределяется в правильную область (более подробная информация в дополнительном примечании 10). На рис. 4б показана зависимость точности распознавания от расстояния и направления рассогласования (а именно, параметров r и θ).

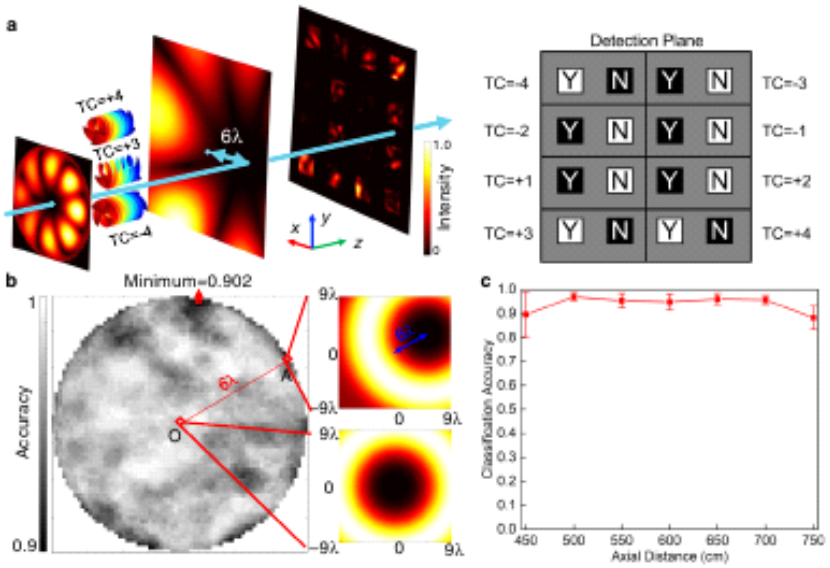


Рис. 4 Распознавание несовмещенных состояний OAM. a Показывает распознавание мультиплексированного луча OAM (с TCs = +3, ± 4 и расстоянием рассогласования 6λ в качестве примера) разработанной метанейронной сетью, которая перераспределяет падающую энергию в плоскости обнаружения таким образом, чтобы можно ли однозначно пометить каждое состояние OAM. b Изображает зависимость точности распознавания от расстояния и направления перекоса. Вставки: пространственные шаблоны несовмещенных (вверху) и выровненных (внизу) лучей OAM с одинаковым порядком OAM. c Показывает смоделированную точность распознавания как функцию от осевого расстояния, а полоса ошибок указывает стандартное отклонение ± 1 от среднего значения точности.

Существенное рассогласование можно наблюдать при сравнении пространственных структур, изображенных на вставках, для выровненного и смещенного луча OAM того же порядка. Мы рассчитали точность распознавания для всех возможных 255 комбинаций среди 8 порядков состояний OAM при различных (r, θ) и построили результаты на рис. 4b, который ясно показывает, что наш механизм эффективен даже на расстоянии между центрами OAM. пучки и мета-нейронный слой достигает 6λ. В процессе обучения мы также приняли во внимание параметр расстояния распространения лучей OAM, пытаясь также расширить возможности разработанной метанейронной сети с высокой устойчивостью к смещению устройства обнаружения по направлению распространения, которое могло бы иметь большое значение для практического

применения связи на основе OAM. Смоделированная точность распознавания как функция расстояния по оси, изображенная на рис. 4c, показывает высокую точность нашей метанейронной сети, сохраняющейся в широком диапазоне расстояний распространения (от 500 см до 700 см, почти 18 λ). В результате такого отличительного механизма мы реализуем в реальном времени и пассивное распознавание каждого взаимно ортогонального состояния OAM с помощью метанейронной сети, которая имеет контролируемые области вывода и высокую устойчивость к рассогласованию как в осевом, так и в поперечном направлениях, что помогает решить давнишние вопросы в высокопроизводительной связи на основе OAM и будут иметь далеко идущие последствия в соответствующих областях, выступая в качестве интеллектуального преобразователя с потенциалом, который может быть расширен для распознавания более сложных объектов при достаточно большой обучающей базе данных и, соответственно, переработанные мета-нейроны, например, для диагностики опухолей при ультразвуковой визуализации или выявления дефектов в промышленных тестах.

Обсуждение

Для наглядной демонстрации физической модели и облегчения практической реализации мы демонстрируем только значительно сокращенную модель мета-нейронной сети с несколькими основными упрощениями, которые, однако, не повлияют на общность предлагаемого нами механизма. В частности, целостная характеристика x y z 6λ 1,0 0 9λ 0 –9λ 0 9λ O A Минимум = 0,902 0,9 1 Точность 9λ –9λ 0 0 9λ a b c Интенсивность Рис. 4 Распознавание смещенных состояний OAM. a Показывает распознавание мультиплексированного луча OAM (c TCs = +3, ± 4 и расстоянием рассогласования 6λ в качестве примера) разработанной метанейронной сетью, которая перераспределяет падающую энергию в плоскости обнаружения таким образом, чтобы можно ли однозначно пометить каждое состояние OAM. b Изображает зависимость точности распознавания от расстояния и направления перекоса. Вставки: пространственные шаблоны несовмещенных (вверху) и выровненных (внизу) лучей OAM с одинаковым порядком OAM. c Показывает смоделированную точность распознавания как функцию осевого расстояния, а полоса ошибок указывает стандартное отклонение ± 1 от среднего значения точности.

Текущая мета-нейронная сеть может быть дополнительно улучшена путем изменения конструкции и обучения метанейронов. Например, можно легко повысить его компактность и эффективность, заменив используемую здесь простую элементарную ячейку из метаматериала некоторыми недавно появившимися конструкциями, такими как метамате-

риал полого типа с длиной волны менее 1/600 46, и позволяет программировать метанейронную сеть с помощью реконфигурируемых метанейронов. Наша схема также применима к более реалистичным приложениям, таким как ультразвуковая визуализация, с использованием метаматериалов на водной основе, таких как мягкие градиентно-пористые среды58, и включения неплоской падающей волны и неоднородной среды в процесс обучения. В заключение мы демонстрируем теоретический дизайн и экспериментальную реализацию пассивной нейронной сети на основе метаматериалов в акустике, выполняющей различные сложные задачи распознавания объектов, такие как распознавание рукописных цифр и смещенных лучей OAM. Помимо отсутствия зависимости от специалистов-людей, как в компьютерных методах глубокого обучения, наша предлагаемая метанейронная сеть не требует сложных массивов датчиков или дорогостоящих компьютеров, и, в частности, выполняет распознавание в реальном времени без источника питания, спасибо его пассивному характеру и параллельному взаимодействию волн, освобождающему от тяжелой нагрузки на вычислительное оборудование в традиционных методах глубокого обучения. Кроме того, мета-нейронные сети имеют небольшую площадь основания благодаря субволновой природе метаматериалов, что имеет жизненно важное значение для их применения в акустике, где акустические волны обычно имеют макроскопическую длину волны, но недостижимы с помощью нейронных сетей на основе дифракционных компонентов. Наш дизайн с простотой, компактностью и эффективностью предлагает возможность миниатюризации и интеграции устройств глубокого обучения и может даже открыть путь к разработке концептуальных акустических устройств нового поколения, таких как портативные и интеллектуальные преобразователи, которые в результате соединения функции обнаружения и вычисления, могут быть в состоянии автоматически анализировать акустические сигналы обратного рассеяния, которые он получает, и впоследствии выполнять сложные задачи, такие как оценка опухолей полностью пассивным образом, без сканирования сенсоров и постобработки. Кроме того, разработанное нами устройство служит новым классом пассивных микросхем глубокого обучения для решения задач в режиме реального времени, не требующего питания, с возможностью вдохновлять соответствующие исследования для других классических волн. Методы. Наша акустическая метанейронная сеть была смоделирована с помощью MATLAB и обучена на настольном компьютере с графическим процессором (GPU) GeForce RTX 2070, процессором Intel (R) Xeon (R) E5-2620 v3 @ 2,40 ГГц и 160 ГБ оперативной памяти. под управлением операционной системы Windows 7 (Microsoft). В эксперименте входной звук генерировался динамиком (Beyma CP380), управляемым генератором сиг-

налов (RIGOL DG1022). Датчик, который мы использовали на плоскости обнаружения, представлял собой 1/4-дюймовый микрофон со свободным полем (BRÜEL & KJÆR, тип 4961) и автономный самописец (BRÜEL & KJÆR, тип 3160-A-022). Эксперименты проводятся в безэховой комнате.

Доступность данных Данные, подтверждающие выводы этого исследования, доступны в документе и в дополнительной информации. Дополнительные данные, относящиеся к этой статье, можно получить у соответствующих авторов по разумному запросу. Исходные данные предоставлены вместе с этой статьей. Доступность кода. Код, подтверждающий выводы этого исследования, можно получить у соответствующего автора по разумному запросу. Поступило: 31 января 2020 г.; Принята в печать: 16 октября 2020 г.;

References

1. Moore, C. L. & Copel, J. A. Point-of-care ultrasonography. N. Engl. J. Med. 364, 749–757 (2011).
2. Guo, X., Zhang, D. & Zhang, J. Detection of fatigue-induced micro-cracks in a pipe by using time-reversed nonlinear guided waves: a three-dimensional model study. Ultrasonics 52, 912–919 (2012).
3. Azimi-Sadjadi, M. R., Yao, D., Huang, Q. & Dobeck, G. J. Underwater target classification using wavelet packets and neural networks. IEEE Trans. Neural Netw. 11, 784–794 (2000).
4. Chen, H. et al. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J. Biomed. health Inform. 19, 1627–1636 (2015).
5. Milletari, F. et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. Comput. Vis. Image Underst. 164, 92–102 (2017).
6. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248 (2017).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436 (2015).
8. Litjens, G. et al. A survey on deep learning in medical image analysis. Med.image Anal. 42, 60–88 (2017).
9. Ting, D. S. W. et al. AI for medical imaging goes deep. Nat. Med. 24, 539 (2018).
10. Mikolov, T. et al. Strategies for training large scale neural network language models[C]. in Proc. 2011 IEEE Workshop on Automatic Speech Recognition &Understanding, 196–201 (IEEE, 2011).
11. Hinton, G. et al. Deep neural networks for acoustic modeling in speechrecognition: The shared views of four research groups. IEEE Signal Process. Mag. 29, 82–97 (2012).

12. Sainath, T. N., et al. Deep convolutional neural networks for LVCSR[C]. in Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 8614–8618 (IEEE, 2013).
13. Schmidt, E. M. & Kim, Y. E. Learning emotion-based acoustic features with deep belief networks[C]. in Proc. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 65–68 (IEEE, 2011).
14. Mao, Q., Dong, M., Huang, Z. & Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimed. 16, 2203–2213 (2014).
15. Han, K., Yu, D. & Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine[C]. in Proc. Interspeech 2014-Fifteenth annual conference of the international speech communication association, 223–227 (ISCA, 2014).
16. Fayek, H. M., Lech, M. & Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 92, 60–68 (2017).
17. Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals. Solid State Commun. 102, 165–173 (1997).
18. Sánchez-Pérez, J. V. et al. Sound attenuation by a two-dimensional array of rigid cylinders. Phys. Rev. Lett. 80, 5325–5328 (1998).
19. Yang, S. et al. Focusing of sound in a 3D phononic crystal. Phys. Rev. Lett. 93, 024301 (2004).
20. Yang, Z. et al. Topological acoustics. Phys. Rev. Lett. 114, 114301 (2015).
21. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. Nat. Photonics 11, 441 (2017).
22. Ding, Y. et al. Experimental demonstration of acoustic chern insulators. Phys. Rev. Lett. 122, 014302 (2019).
23. Liu, Z. et al. Locally resonant sonic materials. Science 289, 1734 (2000).
24. Fang, N. et al. Ultrasonic metamaterials with negative modulus. Nat. Mater. 5, 452–456 (2006).
25. Pendry, J. B. Negative refraction makes a perfect lens. Phys. Rev. Lett. 85, 3966–3969 (2000).
26. Li, Y. et al. Experimental realization of full control of reflected waves with subwavelength acoustic metasurfaces. Phys. Rev. Appl. 2, 064002 (2014).
27. Chen, H., Wu, B.-I., Zhang, B. & Kong, J. A. Electromagnetic wave interactions with a metamaterial cloak. Phys. Rev. Lett. 99, 063903 (2007).
28. Zhu, X., Liang, B., Kan, W., Zou, X. & Cheng, J. Acoustic cloaking by a superlens with single-negative materials. Phys. Rev. Lett. 106, 014301 (2011).
29. Liang, B., Yuan, B. & Cheng, J.-C. Acoustic diode: rectification of acoustic energy flux in one-dimensional systems. Phys. Rev. Lett. 103, 104301 (2009).
30. Liang, B., Guo, X. S., Tu, J., Zhang, D. & Cheng, J. C. An acoustic rectifier. Nat. Mater. 9, 989–992 (2010).
31. Li, L. et al. Machine-learning reprogrammable metasurface imager. Nat. Commun. 10, 1082 (2019).
32. Freitas, G. M. F., Rego, S. L. & Vasconcelos, C. F. L. Design of metamaterials using artificial neural networks[C]. in Proc. 2011 SBMO/IEEE MTT-S Internation-

al Microwave and Optoelectronics Conference (IMOC 2011), 541–545 (IEEE, 2011).
33. Liu, Z., Zhu, D., Rodrigues, S. P., Lee, K.-T. & Cai, W. Generative model for the inverse design of metasurfaces. Nano Lett. 18, 6570–6576 (2018).
34. Ma, W., Cheng, F. & Liu, Y. Deep-learning-enabled on-demand design of chiral metamaterials. ACS Nano 12, 6326–6334 (2018).
35. Malkiel, I. et al. Plasmonic nanostructure design and characterization via Deep Learning. Light.: Sci. Appl. 7, 60 (2018).
36. Lin, X. et al. All-optical machine learning using diffractive deep neuralnetworks. Science 361, 1004 (2018).
37. Mellin, S. D. & Nordin, G. P. Limits of scalar diffraction theory and an iterative angular spectrum algorithm for finite aperture diffractive opticalelement design. Opt. Express 8, 705–722 (2001).
38. Wu, Z., Zhou, M., Khoram, E., Liu, B. & Yu, Z. Neuromorphic metasurface. Photonics Res. 8, 46–50 (2020).
39. Khoram, E. et al. Nanophotonic media for artificial neural inference. Photonics Res. 7, 823–827 (2019).
40. Born, M. A. X. & Wolf, E. Principles of optics: electromagnetic theory of propagation, interference and diffraction of light[M]. (Elsevier, 2013).
41. Yu, N. et al. Light propagation with phase discontinuities: generalized laws of reflection and refraction. Science 334, 333 (2011).
42. Li, Y., Liang, B., Zou, X.-Y. & Cheng, J.-C. Extraordinary acoustic transmission through ultrathin acoustic metamaterials by coiling up space. Appl. Phys. Lett. 103, 063509 (2013).
43. Zhu, X. et al. Implementation of dispersion-free slow acoustic wave propagation and phase engineering with helical-structured metamaterials. Nat. Commun. 7, 11731 (2016).
44. Zhu, Y., Fan, X., Liang, B., Cheng, J. & Jing, Y. Ultrathin acoustic metasurfacebased schroeder diffuser. Phys. Rev. X 7, 021034 (2017).
45. Zhu, Y. et al. Fine manipulation of sound via lossy metamaterials with independent and arbitrary reflection amplitude and phase. Nat. Commun. 9,1632 (2018).
46. Tang, H. et al. Hollow-out patterning ultrathin acoustic metasurfaces for multifunctionalities using soft fiber/rigid bead networks. Adv. Funct. Mater. 28, 1801127 (2018).
47. Jiang, X.-J. et al. A microscale linear phased-array ultrasonic transducer based on PZT ceramics. Sensors. 19, 1244 (2019).
48. Rubinstein, R. Y. & Kroese, D. P. The Cross-Entropy Method. (Springer-Verlag New York, 2013).
49. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT press, 2016).
50. Li, Y., Jiang, X., Liang, B., Cheng, J.-C. & Zhang, L. Metascreen-based acoustic passive phased array. Phys. Rev. Appl. 4, 024003 (2015).
51. Willner, A. E., Wang, J. & Huang, H. A different angle on light communications. Science 337, 655 (2012).

52. Wang, J. et al. Terabit free-space data transmission employing orbital angular momentum multiplexing. Nat. Photonics 6, 488 (2012).
53. Jiang, X., Li, Y., Liang, B., Cheng, J.-C. & Zhang, L. Convert acoustic resonances to orbital angular momentum. Phys. Rev. Lett. 117, 034301(2016).
54. Shi, C., Dubois, M., Wang, Y. & Zhang, X. High-speed acoustic communication by multiplexing orbital angular momentum. Proc. Natl Acad. Sci. USA 114, 7250–7253 (2017).
55. Jiang, X., Liang, B., Cheng, J. C. & Qiu, C. W. Twisted acoustics: metasurface-enabled multiplexing and demultiplexing. Adv. Mater. 30, 1800257 (2018).
56. Xie, G. et al. Performance metrics and design considerations for a free-space optical orbital-angular-momentum–multiplexed communication link. Optica 2, 357–365 (2015).
57. Willner Alan, E. et al. Recent advances in high-capacity free-space optical and radio-frequency communications using orbital angular momentum multiplexing. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 375, 20150439 (2017).
58. Jin, Y., Kumar, R., Poncelet, O., Mondain-Monval, O. & Brunet, T. Flat acoustics with soft gradient-index metasurfaces. Nat. Commun. 10, 143 (2019).

НА ПУТИ К ИНТЕЛЛЕКТУАЛЬНОЙ ФОТОННОЙ СИСТЕМЕ

Цзо В., Ма Б., Сю Ш., Цзо С., Ван С

https://doi.org/10.1007/s11432-020-2863-y

Технологии, представленные глубоким обучением, расширили свое применение в различных областях. Помимо традиционных систем обработки на основе электроники, конвергенция технологий фотоники и искусственного интеллекта (ИИ) повышает производительность и обучаемость ИИ.

В этом обзоре мы предлагаем концепцию интеллектуальной фотонной системы (IPS), иллюстрирующую ее как развивающуюся архитектуру с тремя различными версиями. Для каждой версии IPS мы рассматриваем несколько репрезентативных исследований. Кроме того, мы обсуждаем проблемы, стоящие перед IPS, и даем некоторые перспективы для будущего развития.

Ключевые слова: интеллектуальная фотонная система, глубокое обучение, искусственный интеллект, оптическая нейронная сеть, нейроморфная фотоника

1 Введение

Фотоника способна преодолеть узкие места в электронике благодаря преимуществам широкополосного доступа. ширина, высокая скорость работы, устойчивость к электромагнитным помехам и низкие перекрестные помехи. Типичные приложения варьируются от оптической связи [1], микроволновой фотоники (MWP) [2], оптического изображения [3] до оптического зондирования [4]. По мере развития обычных систем постепенно возникает узкое место в производительности. Например, в системах микроволновой фотоники производительность ухудшается с увеличением сложности. Потому что несовершенство оптоэлектронных устройств вызывает несоответствие между параллельными каналами [5]. Кроме того, обычная фотонная система часто разрабатывается для специфического применения с относительно низкой гибкостью. На ранней стадии фотонные процессоры, которые реализуют переключаемые и программируемые функции, привлекательны для исследователей [6–13]. Теперь фотонные системы требуют не только повышения производительности, но и достаточной гибкости для выполнения интеллектуальных задач. Как было отмечено AlphaGo [14, 15], искусственный интеллект (ИИ), основанный на глубоком обучении и искусственных нейронных сетях (ИНС), переживает бум. ИИ привносит инновации и революции в различные области, такие как медицина [16], взаимодей-

ствие человека с компьютером [17], игры [18], робототехника [19] и автономное вождение [20]. Исследователи в различных областях пытаются использовать ИИ для создания интеллектуальных систем. В настоящее время ИИ внедрен в оптику и фотонику для новых приложений. Системы на базе искусственного интеллекта демонстрируют впечатляющую производительность и надежность по сравнению с традиционными методами оптимизации. Что наиболее важно, ИИ потенциально снижает стоимость и требует ручных усилий обычных систем в практических сценариях.

Как и ожидалось, интеллектуальная фотонная система (IPS), которая подходит для будущих приложений в сложных средах, весьма привлекательна. В этом обзоре мы выдвинули концепцию IPS. Описываются характеристики IPS в различных версиях. Чтобы продемонстрировать возможности различных версий IPS, рассматриваются последние достижения архитектур и компонентов IPS. Кроме того, даются перспективы и выводы для будущего развития IPS.

2 Концепция IPS

С развитием фотоники и интеллектуальной технологии IPS эволюционирует от искусственного интеллекта до нейроморфного интеллекта для универсальности и высокой эффективности. Мы предлагаем иерархию концепции IPS, показанную на рисунке 1, включая IPS на основе AI, IPS с AI с поддержкой фотоники и нейроморфную IPS. IPS на базе искусственного интеллекта призвана объединить искусственный интеллект с существующими фотонными / оптическими системами для интеллектуального анализа и оценки информации. Чтобы полностью раскрыть потенциал ИИ, IPS с ИИ с использованием фотоники фокусируется на вычислительном ускорителе с высокой скоростью и низким энергопотреблением. Нейроморфный IPS, созданный на основе мозга, обладает способностями к адаптации и принятию решений. В этом разделе мы подробно обсудим конфигурации IPS и методы различных версий IPS. На начальном этапе мы пытаемся внедрить технологию искусственного интеллекта, основанную на глубоком обучении, для оптимизации фотонных систем, которые удачно называются IPS на основе искусственного интеллекта. Как показано на рис. 1 (верхняя вставка), IPS состоит из аналогового широкополосного интерфейса и цифрового интеллектуального внутреннего интерфейса. Первый включает в себя лазерный источник, оптический модулятор, процессор оптических сигналов и фотоприемник. Сначала входной радиочастотный (RF) сигнал преобразуется в оптическую область через оптический модулятор. Затем модулированный оптический сигнал обрабатывается процессором оптических сигналов. Различные функции обработки, такие как фильтрация

и преобразование частоты, предоставляются при изучении обработки сигналов MWP [21]. После обработки оптического сигнала сигнал преобразуется в электрическую цифровую область с помощью фотодетектора и аналого-цифрового преобразователя (АЦП). В практических аппаратных реализациях несовершенства внешнего интерфейса создают проблемы накопления шума, нелинейности и рассогласования между параллельными каналами [22]. По мере того, как интерфейс становится более сложным для обработки высокого уровня, отношение сигнал / шум (SNR) сильно ухудшается. Предлагаются методы оптимизации для улучшения отношения сигнал / шум с помощью аппаратного или программного обеспечения [23,24], что будет иметь серьезные ограничения в практических приложениях из-за низкой гибкости и дорогостоящего вспомогательного оборудования. Искусственный интеллект – многообещающий кандидат в качестве гибкой схемы повышения производительности внешнего интерфейса. Глубокое обучение с помощью ИНС позволяет установить сложную взаимосвязь отображения, которая находится между опорным сигналом (идеальным) и выходным сигналом (неидеальным). На схемах изображены полносвязные нейронные сети, рекуррентные нейронные сети и сверточные нейронные сети (CNN), которые являются часто используемыми сетевыми моделями для настройки архитектуры ИНС. После обучения сети выходной сигнал внешнего интерфейса с различной полосой пропускания и амплитудами может быть оптимизирован с небольшими затратами. Таким образом, IPS с питанием от искусственного интеллекта может обрабатывать радиочастотные сигналы с улучшенными характеристиками, чем обычная фотонная система. Обобщены общие процедуры разработки IPS на базе искусственного интеллекта. Во-первых, получите обучающий набор и набор для тестирования. Обучающий набор состоит из выходного сигнала и соответствующего опорного сигнала. Во-вторых, тренируйте сеть. В общем, эффект обучения улучшается с увеличением размера обучающей выборки. Наконец, протестируйте сеть. Если сеть усвоила особенности отображения между выходным сигналом и опорным сигналом, она будет эффективно оптимизировать тестовые сигналы, даже если они представлены в первый раз. Следует отметить, что сетевая архитектура, размер и параметры должны быть разумно определены, чтобы максимизировать потенциал ИИ. Следовательно, необходимо тщательно сопоставлять характеристики внешнего и внутреннего интерфейса.

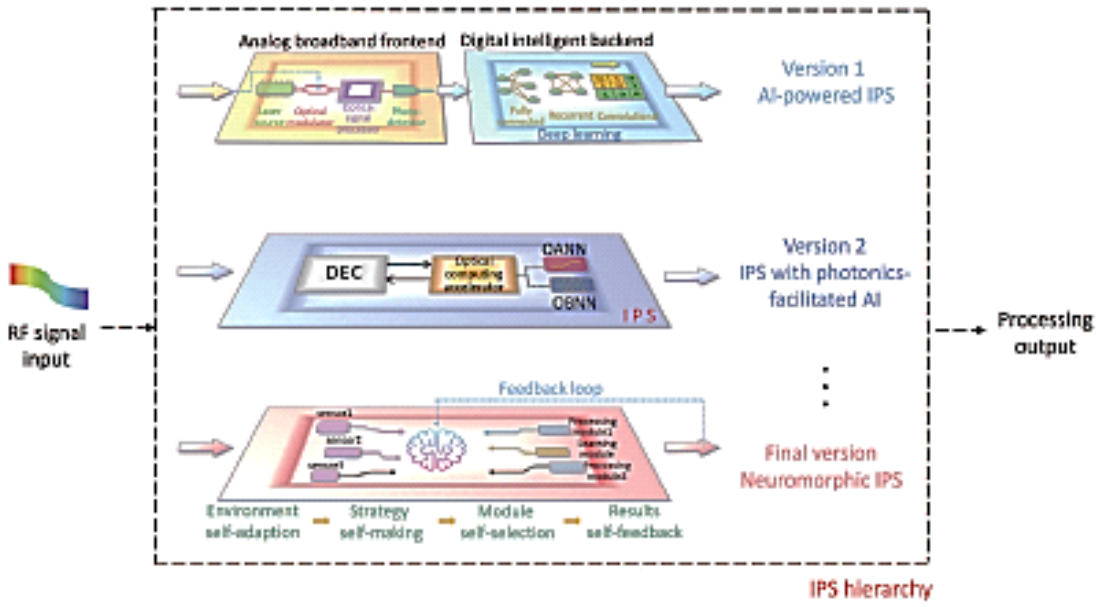


Рисунок 1 (Цветной онлайн) Иерархия концепции IPS

На втором этапе (средняя вставка на рис. 1), IPS с искусственным интеллектом на основе фотоники, мы стремимся избавиться от ограничений, налагаемых традиционной парадигмой электронных вычислений. Ключевыми факторами ИИ на основе глубокого обучения являются данные, алгоритмы и вычислительная мощность. По мере того как масштаб ИНС расширяется для более сложных задач, возрастающие вычислительные операции (например, умножение и накопление, MAC) становятся бременем для нынешнего центрального процессора (ЦП). Во-первых, потребление энергии достигает DEC: цифровая электронная схема. довольно высокий уровень. Расходы на электроэнергию для работы крупномасштабной сети могут составлять тысячи долларов [25]. Во-вторых, скорость вычислений с помощью электронного оборудования предотвращает появление потенциальных приложений, требующих возможности принятия решений в реальном времени. Несмотря на то, что вычислительные блоки, предназначенные для глубокого обучения (например, блоки тензорной обработки, TPU), постепенно заменяют ЦП для работы ИНС, скорость по-прежнему ограничена.

Кроме того, приближается ожидаемый конец закона Мура [26], который требует инноваций в вычислительной среде. Для требовательного к вычислениям искусственного интеллекта фотоника представляет собой захватывающую вычислительную среду благодаря ее высокой скорости и высокой энергоэффективности. Оптическая искусственная нейронная сеть (OANN) является одним из кандидатов на роль ускорителя оптических вычислений. Ожидается, что объединение оптической реализации с архитектурой ИНС позволит OANN значительно повысить

скорость работы и эффективность ИИ. Следовательно, IPS, которая извлекает выгоду из OANN, может применяться в потенциальных сценариях в реальном времени. Помимо полностью подключенной нейронной сети, реализация CNN способствует практической реализации OANN за ее выдающуюся способность извлечения признаков. В то же время необходимая нелинейная функция активации в ИНС может быть реализована с помощью богатых электрооптических или оптических нелинейных эффектов [27], что позволяет создать полную и эффективную OANN. В результате OANN может произвести революцию в вычислительной структуре искусственного интеллекта. Сделаем еще один шаг: оптическая биологическая нейронная сеть (OBNN), которая использует принцип связи и вычислений мозга, является еще одним кандидатом в качестве оптического ускорителя вычислений со сверхнизким энергопотреблением. Мозгу, основанному на биологической нейронной сети, требуется всего 20 Вт мощности для выполнения различных задач обработки [28]. Напротив, ИНС – это модель, управляемая данными, с большим энергопотреблением во время обучения. Жажда мощности IPS будет препятствовать работе некоторых гибких приложений, таких как мобильные сценарии. К счастью, решение, предоставленное биологической нейронной сетью, демонстрирует четкое выражение информации – спайк [29]. Спайк представляет собой возникновение события, которое несет информацию во времени, а не в форме. Исследования в области вычислительной нейробиологии доказывают, что пики – это эффективный и полный формат информации [30]. Благодаря высокоскоростному импульсу и высокой энергоэффективности оптические методы для реализации биологической нейронной сети предлагают возможность создания вычислительной структуры с низким энергопотреблением, как показано на рисунке 1 (средняя вставка). Между тем, по сравнению с электронным аналогом [31–33], компромисс между пропускной способностью и плотностью соединений узла (т.е. нейрона) подчеркивает большую пропускную способность и низкие перекрестные помехи фотоники. Следовательно, ожидается, что OBNN позволит использовать IPS в широком диапазоне приложений обработки. IPS приближается к финальной стадии нейроморфной IPS, которая представляет собой «фотонный мозг», оснащенный мощными функциями восприятия и обработки. Мозг – самая сложная, но интригующая система обработки информации в природе. Функции мозга включают обучение на собственном опыте, адаптацию к окружающей среде, выводы из ситуации и сверхнизкое потребление энергии. Естественно, мозг – это парадигма надежной и интеллектуальной системы обработки. Поэтому многие исследователи пытаются разработать систему с мозговой способностью

принятия решений и способностью к обучению, как интеллектуальная робототехника [34].

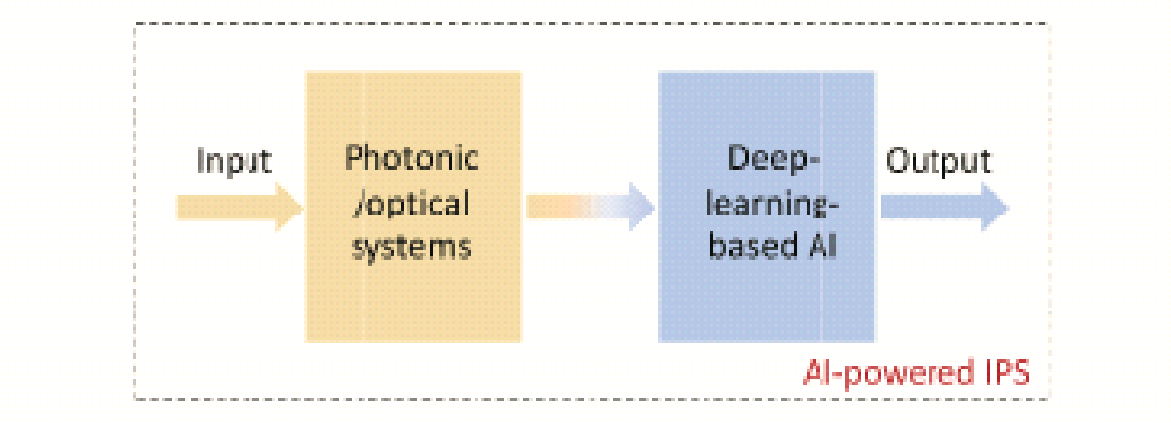


Рисунок 2 (Цветной онлайн) Архитектура IPS на базе искусственного интеллекта.

Характеристики нейроморфной IPS представлены на рисунке 1 (нижняя вставка). Во-первых, нейроморфная IPS демонстрирует приспособляемость к окружающей среде. В разных средах на входной сигнал влияет естественное (например, температура) или искусственное (например, логический сигнал) влияние на входной сигнал, что затрудняет правильное понимание состояния входного сигнала нейроморфной IPS. Выявление смысла состояния сигнала имеет решающее значение для следующей стратегии обработки. Следовательно, должна быть реализована многомерная оценка входного сигнала различными «датчиками». Во-вторых, стратегию обработки данных создает «мозг» нейроморфной IPS. В этой процедуре суть состоит в том, чтобы реализовать цель обработки самым простым способом в соответствии с входным сигналом. В идеале нейроморфная IPS сама устанавливает цель обработки после периода обучения, такого как исправление искажений или маскировка входного сигнала. Основываясь на конкретной цели, нейроморфный IPS «думает» об оптимизированной стратегии обработки входного сигнала. В-третьих, руководствуясь стратегией обработки, нейроморфная IPS выбирает необходимые модули из своей библиотеки обработки. В частности, ключевым моментом является разложение стратегии и объединение модулей в интегрированную функцию. Ожидается, что нейроморфная IPS обеспечит координацию между модулями, а не просто суммирует их. В-четвертых, нейроморфная IPS учится на выходных результатах, из которых создается опыт. Эффективный обучающий модуль распознает ошибки в текущих выходных результатах и корректирует стратегию обработки. Несмотря на то, что нейроморфная IPS на ранней стадии вырабатывает несовершенные стратегии, способность к обучению позволяет себе становиться все более и более разумной по мере накопления опыта. Таким образом, нейроморфная IPS с

возможностью широкополосной обработки, обеспечивающей фотонику, и умной способностью к обучению и принятию решений, подобной мозгу, принесет большую пользу будущей системе обработки информации.

3 Исследования архитектур и компонентов IPS

В последние годы мы стали свидетелями того, что исследователи решают практические проблемы систем с помощью ИИ и разрабатывают оптические ускорители для вычислений ИИ. В соответствии с тремя частями IPS на базе искусственного интеллекта, OANN и OBNN, в этом разделе будут рассмотрены типичные достижения архитектур и компонентов IPS, а также наши усилия.

3.1 IPS на базе искусственного интеллекта

Каноническая архитектура IPS на базе искусственного интеллекта показана на рисунке 2. Возможными решениями для внешнего интерфейса являются различные фотонные / оптические системы, такие как системы оптической визуализации и широкополосные анализаторы сигналов. Специально разработанная ИНС обучена анализировать и обрабатывать исходный вывод внешнего интерфейса, таким образом поддерживая IPS на базе ИИ. Благодаря успеху ИИ в обработке изображений, появляется тенденция сочетать ИИ с оптическими изображениями [35]. В [36] микроскопия глубокого обучения обеспечивает более высокое пространственное разрешение с входом Как показано на рисунке 3, для обучения глубокой нейронной сети (DNN) предоставляются изображения с низким разрешением и соответствующие изображения с высоким разрешением. После оптимизации параметров DNN выходные изображения имеют улучшенное пространственное разрешение, поле обзора и глубину поля, которые сопоставимы с характеристиками объективов с более высокой числовой апертурой. Микроскопия с глубоким обучением не требует дополнительного оборудования или специальной обработки для вывода изображений с высоким разрешением и высокой скоростью отклика.

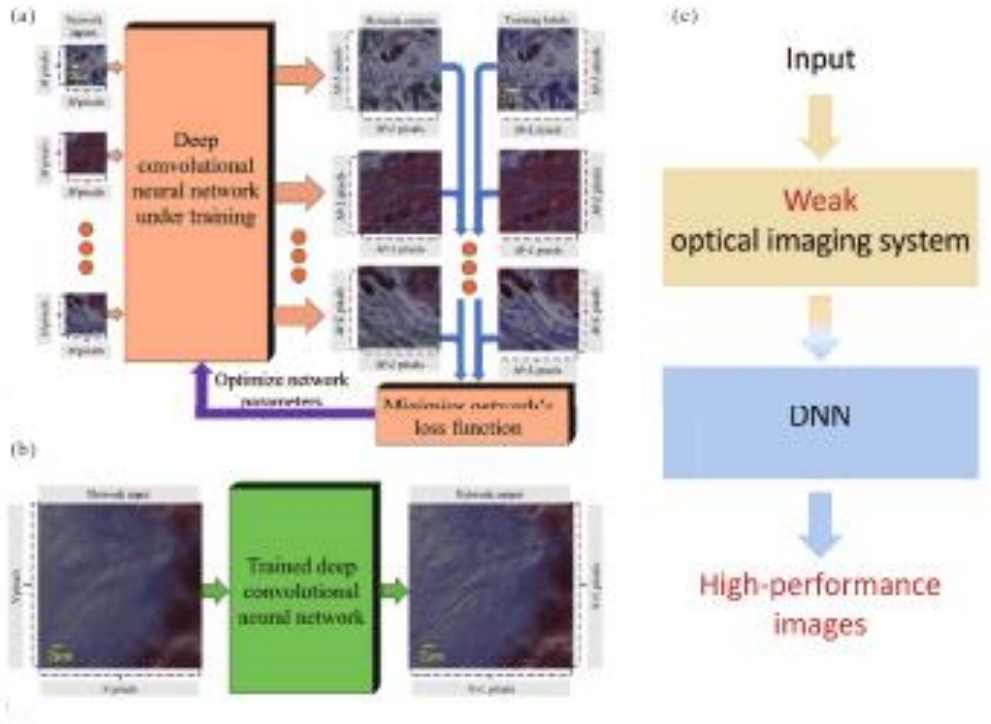


Рисунок 3 (Цветной онлайн) (а) Процесс обучения DNN для микроскопических изображений; (б) после обучения результаты DNN улучшаются [36] @Copyright 2017 The Optical Society; (с) схема микроскопии глубокого обучения. изображение, полученное с помощью обычного оптического микроскопа.

В [37] DNN применяется для решения сквозных обратных задач компьютерного построения изображений в безлинзовой системе построения изображений. Обученная DNN восстанавливает фазовые объекты при задании распространенных паттернов разности интенсивности. Другие приложения с AI в этом поле включают автофокусировку [38], оптическое сечение [39], шумоподавление [40] и так далее. Большинство систем оптической визуализации на базе искусственного интеллекта выделяются улучшенными характеристиками, такими как беспрецедентное количество достоинств или значительно упрощенный процесс эксплуатации. Мы относим эти похожие IPS к анализу результатов. Еще один поразительный тип IPS на базе искусственного интеллекта – это интеллектуальные схемы, которые еще далеко не исследованы. Широкополосные анализаторы сигналов с фотонной поддержкой являются одной из многообещающих платформ для создания IPS со схемным интеллектом [41,42]. Объединив широкополосную природу фотоники и интеллектуальных алгоритмов, можно получить уникальные сведения в обеих областях. Кроме того, ожидается, что новые устройства, такие как оптические частотные гребенки, будут поддерживать IPS на базе искусственного интеллекта, особенно на уровне микросхем [43].

Кроме того, были обнаружены некоторые новые приложения с помощью ИИ, такие как проектирование интегрированных фотонных устройств [44], дизайн наноструктур [45], высокопараллельное моделирование фотонных схем [46] и реконструкция ультракоротких импульсов [47]]. Благодаря способности изучать сложные взаимосвязи карт в будущем можно ожидать появления более интеллектуальных и простых в использовании систем с ИИ. Достижения нашей группы включают, в основном, фотонный аналого-цифровой преобразователь с функцией глубокого обучения (DL-PADC) [48] и повышение точности измерения мгновенной частоты Бриллюэна (BIFM) от CNN [49]. Во-первых, DL-PADC преодолевает узкое место в точности традиционной архитектуры без использования дорогостоящего оборудования или увеличения сложности системы. В частности, принятый DNN изучает внутренние недостатки отклика системы и восстанавливает искаженные данные, которые в основном вызваны нелинейностью оптического модулятора и несовпадением каналов из-за параллельной архитектуры. Схема DL-PADC показана на рисунке 4 и состоит из фотонного интерфейса, электронного квантования и восстановления данных с глубоким обучением. На верхних вставках нелинейность и показаны несовпадения каналов и соответствующие DNN для восстановления данных.

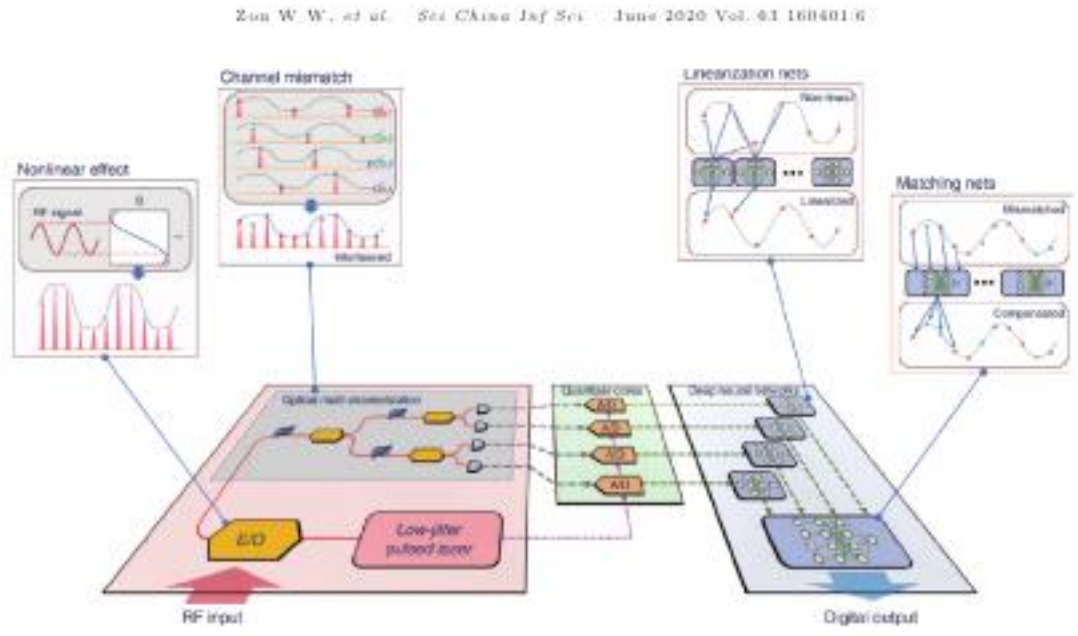


Рис. 4 (Цветной онлайн) Схема архитектуры DL-PADC [48] @Copyright 2019 Springer Nature.

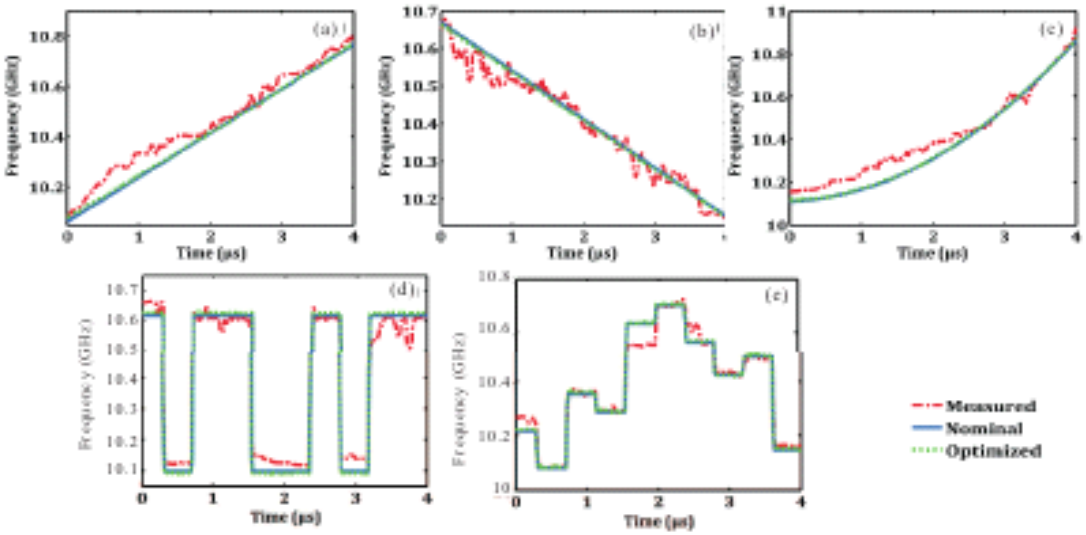


Рисунок 5 (Цветной онлайн) Оптимизированные результаты различных форматов сигналов с использованием метода на основе CNN в BIFM, включая линейную частотную модуляцию (LFM) (восходящий чирп) (a), LFM (нисходящий чирп) (б), нелинейную частотную модуляцию (NLFM) (c), двоичной а частотно-сдвиговой манипуляции (BFSK) (d) и частотной модуляции Костаса (e) [49] @Copyright 2019 The Optical Society.

Обратите внимание, что результаты показывают, что DL-PADC превосходит современные АЦП. Это исследование является важным шагом на пути к новым достижениям в информационных системах следующего поколения, таких как радары, системы визуализации и связи. Во-вторых, мы предлагаем метод оптимизации BIFM на основе CNN для произвольного сигнала с широкой полосой пропускания. Ошибки, вызванные системными дефектами результатов BIFM, восстанавливаются с помощью сопоставления измеренных и номинальных мгновенных частот, установленного CNN. На рисунке 5 экспериментально продемонстрирован эффект оптимизации, соответствующий различным форматам сигналов. Это исследование позволяет точно определять частоту широкополосных сигналов в таких сценариях, как электромагнитное наблюдение.

3.2 OANN

Когда OANN выступает в качестве ускорителя оптических вычислений, архитектура IPS с искусственным интеллектом, поддерживаемым OANN, представлена на рисунке 6. Цифровые электронные схемы (DEC) назначают AI-вычисления задание OAHH на скоростное исполнение. Полносвязные и сверточные OANN являются наиболее типичными формами для изучения. Как показано на рисунке 7, для выполнения матричного умножения была предложена каскадная архитектура с интерферометрами Маха-Цендера (MZI) и фазовращателями [50]. С помощью технологии фотонной интеграции на основе кремния исследователи интегрируют архитектуру в схему OANN. Ожидаемые результаты помех могут быть реали-

зованы путем настройки фазовращателей как настраиваемых параметров, соответствующих различным задачам. Обратите внимание, что после настройки фазовращателя схема потенциально становится полностью пассивной без затрат энергии во время работы. Потребляемая мощность относится к лазерному источнику и оптическому усилителю потенциала. В идеальных условиях энергоэффективность схемы по крайней мере на три порядка выше, чем у электронных компьютеров. Полностью оптическая схема OANN на основе дифракции представлена Lin et al. [51].

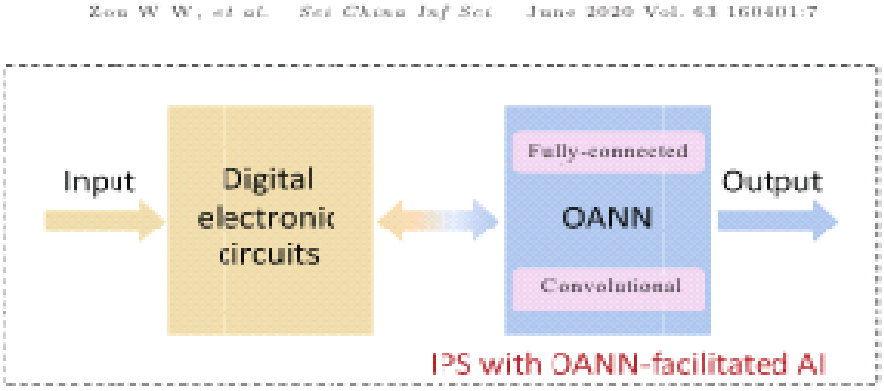


Рисунок 6 (Цветной онлайн) Архитектура IPS с искусственным интеллектом, поддерживаемым OANN.

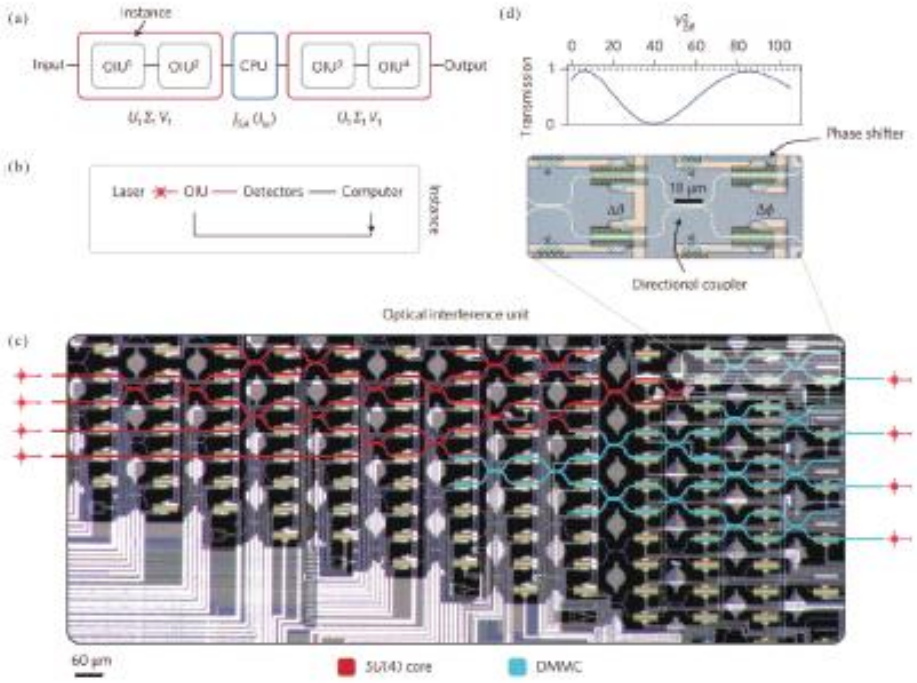


Рис. 7 (Цветной онлайн) (a) Процесс работы двухуровневой OANN. (б) Петля обратной связи, введенная в эксперименте. (c) Архитектура OANN на основе MZI, которая настраивается с помощью прилагаемых фазовращателей, как показано в (d) [50] @ Copyright 2017 Springer Nature.

На рисунке 8 входные данные закодированы в амплитуду входного сигнала, который проходит через пассивную многослойную дифракци-

онную сеть со скоростью света. Диффективные поверхности сети заранее спроектированы и реализованы

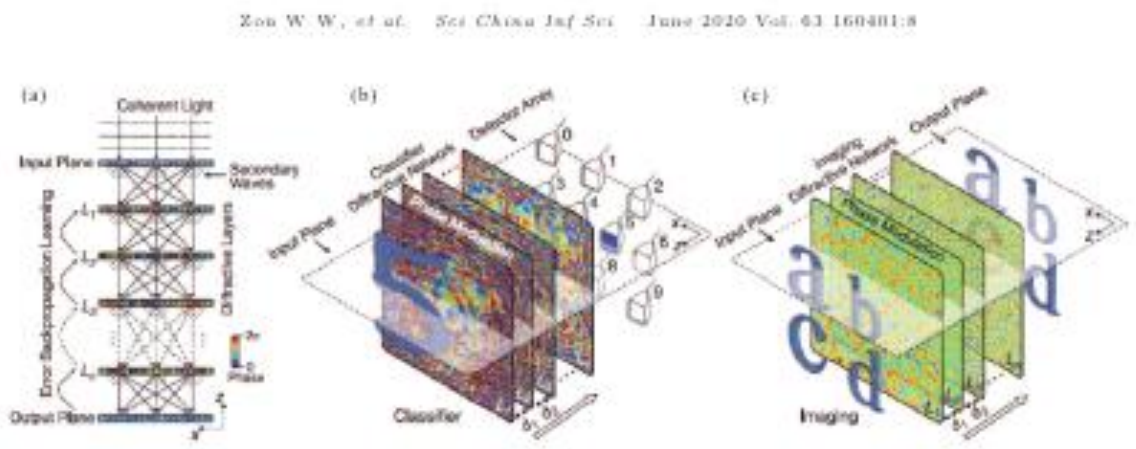


Рис. 8 (Цветной онлайн) (a) Схема OANN на основе дифракции с несколькими дифракционными слоями. OANN реализован в экспериментах как классификатор (b) и формирователь изображений (c) [51] @Copyright 2018 AAAS.

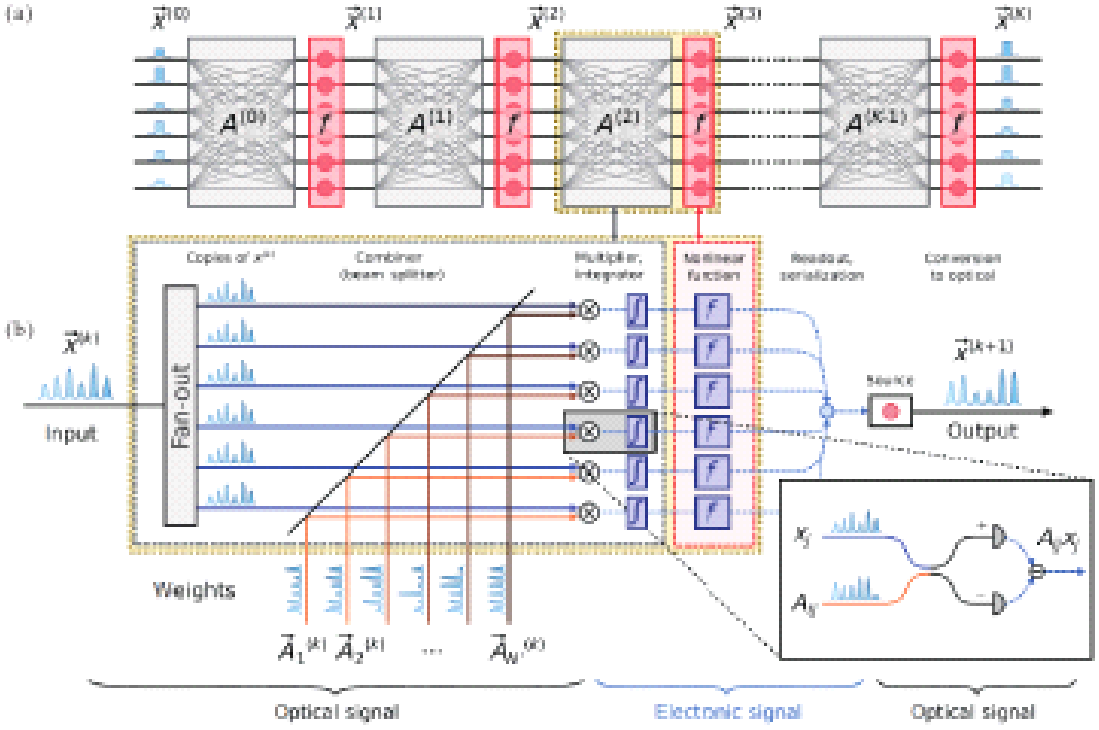


Рис. 9 (Цветной онлайн) (a) Многослойная схема нейронной сети; (b) архитектура однослойной OANN с эхкогерентным обнаружением [52] @Copyright 2019 American Physical Society.

Метод 3D-печати.

Наконец, свет попадает в детекторную матрицу, предназначенную для разных объектов, и выводит результаты классификации. В широко используемом тесте MNIST точность распознавания OANN на основе дифракции составляет 91,75%. Недавно Hamerly et al. [52] изучают

масштабируемую OANN, основанную на когерентном обнаружении с высокой скоростью (уровень ~ ГГц) и низким энергопотреблением (уровень субаттоджоулей). Схема архитектуры представлена на рисунке 9. Входные данные и веса кодируются в амплитуды последовательности оптических импульсов, что является многообещающим для высокоскоростного программирования и обучения. Ключом к низкому энергопотреблению является массивное пространственное мультиплексирование, основанное на оптических компонентах свободного пространства. Стоит отметить, что предлагаемая OANN может реализовать как полностью подключенные, так и сверточные сети. Между тем, обратное распространение и обучение демонстрируются на одном и том же оборудовании. Помимо полностью подключенной сети, исследователи также уделяют внимание новой оптической реализации CNN. В [53] описана CNN цифровой электроники и аналоговой фотоники (DEAP-CNN), использующая микрокольцевые резонаторы (MRR) и мультиплексирование с разделением по длине волны (WDM). MRR служат в качестве модуля входной модуляции и модуля взвешивания. Свертка, выполняемая DEAP, показана на рисунке 10. В этом процессе массив модуляции должен быть скорректирован, в то время как массив весов не нуждается в модификации из-за неизменного ядра. Оптоэлектронные или полностью оптические реализации функции нелинейной активации [54–58] и новое обучение метод [59] привлекает больше интересов. Первый обеспечивает полную реализацию OANN, потому что в предыдущих исследованиях необходимые нелинейные операции в основном выполнялись на компьютере. Метод обучения имеет решающее значение для практического использования OANN для решения различных задач, особенно с учетом недостатков OANN. Наша группа предпринимает некоторые усилия по OANN, включая высокоточное устройство оптической свертки [60] и реализацию CNN с высокой энергоэффективностью [61].

Как показано в [60], блок оптической свертки основан на решетках акустооптических модуляторов, как показано на рисунке 11. Входные данные и сверточное ядро подаются в матрицы модулятора. С помощью схемы многократного использования оборудования, устройства могут выполнять сложные CNN. При применении методов обучения получается более высокая точность. В [61] мы ввели оптические линии задержки для выполнения операций с данными с малой задержкой и энергопотреблением. Как показано на рисунке 12, WDM используется для повышения вычислительной мощности для каналов с параллельными длинами волн и уменьшения требуемой длины оптических линий задержки, что делает эту схему практичной. Согласно результатам теста AlexNet, энергоэффективность предлагаемой архитектуры как минимум в два-восемь раз выше, чем у предыдущих фотонных архитектур.

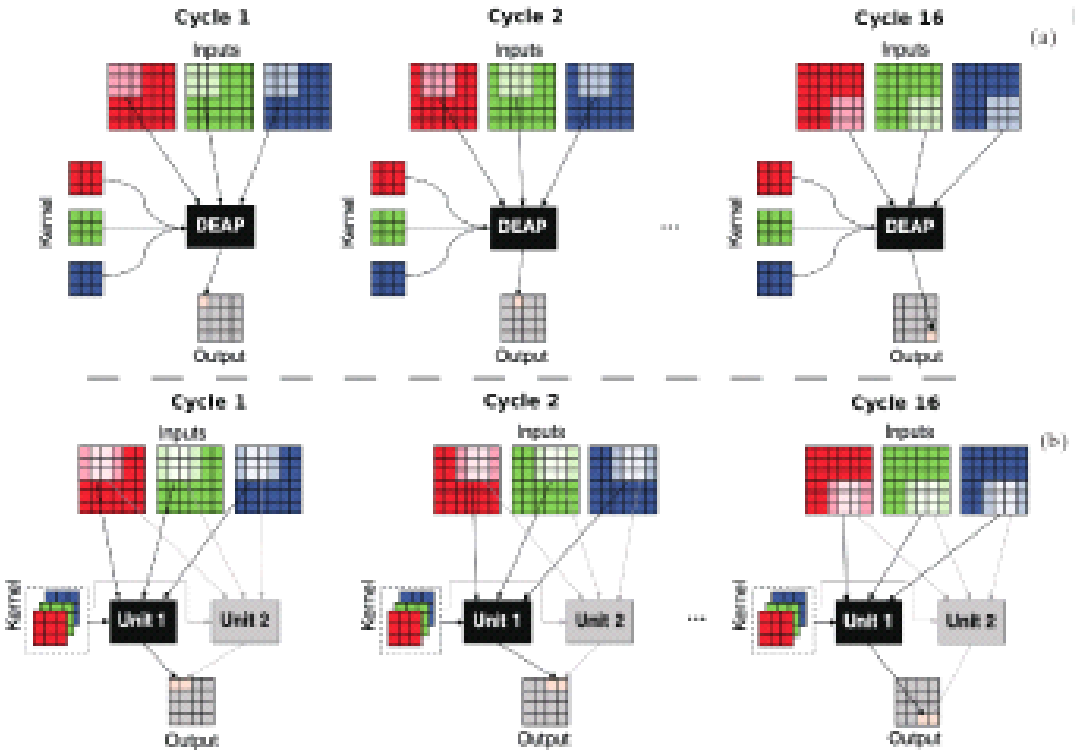


Рис. 10 (Цветной онлайн) (a) Свертка с использованием DEAP; (b) два сверточных модуля для выполнения свертки [53] @Copyright 2020 IEEE.

3.3 OBNN

Типичная архитектура IPS с искусственным интеллектом, поддерживаемым OBNN, представлена на рисунке 13. В отличие от OANN элементами в OBNN являются нейроны и синапсы. Исследования OBNN в основном связаны с темой нейроморфной фотоники, которая охватывает фотонные реализации нейронной возбудимости, синапсов, пластичности, зависящей от времени спайков (STDP), и нейронной сети спайков [62]. Как ключевой компонент OBNN, фотонные нейроны интенсивно исследуются. В [63] Nahmias et al. продемонстрировали, что лазер с вертикальным излучением поверхности резонатора (VCSEL) с внутрирезонаторным насыщающимся поглотителем (SA) аналогичен модели нейрона с утечкой-интеграцией и возгоранием (LIF), в которой нейронная возбудимость является важной особенностью. Возбудимость показана на рисунке 14. Входные возмущения накапливаются, вызывая большой ход фотонного нейрона. Помимо лазера с поглотителем, существует множество других методов реализации фотонных нейронов, включая метод оптической инжекции, метод оптической обратной связи и VCSEL с переключением поляризации [64, 65]. Подробный обзор возбудимых лазеров как фотонных нейронов можно найти в [66]. В послед-

нее время материал с фазовым переходом (PCM) используется для создания примитива пиков [67]. Далее исследуются вычислительные функции фотонного нейрона, такие как XOR. Синапсы играют критически важную роль в нейронной сети с пиками, которая также очень сфокусирована. Ченг и его сотрудники [69] реализовали полностью оптический фотонный синапс с помощью ИКМ. Схема и структура показаны на рисунке 15. Энергонезависимая природа PCM позволяет фотонный синапс без потребления статической энергии. Синаптический вес может непрерывно регулироваться числом входных оптических импульсов. Важно отметить, что фотонный синапс может обеспечить полностью оптический STDP, что является жизненно важным свойством для обучения и тренировки нейронных сетей с импульсными сигналами. Разработка нейронной сети с оптическими пиками имеет огромное значение. Tait et al. [70] предложили архитектуру для масштабируемой обработки фотонных выбросов. Сетевой протокол показан на рисунке 16. Схема широковещательной передачи и взвешивания включает WDM, лазерные нейроны, банки спектральных фильтров и волноводную петлю. Благодаря WDM, высокоплотное соединение между лазерными нейронами осуществляется в едином волноводе. В то же время блоки фильтров обеспечивают эффективный способ взвешивания

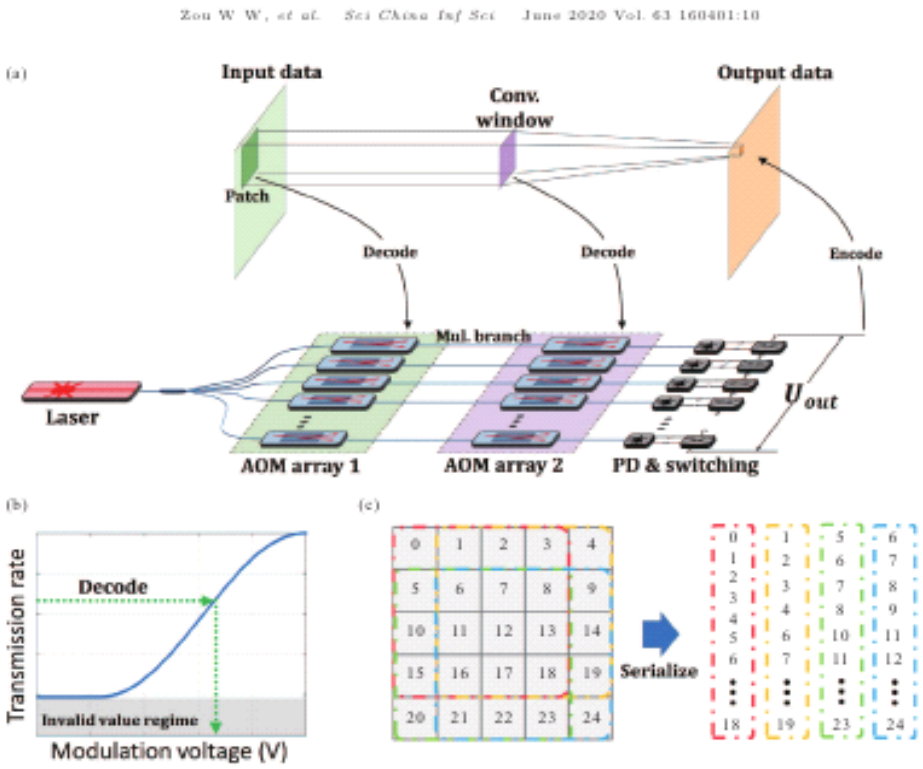


Рис. 11 (Цветной онлайн) (а) Архитектура оптического блока свертки; (б) скорость передачи в зависимости от напряжения модуляции используемых модуляторов; (с) иллюстрация метода сериализации [60] @Copyright 2019 The Optical Society

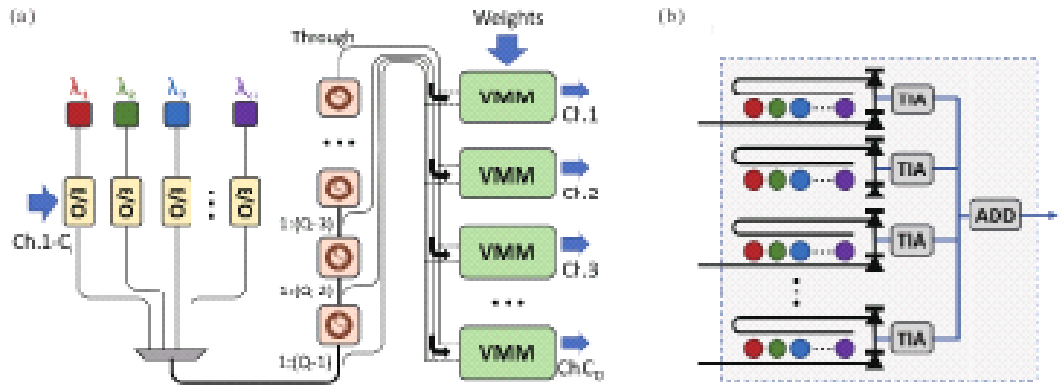


Рис. 12 (Цветной онлайн) (а) Схема реализации CNN с оптическими линиями задержки и WDM; (б) подробная структура ядра вектор-матричного умножения [61]. работа на основе VCSEL-SA [68].

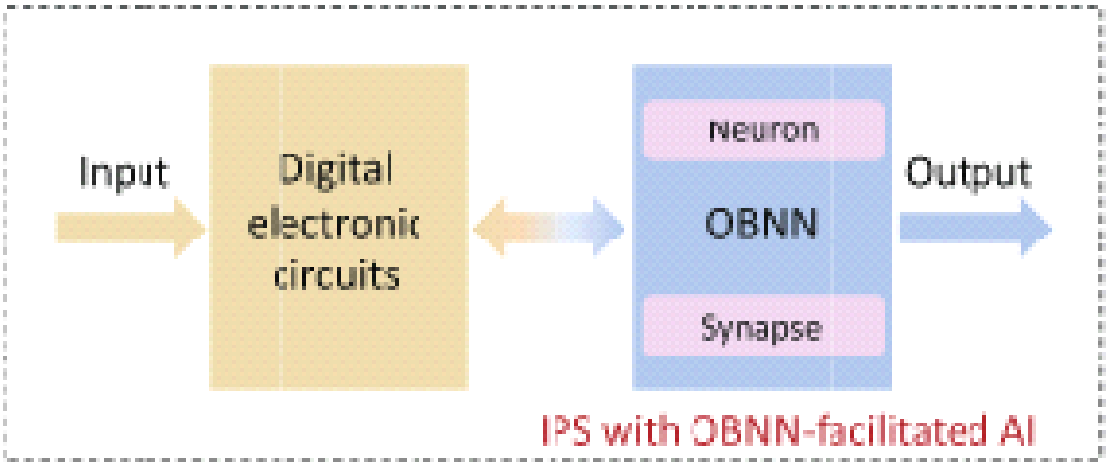


Рисунок 13 (Цветной онлайн) Архитектура IPS с ИИ, поддерживаемым OBNN

Недавно Feldman et al. представили полностью оптическую нейронную сеть на основе PCM. [71]. Как показано на рисунке 17, сеть состоит из весового модуля, нейронного интерфейса и генерации спайков. Обратите внимание, что петля обратной связи спайков адаптирована для изучения веса с использованием STDP. Полностью оптическая реализация обеспечивает сети потенциал сверхширокополосной работы. В отношении большой сети обсуждается масштабируемость. Кроме того, реализуется контролируемый и неконтролируемый процесс обучения для сети, чтобы распознавать шаблон.

Считается, что синаптическая пластичность тесно связана с обучением в нейронных сетях с пиками, поэтому STDP имеет решающее значение для построения обучающего модуля в OBNN. Есть несколько исследований, посвященных фотонной реализации STDP с использованием модуляции перекрестного усиления, насыщающегося поглощения и нелинейного вращения поляризации [72–75].

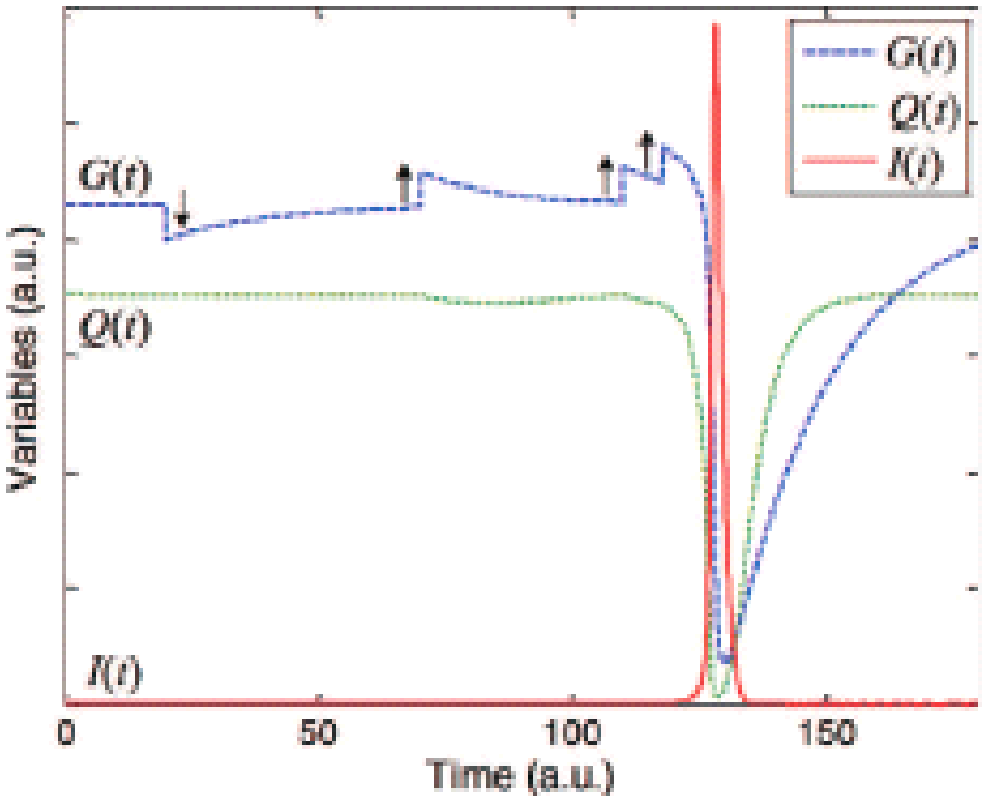
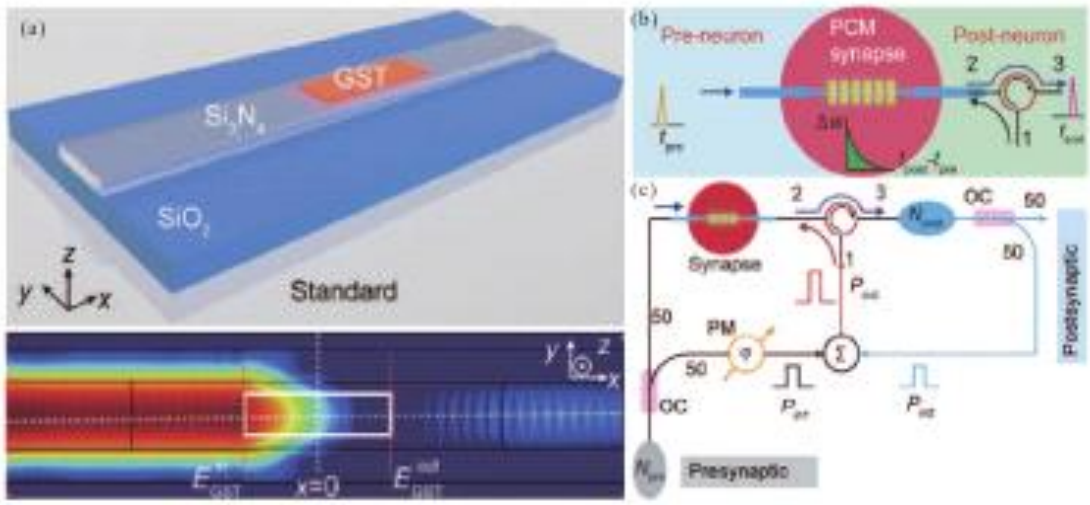


Рис. 14 (Цветной онлайн) Результаты моделирования лазера с СА с возбудимостью. Оптический импульс высвобождается, когда входные возмущения накапливаются до порогового значения [63] @Copyright 2013 IEEE.



(b) (c)

Рисунок 15 (Цветной онлайн) (а) Схема фотонного нейрона с PCM (вверху) и распределением моды TE (внизу); (б) фотонный синапс напоминает функцию биологического синапса; (с) экспериментальная установка реализации STDP фотонным синапсом [69] @Copyright 2017 The AAAS.

STDP-подобный ответ эмулируется для управления весом одного синапса. Обратите внимание, что эти модули основаны на дискретных устройствах с низкой масштабируемостью, чтобы соответствовать высокой плотности использования синапсов.

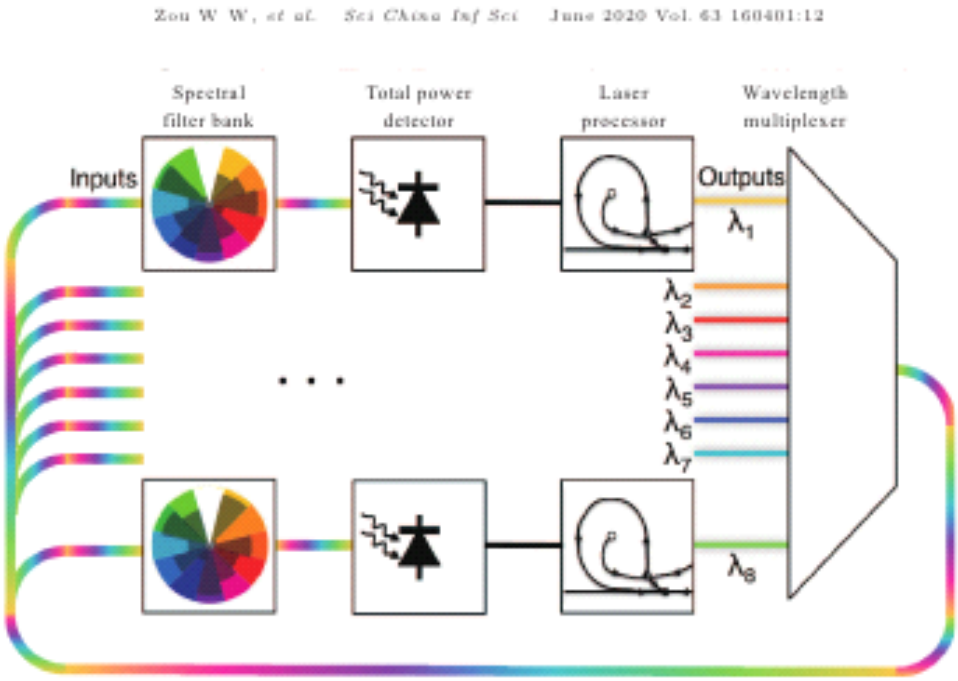


Рис. 16 (Цветной онлайн) Схема широковещательной передачи и взвешивания, включая массив лазеров, WDM, спектральные фильтры и волноводную петлю [70] @Copyright 2014 IEEE.

Одна попытка состоит в том, чтобы разделить модуль STDP между группой синапсов, что может помешать процессу обновления веса в реальном времени. Если STDP может быть встроен в устройства фотонных синапсов в качестве неотъемлемой части, то более удобно построить нейроморфную сеть с потенциалом обучения [69]. Есть авансы, сделанные нашей группой в ОБНН. Мы демонстрируем широко используемый лазерный диод с распределенной обратной связью (DFB-LD), который можно использовать в трех приложениях при обработке нейроморфной информации, для распознавания образов, реализации STDP на одной длине волны и измерения азимута звука. Экспериментальные результаты измерения азимута звука показаны на рисунке 18. Кроме того, мы предлагаем фотонную нейроморфную сеть на основе DFB-LD для распознавания пространственно-временных образов [76]. Как показано на рисунке 19, сеть может изучить целевой шаблон с помощью обучающего модуля STDP, также реализованного с помощью DFB-LD, что удобно для потенциальной интеграции.

4. Крупномасштабная гибридная интеграция для IPS

IPS – это сложная система, включающая активные / пассивные оптические компоненты и ВЧ драйверы. Крупномасштабная гибридная интеграция предлагает возможность низкого энергопотребления и сверхбыстрой обработки в IPS. Кроме того, встроенная IPS с большим масштабом поддерживает решения для более сложных задач, поскольку сложность пропорциональна возможностям сети как в OANN, так и в OBNN. В последнее время появляющаяся крупномасштабная технология гибридной интеграции направлена на объединение активных / пассивных оптических компонентов и радиочастотных / электронных схем на одном кристалле. Мы суммируем репрезентативный прогресс в таблице 1 [6,77–83]. Развитие технологии крупномасштабной гибридной интеграции освещает будущее IPS.

5. Перспективы и выводы

Для IPS на базе искусственного интеллекта задача состоит в том, чтобы спроектировать эффективную ИНС в соответствии с характеристиками фотонных систем. В настоящее время большинство исследований сосредоточено на демонстрации технико-экономического обоснования использования ИНС.

метод, применяемый к системам. IPS на базе искусственного интеллекта привлекает наибольшее внимание своей улучшенной производительностью. Тем не менее, оценка метода, основанного на ИНС, часто игнорируется, например, потенциал реального времени, стоимость обучения и доступность данных в реальных приложениях. Мы уверены, что появятся новые IPS на базе искусственного интеллекта. Следующим важным вопросом является сопоставление характеристик систем с соответствующими структурами ИНС и методами обучения. Таким образом, IPS на базе искусственного интеллекта раскрывает потенциал фотонных технологий и позволяет применять более мощную систему в реальных сценариях.

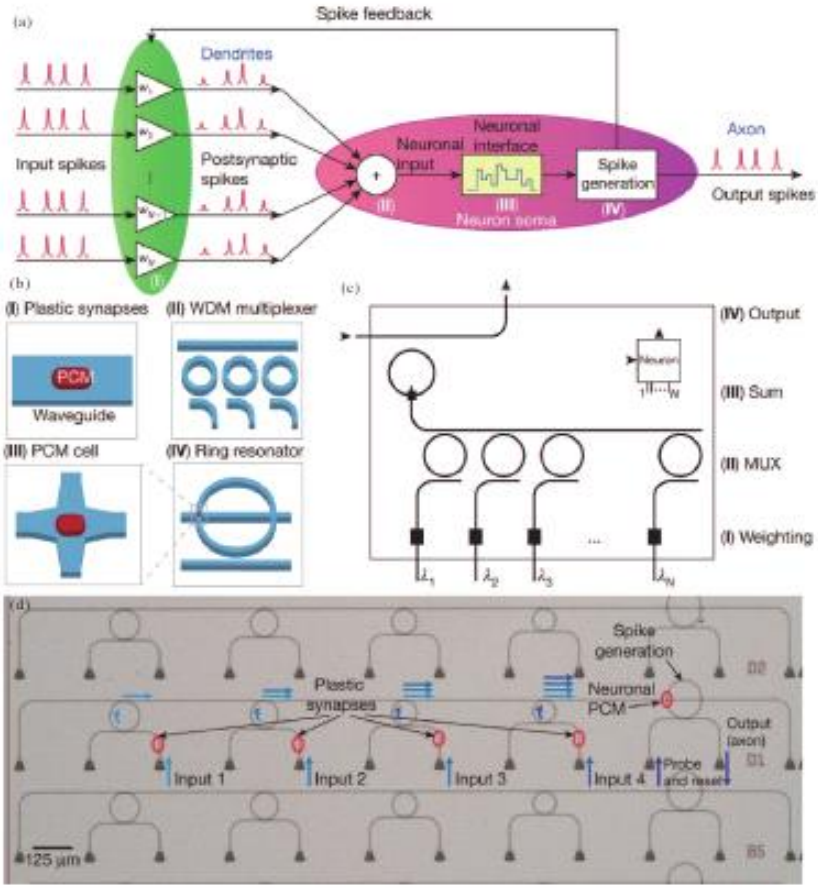


Рис. 17 (Цветной онлайн) (a) и (b) Схема нейронной сети на основе импульсов PCM; (в) схема интегрированного оптического нейрона; (d) оптическая микрофотография трех оптических нейронов с четырьмя входными портами соответственно [71] @Copyright 2019 Springer Nature.

По мнению OANN, эффективные, компактные, программируемые, энергопотребляющие сетевые архитектуры и реализации заслуживают постоянного изучения. Кроме того, смещение фокуса с численного моделирования и экспериментов по проверке концепции на изготовленное оборудование имеет решающее значение для OANN в направлении практических приложений. Обратите внимание, что сложность интеграции возрастает с увеличением масштаба OANN. Это серьезная проблема – найти баланс между практичностью интеграции и необходимыми вычислительными возможностями. Мы также должны следить за построением экосистемы OANN, например, за интерфейсом с популярной вычислительной структурой. Для OBNN фотонные нейроны и синапсы должны быть краткими и диверсифицированными в соответствии с биологической нейронной сетью. Высокоплотные взаимосвязи различных нейронов и синапсов – замечательные особенности мозга. Однако недавние исследования практически не концентрируются на каскадируемости и масштабируемости компонентов. В то же время типы имеющихся комплектующих далеки от адекватных,

В заключение, мы сначала представили необходимость разработки IPS. Затем были представлены объяснения концепции IPS и потенциальные преимущества, соответствующие различным версиям, а именно IPS на основе искусственного интеллекта, IPS с искусственным интеллектом, поддерживаемым фотоникой, и нейроморфная IPS. Кроме того, мы рассмотрели значительные достижения архитектур и компонентов IPS, а также исследования, проведенные нашей группой. Наконец, обсуждались перспективы дальнейшего развития. Мы надеемся на продолжение Zou W W, et al. Sci China Inf Sci Июнь 2020 г. 63 160401: 15 следовательно, не хватает богатства вычислительных свойств. Кроме того, OBNN – важный путь к нейроморфной IPS.

Экспериментальные результаты измерения азимута звука с помощью DFB-LD. Выходы при задержке 3 мкс (а) и 4 мкс (б). Разница 2-го пика

распознавания пространственно-временных образов на основе DFB-LD с обучающим модулем STDP [76] @Copyright 2020 Авторы.

Таблица 1

Прогресс в крупномасштабной гибридной интеграции Гибридный тип Справочная информация Основные моменты Активный / Пассивный [77] Пассивные строительные блоки в общей технологии интеграции [78] Первая однослойная активно-пассивная фотонная интеграция Al2O3 Гетерогенная [79] Систематические обзоры по III-V / кремниевой интеграции [80] Гетерогенная 2D / 3D фотонная интеграция Цифровая / аналоговая [81] Обсуждение цифровых фотонных интегральных схем [6] 70 миллионов транзисторов и 850 фотонных компонентов на микросхеме Photonic / Electronic [82] Способ интеграция фотоники с современной наноэлектроникой [83]

Терагерцовые интегрированные гибридные электронно-фотонные системы вносят свой вклад в повышение практичности и интеллектуальности IPS для будущих задач обработки информации. Внимание будет постоянно уделяться топологическим структурам, схемам обучения и принципам эволюции в биологической нейронной сети, которые могут быть ключом к окончательной версии нейроморфной IPS.

Благодарности: Эта работа была поддержана Национальной программой ключевых исследований и разработок Китая (грант № 2019YFB2203700) и Национальным фондом естественных наук Китая (грант № 61822508).

Литература

1 Кикучи К. Основы когерентной оптоволоконной связи. J Lightw Technol, 2016, 34: 157–179

2 Яо Дж. П. Микроволновая фотоника. J Lightw Technol, 2009, 27: 314–335

3 Лян Дж. И, Ван Л. В. Однократная сверхбыстрая оптическая визуализация. Optica, 2018, 5: 1113–1127
4 Чен Дж. Х., Ли Д. Р., Сюй Ф. Оптические микроволоконные датчики: механизмы обнаружения и последние достижения. J Lightw Technol, 2019, 37: 2577–2589
5 Capmany J, Novak D. Микроволновая фотоника объединяет два мира. Nat Photon, 2007, 1: 319–330
6 Sun C, Wade M. T., Lee Y, et al. Однокристальный микропроцессор, который взаимодействует напрямую с помощью света. Nature, 2015, 528: 534–538
7 Хан М. Х, Шен Х, Сюань И и др. Генерация сигналов произвольной радиочастоты со сверхшироким диапазоном частот с помощью формирователя спектра на основе кремниевого фотонного чипа. Нат Фотон, 2010, 4: 117–122
8 Чжуанг Л. М., Роэлозен Ч. Г. Х., Хукман М. и др. Программируемый чип процессора фотонных сигналов для радиочастотных приложений. Optica, 2015, 2: 854–859
9 Миллер Д. А. Самоконфигурирующийся универсальный линейный оптический компонент. Photon Res, 2013, 1: 1–15
10 Перес Д., Гасулла И., Круджингтон Л. и др. Ядро многоцелевого кремниевого процессора сигналов фотоники. Nat Commun, 2017, 8: 636
11 Перес Д., Гасулла И., Кэпмани Дж. К программируемым микроволновым фотонным процессорам. J Lightw Technol, 2018, 36: 519–532
12 Чжан Дж. Дж., Яо Дж. П. Процессор микроволновых фотонных сигналов для генерации микроволновых сигналов произвольной формы и сжатия импульсов. J Lightw Technol, 2016, 34: 5610–5615
13 Гарсия-Мека С., Лехаго С., Бримонт А. и др. Встроенная в кристалл беспроводная кремниевая фотоника: от реконфигурируемых межсоединений до устройств «лаборатория на кристалле». Light Sci Appl, 2017, 6: e17053
14 Сильвер Д., Хуанг А., Мэддисон С. Дж. И др. Освоение игры в го с глубокими нейронными сетями и поиском по дереву. Nature, 2016, 529: 484–489
15 Silver D, Schrittwieser J, Simonyan K et al. Освоение игры в го без человеческого знания. Nature, 2017, 550: 354–359
16 Тополь Э. Дж. Высокоэффективная медицина: конвергенция человеческого и искусственного интеллекта. Nat Med, 2019, 25: 44–56
17 Абдель-Хамид О., Мохамед А., Цзян Х. и др. Сверточные нейронные сети для распознавания речи. IEEE / ACM Trans Audio Speech Lang Process, 2014, 22: 1533–1545
18 Браун Н., Сандхольм Т. Сверхчеловеческий ИИ для многопользовательского покера. Science, 2019, 365: 885–890
19 Уинфилд А. Этические стандарты робототехники и искусственного интеллекта. Nat Electron, 2019, 2: 46–48
20 Ван Дж. Г., Чжоу Л. Б. Распознавание света с помощью визуализации с расширенным динамическим диапазоном и глубокого обучения. IEEE Trans Intell Transp Syst, 2019,20: 1341–1352
21 Минасиан Р. А. Фотонная обработка микроволновых сигналов. IEEE Trans Microw Theor Techn, 2006, 54: 832–846

22 Миллер Д. А. Б. Идеальная оптика с несовершенными компонентами. Optica, 2015, 2: 747–750
23 Ян Г., Цзоу В. В., Ю. Л. и др. Компенсация многоканальных рассогласований в высокоскоростном фотонном аналого-цифровом преобразователе высокого разрешения. Opt Express, 2016, 24: 24061–24074
24 Минзиони П., Альберти Ф., Скини А. Методы устранения нелинейности во встроенных каналах с помощью оптического фазового сопряжения. J Lightw Technol, 2005, 23: 2364–2370
25 Park S. W., Park J Y, Bong K, et al. Энергоэффективный и масштабируемый процессор глубокого обучения / логического вывода с тетрапараллельной архитектурой MIMD для приложений с большими данными. IEEE Trans Biomed Circ Syst, 2015, 9: 838–848 Zou W. W. et al. Sci China Inf Sci Июнь 2020 г. 63 160401: 16
26 Уолдроп М. М. Фишки отменены по закону Мура. Nature, 2016, 530: 144–147
27 Тейт А.Н., де Лима Т.Ф., Нахмиас М.А. и др. Кремниевый фотонный модулятор нейрона. Phys Rev Appl, 2019, 11: 064043
28 Денев С., Алеми А., Бурдукан Р. Мозг как эффективный и надежный адаптивный обучающийся. Neuron, 2017, 94: 969–977
29 Рой К., Джайсвал А., Панда П. К машинному интеллекту на основе шипов с нейроморфными вычислениями. Nature, 2019, 575: 607–617
30 Маасс В., Натшлаггер Т., Маркрам Х. Вычисления в реальном времени без стабильных состояний: новая структура для нейронных вычислений, основанная на возмущениях. Neural Comput, 2002, 14: 2531–2560
31 Ма В., Зидан М. А., Лу В. Д. Нейроморфные вычисления с мемристивными устройствами. Sci China Inf Sci, 2018, 61: 060422
32 Ву Н. Дж. Нейроморфные чипы зрения. Sci China Inf Sci, 2018, 61: 060421
33 Ян Б. Н., Чен И. Р., Ли Х. Проблемы нейроморфной вычислительной системы на основе мемристора. Sci China Inf Sci, 2018, 61: 060425
34 Cully A, Clune J, Tarapore D, et al. Роботы, которые могут адаптироваться, как животные. Nature, 2015, 521: 503–507
35 Барбастатис Дж., Озкан А., Ситу Дж. Об использовании глубокого обучения для построения компьютерных изображений. Optica, 2019, 6: 921–943
36 Ривенсон Y, Goro¨cs Z, Gu¨naydin H, et al. Микроскопия глубокого обучения. Optica, 2017, 4: 1437–1443
37 Синха А., Ли Дж., Ли С. и др. Вычислительная визуализация без линз с помощью глубокого обучения. Optica, 2017, 4: 1117–1125
38 Wu Y C, Rivenson Y, Zhang Y B и др. Увеличенная глубина резкости при формировании голографических изображений с использованием автофокусировки на основе глубокого обучения и восстановления фазы. Optica, 2018, 5: 704–710
39 Zhang X Y, Chen Y F, Ning K F и др. Метод оптического сечения с глубоким обучением. Opt Express, 2018, 26: 30762–30772
40 Manifold B, Thomas E, Francis A.T, et al. Устранение шумов на изображениях микроскопии вынужденного комбинационного рассеяния посредством глубокого обучения. Biomed Opt Express, 2019, 10: 3860–3874

41 Эсман Д. Дж., Атаи В., Куо Б. П. и др. Циклостационарный анализ широкополосных радиочастотных сигналов с помощью гребенки. J Lightw Technol, 2017, 35: 3705–3712
42 Ma M., Адамс Р., Чен Л. Р. Интегрированный фотонный чип позволил синхронизировать многоканальный широкополосный анализатор радиочастотного спектра. J Lightw Technol, 2017, 35: 2622–2628
43 Фортье Т., Бауманн Э. 20 лет разработок в области технологии и приложений оптических частотных гребенок. Commun Phys, 2019, 2: 153
44 Хаммонд А. М., Камачо Р. М. Разработка интегрированных фотонных устройств с использованием искусственных нейронных сетей. Opt Express, 2019, 27: 29620–29638
45 Malkiel I., Mrejen M, Nagler A, et al. Дизайн и характеристика плазмонных наноструктур с помощью глубокого обучения. Light Sci Appl, 2018, 7: 60
46 Laporte F, Dambre J, Bienstman P. Высокопараллельное моделирование и оптимизация фотонных схем во временной и частотной области на основе фреймворка глубокого обучения PyTorch. Sci Rep, 2019, 9: 5918
47 Захавы Т., Дикопольцев А., Мосс Д. и др. Реконструкция ультракоротких импульсов методом глубокого обучения. Optica, 2018, 5: 666–673
48 Xu S F, Zou X T, Ma B W, et al. Фотонное аналого-цифровое преобразование на основе технологии глубокого обучения. Light Sci Appl, 2019, 8:66
49 Zou X T, Xu S F, Li S J, et al. Оптимизация мгновенного измерения частоты Бриллюэна с помощью сверточных нейронных сетей. Opt Lett, 2019, 44: 5723–5726
50 Шен И. К., Харрис Н. К., Скирло С. и др. Глубокое обучение с когерентными нанофотонными схемами. Нат Фотон, 2017, 11: 441–446
51 Лин Х, Ривенсон Й., Ярдимчи Н. Т. и др. Полностью оптическое машинное обучение с использованием глубоких нейронных сетей. Science, 2018, 361: 1004–1008
52 Хамерли Р., Бернштейн Л., Слуддс А. и др. Крупномасштабные оптические нейронные сети на основе фотоэлектрического умножения. Phys Rev X, 2019, 9: 021032
53 Bangari V, Marquez B.A, Miller H, et al. Цифровая электроника и аналоговая фотоника для сверточных нейронных сетей (DEAP-CNN). IEEE J Sel Top Quantum Electron, 2020, 26: 1–13
54 Уильямсон И. Д., Хьюз Т. В., Минков М. и др. Перепрограммируемые электрооптические нелинейные функции активации для оптических нейронных сетей. IEEE J Sel Top Quantum Electron, 2020, 26: 1–12
55 Джордж Дж. К., Мехрабиан А., Амин Р. и др. Нейроморфная фотоника с модуляторами электропоглощения. Opt Express, 2019, 27: 5181–5191
56 Zuo Y, Li B H, Zhao Y J, et al. Полностью оптическая нейронная сеть с нелинейными функциями активации. Optica, 2019, 6: 1132–1137
57 Мургиас-Александрис Г., Цакиридис А., Пассалис Н. и др. Полностью оптический нейрон с функцией активации сигмовидной кишки. Opt Express, 2019, 27: 9620–9630

58 Miscuglio M, Mehrabian A, Hu Z B и др. Полностью оптическая нелинейная функция активации для фотонных нейронных сетей. Opt Mater Express, 2018, 8: 3851–3863
59 Хьюз Т. В., Минков М., Ши Ю. и др. Обучение фотонных нейронных сетей с помощью обратного распространения ошибки на месте и измерения градиента. Optica, 2018, 5: 864–871
60 Xu S. F, Wang J, Wang R, et al. Архитектура высокоточного блока оптической свертки для сверточных нейронных сетей за счет каскадных массивов акустооптических модуляторов. Opt Express, 2019, 27: 19778
61 Xu S. F, Wang J, Zou W. W. Интегрированные фотонные сверточные нейронные сети с высокой энергоэффективностью. ArXiv: 1910.12635
62 Прукнал П. Р., Шастри Б. Дж. Нейроморфная фотоника. Бока-Ратон: CRC Press, 2017
63 Нахмиас М.А., Шастри Б.Дж., Тейт А.Н. и др. Излучающий лазерный нейрон с функцией интеграции и зажигания для сверхбыстрых когнитивных вычислений. IEEE J Sel Top Quantum Electron, 2013, 19: 1–12 Zou W W, et al. Sci China Inf Sci Июнь 2020 г. 63 160401: 17
64 Robertson J, Wade E, Kopp Y, et al. К нейроморфным фотонным сетям сверхбыстрых импульсных лазерных нейронов. IEEE J Sel Top Quantum Electron, 2020, 26: 1–15
65 Xiang S Y, Zhang H, Guo X X и др. Каскадная нейроноподобная динамика спайков в связанных VCSEL, подверженных инжекции ортогонально поляризованных оптических импульсов. IEEE J Sel Top Quantum Electron, 2017, 23: 1–7
66 Прукнал П. Р., Шастри Б. Дж., Де Лима Т. Ф. и др. Недавний прогресс в полупроводниковых возбуждаемых лазерах для обработки фотонных пиков. Adv Opt Photon, 2016, 8: 228–299
67 Чакраборти И., Саха Дж., Рой К. Фотонный вычислительный примитив в памяти для пиков нейронных сетей с использованием материалов с фазовым переходом. Phys Rev Appl, 2019, 11: 014063
68 Xiang S Y, Ren Z X, Zhang Y H, et al. Полностью оптическая нейроморфная операция XOR с ингибирующей динамикой одиночного фотонного импульсного нейрона на основе VCSEL-SA. Opt Lett, 2020, 45: 1104–1107
69 Ченг З. Г., Риос С., Пернис В. Х. П. и др. Фотонный синапс на кристалле. Sci Adv, 2017, 3: e1700160
70 Tait A N, Nahmias M. A, Shastri B.J. и др. Широковещательная передача и вес: интегрированная сеть для масштабируемой обработки фотонных пиков. J Lightw Technol, 2014, 32: 4029–4041
71 Фельдманн Дж., Янгблад Н., Райт К. Д. и др. Полностью оптические нейросинаптические сети с возможностью самообучения. Nature, 2019, 569: 208–214
72 Xiang S Y, Zhang Y L, Gong J K, et al. Обучение неконтролируемых спайковых паттернов на основе STDP в нейронной сети с фотонными пиками с VCSEL и VCSOA. IEEE J Sel Top Quantum Electron, 2019, 25: 1–9
73 Ren Q S, Zhang Y L, Wang R, et al. Оптическая пластичность, зависящая от времени спайка, с зависящим от веса окном обучения и модуляцией вознаграждения. Opt Express, 2015, 23: 25247–25258

74 Тул Р., Тейт А. Н., де Лима Т. Ф. и др. Фотонная реализация алгоритмов пластичности и обучения биологических нейронных систем, зависящих от времени спайков. J Lightw Technol, 2016, 34: 470–476
75 Фок М. П., Тиан Й., Розенблут Д. и др. Обнаружение опережения / запаздывания импульса для адаптивной обратной связи и управления на основе оптической пластичности, зависящей от времени всплеска. Opt Lett, 2013, 38: 419–421
76 Ma B W, Chen J P, Zou W. W. Фотонная нейроморфная сеть на основе DFB-LD для распознавания пространственно-временных паттернов. B: Proceedings of Optical Fiber Communication Conference, San Diego, 2020. M2K.2
77 Smit M, Leijtens X. Интеграция пассивных и активных компонентов в PIC на основе InP B: Proceedings of Advances in Optical Sciences Congress, Гонолулу, 2009. ITuB2
78 ван Эммерик К.И., Дейкстра М., де Гёде М. и др. Однослойная активно-пассивная платформа для фотонной интеграции Al2O3. Opt Mater Express, 2018, 8: 3049–3054
79 де Валикур Дж., Чанг С. М., Эгглстон М. С. и др. Фотонная интегральная схема на основе гибридной интеграции III-V / кремния. J Lightw Technol, 2018, 36: 265–273
80 Ю С. Дж.Б, Гуан Б. Б., Скотт Р. П. Гетерогенные 2D / 3D фотонные интегрированные микросистемы. Microsyst Nanoeng, 2016, 2: 16030
81 Хилл М., Смит М., Кромбез П. и др. Цифровая и аналоговая фотонная интеграция B: Proceedings of Integrated Photonics and Nanophotonics Research and Applications, Boston, 2008. IWC1
82 Atabaki A H, Moazeni S, Pavanello F, et al. Интеграция фотоники с кремниевой наноэлектроникой для следующего поколения систем на кристалле. Nature, 2018, 556: 349–354
83 Сенгупта К., Нагацума Т., Миттлман Д. М. Терагерцовые интегрированные электронные и гибридные электронно-фотонные системы. Нат Электрон, 2018, 1: 622–635.

НЕЗАБЫВАЕМЫЕ КОНВОЛЮЦИОННЫЕ КЛАССИФИКАТОРЫ НЕЙРОННЫХ СЕТЕЙ ЧЕРЕЗ УСИЛЕНИЕ ИЗУЧЕНИЕ ПРЕДПОЧТЕНИЯ

Малашин Роман О.

Реферат

Обучение ансамбля нейронной сети (CNN) с целевой функцией, основанной на принципе наименьшего действия; он включает компонент потребления ресурсов. Мы обучаем агента воспринимать изображения с помощью набора предварительно обученных классификаторов и хотим, чтобы результирующая динамически сконфигурированная система развернула вычислительный граф с траекторией, которая относится к минимальному количеству операций и максимальной ожидаемой точности. Предлагаемая архитектура агента неявно аппроксимирует требуемую функцию выбора классификатора с помощью обучения с подкреплением. Наши экспериментальные результаты доказывают, что, если агент использует динамическую (и контекстно-зависимую) структуру вычислений, он превосходит обычное ансамблевое обучение. Ключевые слова Ансамблевое обучение Обучение в условиях вычислительных ограничений Мета-обучение Динамически конфигурируемые системы

1 Введение

Ансамблевое обучение – это подход к машинному обучению, который относится к получению предиктора (сильного классификатора или комитета), который имеет форму взвешенной комбинации базовых моделей (слабые ученики). Бэггинг, бустинг и укладка – хорошо известные ансамблевые методы с практическим применением. Ансамблевое обучение предполагает, что все слабые ученики используются для прогнозирования. Это нарушает принцип минимального потребления энергии. Мы называем этот фундаментальный принцип наименьшего действия [Малашин, 2019]. Шелепин и др. Соавторы показали, что принцип наименьшего действия можно рассматривать как принцип познания в зрении [Шелепин, Красильников, Шелепин и др., 2006]. В физике принцип наименьшего действия гласит, что объекты в космосе следуют траекториям, которые удовлетворяют минимуму двухкомпонентного функционала, называемого действием. Мы адаптируем этот принцип: вычислительный граф должен проходить по траектории, которая удовлетворяет максимальной ожидаемой точности и минимальным вычислительным затратам. С точки зрения ансамблевого обучения, если пример прост, мы предпочитаем полагаться на реакцию всего нескольких слабых учеников

(используйте короткий путь в динамическом вычислительном графе), в то время как вычислительно тяжелый анализ оправдан для сложных случаев. Обычное разреженное усиление предполагает, что некоторые функции могут отсутствовать во время прогнозирования, но не учитывает желательность такого «отсутствия». Простой, но популярный подход, включающий принцип наименьшего действия, – это списки решений, когда большинство простых случаев могут быть отклонены ранними тестами [Viola and Jones, 2001]. Тем не менее, этот подход применим только для двоичной классификации; в нем отсутствуют многие желательные возможности конфигурирования динамических графов [Малашин, 2019]. Мы ставим задачу изучения разреженных ансамблевых классификаторов с учетом принципа наименьшего действия. Проблему можно решить с помощью обучения с подкреплением, научив агента воспринимать изображение с помощью набора классификаторов CNN, которые обучаются извне. Конечное вознаграждение агента складывается из точности за вычетом затрат времени. В этой работе мы концентрируемся на задаче классификации изображений, хотя этот подход можно естественным образом расширить в более широкие области анализа данных.

Выбранный фрагмент x Агент Выбранный классификатор f Пул обученных классификаторов f (x) Решение

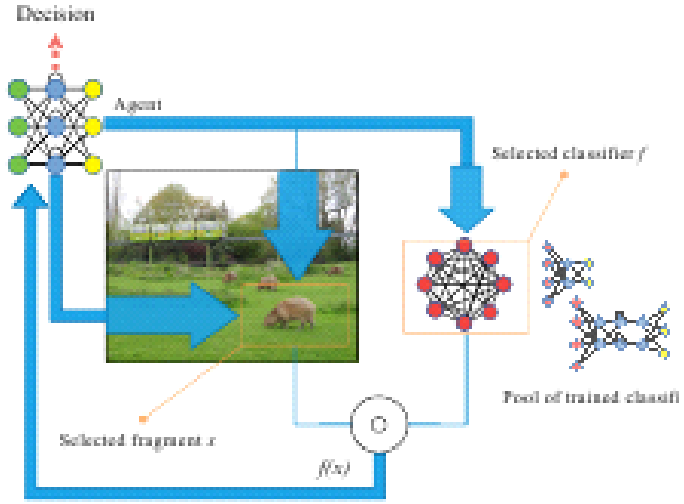


Рисунок 1. Изученная схема взаимодействия агента с изображением через пул обученных классификаторов с визуальным вниманием механизм. Пример фотографии Гарета Джеймса [cc-by-sa / 2.0] (geograph.org.uk/p/6128774)

Цель агента – изучить политику оптимального выбора и интерпретации классификаторов на каждом этапе с учетом уже выявленных особенностей Изображение. Агент изучает своего рода механизм внимания, который можно естественным образом комбинировать с жестким визуальным вниманием, чтобы выбрать правильную область изображения для анализа. На рисунке 1 изображена общая идея взаимодействия аген-

та и образа через пул классификаторов с пространственным механизмом внимания. Мы обнаружили, что одновременное обучение визуальному вниманию и политике выбора классификатора затруднено (из-за взаимной зависимости обеих задач). В экспериментальной части мы концентрируемся только на изучении политики отбора классификаторов.

2 Связанные работы

2.1 Повышение уровня нейронных сетей

Деревья классификации и регрессии, вейвлеты Хаара являются подходящими слабыми учениками для повышения, но усиление CNN менее изучено. Одна из причин заключается в том, что классификаторы CNN, снабженные достаточным количеством обучающих данных, хорошо работают без ансамблевого обучения, в то время как классификация является основной областью повышения. Еще важнее то, что сама нейронная сеть неявно представляет собой ансамбль (где скрытые единицы – слабые ученики, а конечная единица – ансамбль [Мерфи]), будучи более мощной, чем поэтапная аддитивная модель (на которой полагается обычное усиление). Могими и Ли [Mohammad Moghimi and Li, 2016] применяют GD-MC Boosting [Saberian and Vasconcelos, 2011] к CNN и показывают, что это предпочтительнее пакетного обучения для ансамблевого обучения с CNN. В [Mosca and Magoulas, 2017] авторы утверждают, что случайная инициализация сети на каждом этапе повышения не требуется; они выступают за перенос веса с предыдущего шага повышения. Лю и др. [Liu et al., 2018] используют для маркировки данных для онлайн-повторного изучения каскада сильных классификаторов с функциями Хаара.

2.2. Динамически конфигурируемые нейронные сети.

Во многих исследованиях изучаются способы расширения нейронных сетей с помощью эффективного динамически конфигурируемого графа вычислений. Одна из целей – сэкономить вычислительные ресурсы за счет выделения сложных и простых примеров. Graves [Graves, 2016] модифицирует архитектуру рекуррентной нейронной сети, чтобы обеспечить адаптивное время вычислений (ACT). Фигурнов и др. [Фигурнов и др., 2016] использовали ACT в остаточных блоках сверточных нейронных сетей и применили их для обнаружения объектов. В [McGill and Perona, 2017] сеть решает, продолжает ли она обрабатывать изображение с помощью сигналов «стоп» и «идти». Процесс классификации инкапсулирован в единую сетевую архитектуру, которая разделяет внутреннее представление отдельных подмодулей. В отличие от нашего подхода, «функция выбора классификатора» (определенная в разделе 3) не

может быть изучена явно во всех случаях. В [Neshatpour et al., 2018] последовательно запускаются несколько отдельных сетей разного размера; классификация останавливается на произвольном шаге на основе предполагаемой уверенности. Каждая сеть берет отдельный поддиапазон, сгенерированный дискретным вейвлет-преобразованием входного изображения. Первые сети работают с более низким разрешением, поэтому потребляют меньше вычислительных ресурсов, чем последующие. Подобный эффект «грубого анализа» может быть достигнут с помощью быстрых саккадических движений в механизме жесткого зрительного внимания, которому можно научиться с помощью обучения с подкреплением. Первая работа в этом направлении – «Рекуррентное зрительное внимание» (RAM) [Mnih et al., 2014]; на каждом временном шаге агент наблюдает только часть изображения и управляет направлением зрения, чтобы сосредоточиться на наиболее информативных областях. [Liu et al., 2018] показали, что оперативную память можно улучшить с помощью динамического вычислительного времени (DT-RAM), предоставив возможность сети генерировать сигнал остановки; в среднем DT-RAM требует меньше шагов, чтобы обеспечить такие же или лучшие результаты на MNIST. В [Bellver et al., 2016] и [Wang et al., 2017] агент учится управлять не только положением, но и размером окна, что позволяет фокусироваться на объектах разных размеров. Кроме того, в [Wang et al., 2017] агенты-наблюдатели VGG имеют пространство функций вместо необработанных пикселей. Однако жесткое визуальное внимание не подразумевает разветвления внутренней структуры вычислений, что и является целью нашего исследования. Концептуально близкими к принципу наименьшего действия являются сети улучшения изображений с динамически конфигурируемыми вычислениями [Yu et al., 2018, 2019]. Их ключевая идея заключается в том, что некоторые части изображения однородны и их легче удалить, поэтому их следует обрабатывать иначе, чем безмолвные. Yu и др. [Yu et al., 2018, 2019] адаптируют обучение с подкреплением и обучают различные цепочки инструментов, которые может использовать агент. В [Huang et al., 2017] авторы аналогичным образом учат агента пропускать слои нейронных сетей в задаче отслеживания визуальных объектов. В последнее время механизм самовнимания, обеспечиваемый трансформерами, показывает многообещающие результаты в применении к проблемам компьютерного зрения [Алексей Досовицкий, Карион и др., 2020], хотя эти работы концентрируются на преимуществах производительности, а не адаптируют жесткое внимание; принцип наименьшего действия игнорируется.

2.3 Мета-обучение

Проблема обучения политике для выбора алгоритма из списка известна как задача выбора алгоритма (AS) [Rice, 1976]. Недавно представленная конфигурация динамических алгоритмов (DAC) [Biedenkapp1 et al., 2020] в отличие от традиционной AS предлагает использовать итеративный характер реальных задач, когда агент должен итеративно раскрывать важные детали конкретного примера. Биденкапп и др. [Biedenkapp1 и др., 2020] формулируют проблему как контекстный марковский процесс принятия решений (контекстный MDP), исходя из того факта, что контекст играет решающую роль в точной конфигурации. Они показывают, что обучение с подкреплением – надежный кандидат для получения политик конфигурации: оно превосходит стандартные подходы к оптимизации параметров. Само по себе разреженное ансамблевое обучение использует итеративный характер, поэтому наш подход можно рассматривать как частный случай DAC: контекст – это конкретное изображение, вознаграждение учитывает экономию вычислений, пространства действий и наблюдения имеют определенные формы. Эти аспекты приводят к другой архитектуре агента и функции потерь, чем в [Biedenkapp1 et al., 2020]. Наша установка также относится к контекстным бандитам, но вместо того, чтобы просто изучать политику выбора действий, алгоритм должен интерпретировать ответ бандита (классификатора). В этой работе мы создаем набор классификаторов, которые полезны для агента, вместо того чтобы изучать один классификатор с подмодулями внутри. Введение недифференциальных операций может показаться чрезмерным усложнением, поскольку контролируемый сигнал богаче, а процедура обучения проще. Но с помощью отдельных модулей мы можем контролировать, чтобы оптимальная политика использовала динамически конфигурируемые вычисления, а «функция выбора» лучше усваивалась с помощью сигнала подкрепления [Mnih et al., 2014]. Редкое ансамблевое обучение позволяет плавно итеративно увеличивать сложность без переобучения с нуля, потому что сам агент можно рассматривать как «инструмент»; это может помочь в создании систем, которые постепенно усложняются. Мы рассматриваем наш вклад как двойной:

1. Мы формулируем проблему обучения разреженному ансамблю, основанную на принципе наименьшего действия, как частный случай конфигурации динамического алгоритма.

2. Мы предлагаем архитектуру сети классификатора наименьшего действия и должным образом спроектированную функцию потерь для решения указанной проблемы.

Мы показываем экспериментально, что классификатор наименьшего действия имеет преимущество перед обычным ансамблевым обучением (суммированием) при ограничениях вычислительных затрат.

3 Классификация по наименьшему действию

В предыдущей работе [Малашин, 2019] мы показали, что вычислительно эффективный ансамбль классификаторов при некоторых предположениях должен реализовывать две ключевые функции:

1. Функция выбора классификатора Ф1: S (t) 迸 → aˆ (t).

2. Функция обновления состояния Ф2: {S (t), y (t)} 覞 → S (t+1).

В случае, когда S – внутреннее (скрытое) представление текущего статуса задачи на шаге t, y (t) – это выбранный ответ классификатора, aˆ (t) – «ключ» (индекс) классификатора. Функция выбора классификатора принимает состояние в качестве входных данных и создает «ключ» классификатора в пуле. Назначение функции обновления состояния включает в себя информацию об ответе классификатора в представлении состояния. Проблема может быть представлена марковским процессом принятия решения, показанным на рисунке 2 [Малашин, 2019].

Из-за взаимной зависимости ответов классификаторов найти оптимальные Ф1 и Ф2 сложно, как и в исходной задаче классификации, но приближения могут быть изучены в системе обучения с подкреплением со следующими элементами:

1. Среда – это а) образ и б) пул классификаторов.

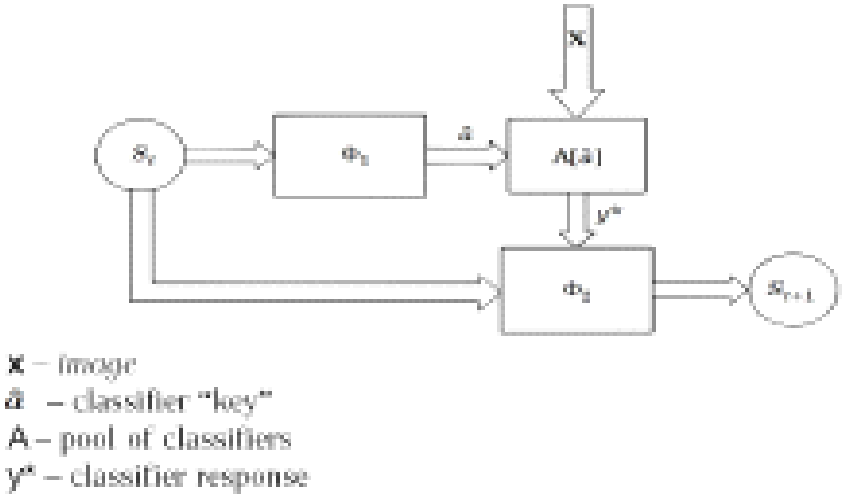


Рисунок 2: Марковская диаграмма процесса принятия решений [Малашин, 2019]

2. Пространство действий состоит из а) индексов классификаторов в пуле и б) прогноза (метки или распределения вероятностей по классам).

3. Наблюдение – это ответы классификаторов.

4. Эпизод классифицирует одно изображение.

5. Вознаграждение учитывает точность решения и вычислительную сложность выбранных классификаторов.

Мы можем рассматривать этот подход как «разреженное» сложенное обобщение [Wolpert, 1992], когда метаученик – это агент, а базовые модели – это CNN. Мы хотим, чтобы учащийся присвоил нулевой вес большинству прогнозов классификаторов, но точные «нули» обнаруживаются итеративно и индивидуально для каждого изображения. По всему распределению выборки тяжелые в вычислительном отношении классификаторы следует обнулять чаще, чем облегченные классификаторы. Это могло бы улучшить обобщение, потому что большие модели часто имеют тенденцию к чрезмерной подгонке.

3.1 Пул классификаторов

Чтобы изучить политики агента, нам нужно создать начальный пул классификаторов, с помощью которого агент сможет взаимодействовать с изображением. Интуитивно желательными свойствами классификаторов являются декоррелированные ответы и вычислительная изобилие архитектур. Мы рассматриваем два типа классификаторов:

1. CNN, полученные путем итеративного увеличения веса изображений, которые были неправильно классифицированы на предыдущем шаге (повышение).

2. CNN изучили различные подмножества классов.

Эти классификаторы обеспечивают хорошую вариабельность сетевых ответов. Повышение Целью обычного повышения квалификации является объединение комитета f, который имеет следующую форму:

$$f(\mathbf{x}) = \sum_{m=1}^{M} w_m f_m(\mathbf{x}), \quad (1)$$

где fm – m-й слабый ученик, а wm – его масса.

Boosting неявно предполагает, что разные классификаторы из комитета сосредотачиваются на разных примерах. Следовательно, {fm} может обеспечить хорошую среду для агента, цель которого – изучить политику, которая избегает использования каждого классификатора для каждого изображения. Мы реализовали BoostCNN [Mohammad Moghimi and Li, 2016], который выполняет оптимизацию градиентным спуском в функциональном пространстве с подходом GD-MC.

Могими и др. Показывают, что GD-MC предпочтительнее для CNN, чем мешковина. Но согласно нашим экспериментам, преимущество BoostCNN в их экспериментах на CIFAR-10 можно объяснить недостаточной адаптацией отдельных сетей во время одной итерации упаковки. Мы оптимизировали некоторые параметры и пришли к выводу, что пакетирование превосходит BoostCNN в этой задаче. Только когда очень

мало итераций повышения (например, 2), BoostCNN иногда предоставляет лучший комитет. Мы приводим более подробную информацию в приложении A. Мы также экспериментировали с Multi-Class Adaboost SAMME [Zhu et al., 2009], который повторно взвешивает обучающие примеры после каждой итерации повышения. SAMME поддерживает произвольную функцию потерь (не только со среднеквадратичной ошибкой), включая кросс-энтропию, обычно используемую для классификации. Но эксперименты показали, что процедура взвешенного обучения плохо сходится для CNN из-за большой дисперсии весов после каждой итерации повышения. Проблему можно решить, сформировав обучающую выборку в соответствии с повышающими весами (Adaboost.M2), но мы не исследовали этот подход. В [Mosca and Magoulas, 2017] авторы предлагают последовательное увеличение глубины сети на каждой итерации повышения. Мы попытались расширить подход, заморозив веса, полученные на предыдущей итерации повышения. В этом случае возможности классификатора на t – 1-й итерации повышения могут быть использованы без повторных вычислений в более глубоком классификаторе t. Однако мы заметили, что без тонкой настройки всех уровней точность комитета не улучшается от итерации к итерации. Мы экспериментировали с идеей неполноценного использования сетей в первой итерации повышения, и влияние было непоследовательным. Таким образом, в наших экспериментах метод простого пакетирования превосходит обычное усиление классификаторов CNN. В то же время классификаторы, полученные в результате пакетирования, не обладают специфичностью, необходимой для изучения способности агента производить контекстно-зависимую последовательность действий. Классификаторы обучаются с разными подмножествами классов. Подход к обучению классификаторов по разным подмножествам классов гарантирует конкретность и (по крайней мере, частичную) декорреляцию ответов. Как отрицательное последствие сокращение пространства признанных классов приводит к ухудшению градиентов [Малашин, 2016] и, следовательно, вредит обучению. Однако для исследовательских целей разные «задачи» заставляют классификаторов иметь менее коррелированные ответы. Для больших задач может возникнуть естественная специфика различных модулей. Пусть набор данных D состоит из N изображений xi с соответствующими метками yi:

$$\mathbf{D = \{(x_i, y_i), i \in [1, N], x \in X, y \in Y\}.} \quad (2)$$

Подмножества классов $Y\ k \subset Y$ разбивают D на перекрывающиеся наборы данных Dk:

$$\mathbf{Dk = \{(xk, yk) \in D, yk \in Y\ k\}.} \quad (3)$$

Отдельные классификаторы, изученные на каждом Dk, образуют пул классификаторов.

3.2. Классификатор наименьшего действия.

Нейронные сети могут быть хорошими кандидатами для аппроксимации функций Ф1 и Ф2. Мы пришли к классификатору наименьшего действия (LAC), изображенному на рисунке 3. LAC состоит из следующих пяти основных компонентов:

1. Генератор ответа среды, недифференцируемый элемент, который принимает изображение и индекс запрошенного классификатора, и возвращает ответ классификатора.

2. Обновление состояния, реализующее функцию Ф2; на шаге t он вводит скрытый вектор состояния и закодированный ответ классификатора; возвращает новый вектор скрытого состояния.

3. Генератор действий, реализующий функцию Ф1; он вводит скрытый вектор состояния и возвращает «ключ» классификатора.

4. Лицо, принимающее решения, которое вводит скрытый вектор состояния и выводит текущее решение.

5. Кодер ответного действия, который кодирует действие и ответ классификатора в формате, подходящем для обновления состояния.

Архитектура LAC гибка в выборе компонентов. Например, некоторые существующие архитектуры визуального внимания могут быть реализованы как классификаторы LAC путем замены генератора действий сетью политик определения местоположения. Ключевое отличие состоит в том, что LAC использует явно изученные классификаторы CNN, которые могут быть более глубокими, чем те, которые были изучены посредством обучения с подкреплением.

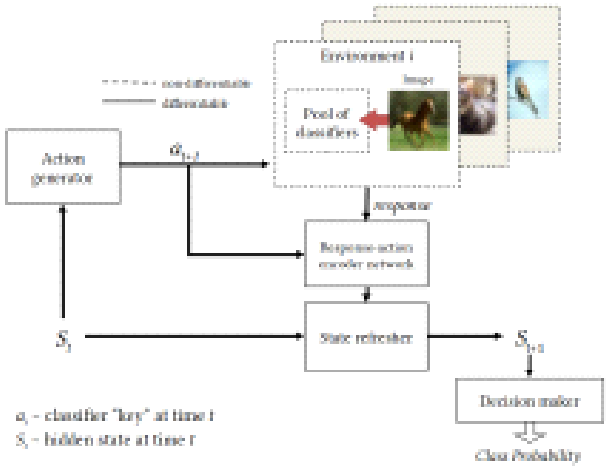


Рис. 3. Сетевая архитектура классификатора наименьшего действия

Блоки сплошных линий дифференцируемы. Образцы изображений взяты из набора данных CIFAR-10, который был собран Алексом Крижевски, Винодом Наиром и Джеффри Хинтоном [лицензия Массачусет-

ского технологического института] (https://www.cs.toronto.edu/ kriz / cifar.html)

State refresher can может быть представлена LSTM или другим типом рекуррентной сети, но в наших экспериментах выяснилось, что более надежные динамически конфигурируемые вычисления обучаются с помощью обновления состояния с короткой памятью. Мы называем LAC с обновлением состояния этого типа LAC-sm. На рисунке 4 показана его структура скрытого состояния и схема алгоритма обновления.

Вместо кодирования состояния с помощью скрытых единиц мы сохраняем ответы каждого классификатора в таблице. Кроме того, мы расширяем представление состояния таблицей масок, которые устанавливаем на единицы при сохранении соответствующего ответа классификатора в первую таблицу. Маски служат четким маркером того, кого уже называли классификаторами; они помогают избежать дублирования действий. Две таблицы составляют скрытое состояние. Таким образом, размер вектора состояния равен N × C × 2, где N – количество откликов, которые необходимо запомнить, а C – размер вектора отклика. В начале «эпизода» обе таблицы заполняются нулями.

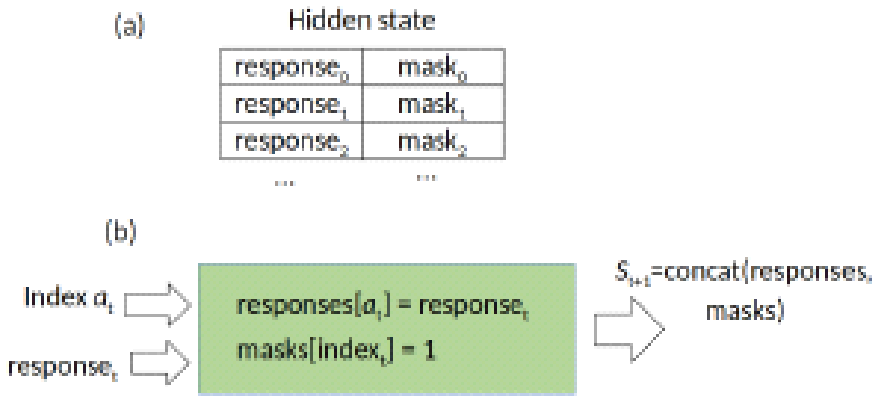


Рисунок 4: Обновление состояния с короткой памятью, структура скрытого состояния (a) и диаграмма его алгоритма обновления (b)

Кодер ответ-действие для LAC-sm – это тождественное отображение ответа классификатора и индекса классификатора. LAC-sm вообще не нуждается в повторяющихся соединениях, поскольку память жестко запрограммирована недифференцируемым образом. Структура генератора действий состоит из двух полностью связанных слоев с активацией RELU. Он возвращает вероятность вызова классификаторов на следующем шаге. Лицо, принимающее решение, имеет три полностью связанных слоя с RELU и возвращает распределение вероятностей по классам изображений.

3.3 Функция потерь

Подобно модели повторяющегося зрительного внимания [Mnih et al., 2014] LAC изучается с помощью гибридной функции:

Loss = γLRL + LossS, (4)

где LRL относится к потере подкрепления, а LossS относится к к стандартной кросс-энтропийной потере (с наземной меткой истинности) γ является гиперпараметром (в наших экспериментах мы используем γ = 0,01). Мы применяем промежуточный надзор [Li et al., 2017], вычисляя контролируемые потери на каждом этапе эпизода. Потеря подкрепления – это сумма потери действия Laction и энтропийного бонуса LH:

LossRL = Laction + αLH (5)

$$L_{action} = \sum_{k}^{K} \sum_{t}^{T} A_{k,t} log(\pi(a_{k,t}|s_{k,t-1}; \theta_a)), \quad (6)$$

где K – количество изображений в пакете, T – количество действий, предпринятых в каждом «эпизоде», π – политика действий, θa – вектор весов генератора действий, Ak, t = Rk, t – b (sk, t – 1) – преимущество, дополнительное вознаграждение R по сравнению с предсказанием базовой сети b, не зависящее от предпринятого действия. эксперименты только с неглубокими одноуровневыми базовыми сетями обеспечивали политику обучения с динамически конфигурируемыми вычислениями. Мы обнаружили, что в качестве альтернативы мы можем использовать более глубокую двухуровневую сеть с выпадением. Формула 6 относится к потерям A3C, поскольку пакеты изображений аналогичны множеству сред. Энтропийный бонус имеет следующий вид:

$$L_H = \sum_{i=1,t=2} log P(a_{i,t}) P(a_{i,t}) + \beta \sum_{k=1,t=1} log P(a_{k,i,t}) P(a_{k,i,t}), \quad (7)$$

где β – гиперпараметр, P (ai, t) – вероятность выбора классификатора i на этапе t, усредненного по всем K изображениям в пакете. Чтобы заставить агента использовать разные классификаторы на разных этапах, в первом члене (7) мы используем энтропию действий, выбранных в ходе каждого эпизода, начиная со второго шага, потому что первый шаг контекстно-независимый. Второй член смягчает прогнозируемое распределение действий, избегая безальтернативных решений во время обучения. В экспериментах β = 10−4. Награда за каждый эпизод имеет вид:

$$R = r - \lambda \sum_{i \in [1,c]} T(a_i), \quad (8)$$

где r равно 1, если изображение классифицировано правильно, и 0, в противном случае T (ai) равно время, необходимое для выполнения

классификатора, связанного с действием ai, $\lambda \geq 0$ – гиперпараметр, а c – количество классификаторов, которые агент использовал перед тем, как дать окончательный ответ. В экспериментах мы использовали фиксированный c, который меньше числа всех классификаторов в пуле, поэтому мы предположили, что $\lambda = 0$.

4 Эксперименты

В экспериментах мы использовали CIFAR-10, который имеет 50000 поездов и 10000 тестов 32 × 32 цветные изображения 10 классов объектов.

4.1. Пул классификаторов

В наших экспериментах мы использовали две простые архитектуры CNN. Первый имеет два конституциональных слоя с 6 и 16 фильтрами, за которыми следуют три полностью связанных слоя с 120, 84 и 10 нейронами соответственно. За каждым сверточным слоем следует max-pooling. Вторая архитектура не имеет полносвязных слоев. Он состоит из трех сверточных слоев с max-pooling (после 1-го слоя) и средним пулом (после 2-го и 3-го слоев). Активация RELU везде кроме топовых сетей. Мы провели случайный поиск параметров обучения и использовали их для каждой сети CNN в нашей среде. Наилучшие результаты в среднем были получены с оптимизатором SGD, геометрическим увеличением, размером пакета 128 и пошаговым расписанием обучения с начальной скоростью 0,01, уменьшающейся в процессе обучения. В таблице 1 представлены шесть классификаторов, которые мы изучили на случайно выбранных подмножествах из 10 исходных классов CIFAR-10; Мы также случайно выбрали сетевую архитектуру для классификатора.

Таблица 1: Пул 1 классификаторов CNN, изученных на подмножестве классов изображений набора данных CIFAR-10

#	Image classes	Arch. type	Test acc (10 classes)
0	{0,1,8,4}	1	35.6
1	{1,2,3,5,6,7,9}	2	57.09
2	{3,2,4}	2	24.64
3	{7,2}	2	18.26
4	{0,1,6,7,8,9}	1	51.02
5	{0,2,3,5}	1	29.49

Классы изображений Arch. тип Испытание в соответствии с (10 классов) 0 {0,1,8,4} 1 35,6 1 {1,2,3,5,6,7,9} 2 57,09 2 {3,2,4} 2 24,64 3 {7, 2} 2 18,26 4 {0,1,6,7,8,9} 1 51,02 5 {0,2,3,5} 1 29,49

4.2 Разрезанное обучение ансамбля

Мы обучаем LAC для 200 эпох с помощью оптимизатора Adam. Скорость обучения снижается в 10 раз после эпох 170 и 190. В первом

эксперименте мы устанавливаем пороговое значение количества действий для LAC-sm.Таблица 2 показывает результаты.

Таблица 2: Производительность LAC на CIFAR-10 с различным количеством действий в пуле 1

Method	Accuracy, %
averaging all responses	62
LAC-sm with 1 action	67.8
LAC-sm with 2 action	75.81
LAC-sm with 3 action	77.81
LAC-sm with 4 action	78.62
LAC-sm with 5 action	79.1
LAC-sm with 6 action	79.29

Точность метода,% при усреднении всех ответов 62 LAC-см с 1 действием 67,8 LAC-см с 2 действием 75,81 LAC-см с 3 действием 77,81 LAC-см с 4 действиями 78.62 LAC-sm с 5 действиями 79.1 LAC-sm с 6 действиями 79.29

Мы пришли к выводу, что агент может включать информацию от нескольких классификаторов, однако неясно, изучает ли агент эффективную контекстно-зависимую функцию выбора классификатора Ф1. Мы сравниваем LAC с контекстно-независимой базовой линией, чтобы убедиться в этом. Во-первых, мы нашли наиболее подходящий алгоритм для суммирования ответов классификаторов из пула. Среди различных алгоритмов машинного обучения нейронная сеть с 5 полносвязными слоями дала лучший результат (79,5% точности), что немного лучше, чем классификатор наименьшего действия с шестью действиями. В экспериментах ниже мы использовали более мелкий многослойный персептрон (MLP) с 3 полносвязными слоями в качестве базовой линии. Он дал почти такой же результат, имея почти вдвое меньшее количество свободных параметров. Результаты всех остальных методов приведены в приложении. Для следующего эксперимента мы формируем пул 2, выбирая классификаторы (с индексами 0,2,3,5), которые дополняют друг друга в данных, на которых они обучались. Затем мы тренируем базовый уровень для каждой комбинации классификаторов в пуле 2 и сравниваем его с LAC в таблице 3.

Таблица 3: Точность по CIFAR-10 с использованием классификаторов из пула 2

Number of classifiers used	MLP (best combination)	LAC-sm
4	72.4	**72.9**
3	69.7	**71.6**
2	66.3	**68.1**
1	59.8	**60.0**

Количество используемых классификаторов MLP (лучшая комбинация) LAC- sm 4 72,4 72,9 3 69,7 71,6 2 66,3 68,1 1 59,8 60,0

Как и ожидалось, исключение любого классификатора снижает точность предоставления ресурсов и конфликт точности в пуле. Таблица 3 показывает, что при вычислительных ограничениях агент учится динамически адаптироваться к содержимому изображения и может с большим запасом нивелировать падение точности. На тестовом наборе LAC-sm с четырьмя разрешенными действиями (LAC-sm-4) использует каждый классификатор равномерно, а LAC-sm-1 использует только лучший из них. Эти политики, естественно, независимы от контекста и должны были дать те же результаты, что и исходные

На удивление, LAC-sm-4 превосходит базовый уровень более чем на 0,5%. Одно из объяснений состоит в том, что промежуточный контроль и шумное обучение обеспечивают эффект регуляризации, похожий на эффект выпадения, заставляя лицо, принимающее решение, угадывать при отсутствии некоторых ответов. Однако у нас есть свидетельства того, что LAC-sm-2 и LAC-sm-3 усвоили контекстно-зависимую политику: они значительно превосходят базовые показатели. На рисунке 5 показано, что LAC-sm-2 использует каждый классификатор с разной частотой, что показывает его способность использовать контекст.

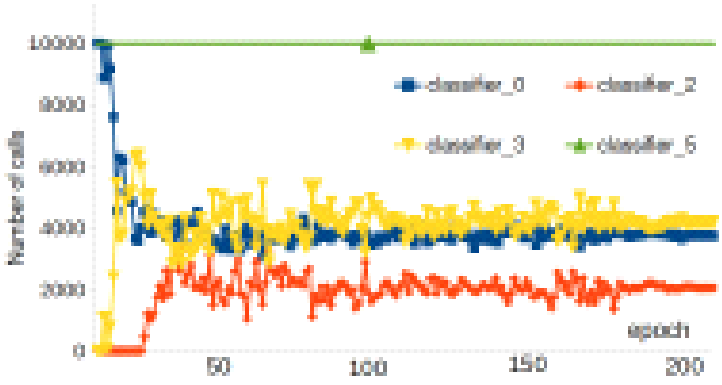


Рисунок 5: Частота вызовов классификатора на тестовом наборе во время обучения LAC-sm-2.

Найти лучшую комбинацию классификаторов для LAC легко, но выявление хорошей контекстно-зависимой политики часто занимает много эпох.

На рисунке 5 показано, что до двадцатой эпохи агент игнорировал классификатор 2. В наших экспериментах динамические вычисления являются ключевым фактором для получения разницы в точности классификатора наименьшего действия и базовой линии, показанной в таблице 3. На рисунке 6 показаны вычислительные графики двух версий. LAC-sm-3 обучены с разными параметрами.

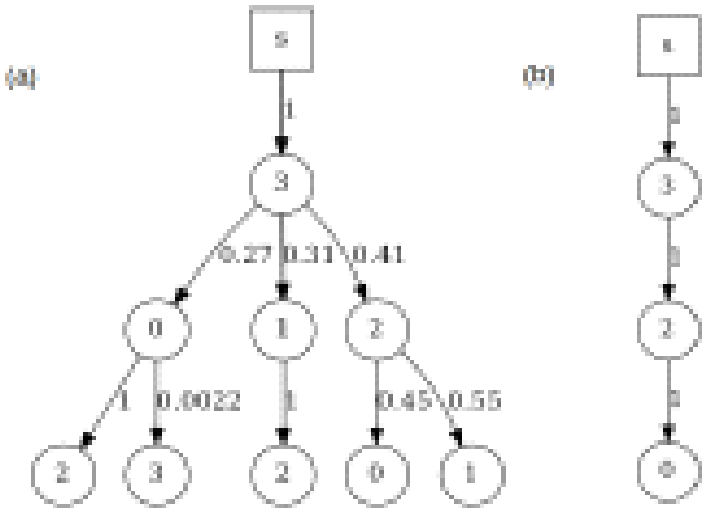


Рисунок 6: Диаграмма, представляющая граф вычислений.

Ребра – это вероятности, узлы – классификаторы, узел «s» относится к началу. (a) LAC-sm-3 обучен с бонусом энтропии и 128 единиц в скрытых слоях лица, принимающего решения (точность 71,6% на тестовом наборе), (b) LAC-sm-3 обучен без бонуса энтропии и 128 единиц в скрытых слоях лица, принимающего решения. слои (точность 69,5%)

Без энтропийного бонуса и чрезмерно большого числа лиц, принимающих решения, классификатор наименьшего действия изучает вычислительный граф, показанный на рисунке 6b, который включает только одну траекторию; он просто игнорирует классификатор №1. Полученная точность теста находится на одном уровне с базовым уровнем, не зависящим от контекста. При правильных параметрах LAC использует пять различных траекторий (рис. 6a) и превосходит базовый уровень почти на 2%

5 Заключение

В этой работе мы формулируем задачу обучения разреженному ансамблю CNN, когда агента учат использовать знания нескольких предварительно обученных классификаторов с учетом их вычислительной сложности. Цель агента – изучить контекстно-зависимую политику для развертывания вычислительного графа таким образом, чтобы обеспечить максимальную ожидаемую точность при условии ограниченного числа действий. Мы представляем архитектуру классификатора наименьшего действия с короткой памятью и соответствующей функцией потерь. Мы показываем экспериментально, что классификатор наименьшего действия изучает политику, которая превосходит традиционный подход группирования классификаторов CNN. Редкое ансамблевое обучение позволяет плавно итеративно увеличивать сложность без переобучения с нуля, потому что сам агент можно рассматривать как

«инструмент»; это может помочь в создании систем, которые постепенно усложняются.

6 Благодарности

Исследование финансировалось Российским научным фондом (проект 19-71-00146).

References

R. Malashin. Principle of least action in dynamically configured image analysis systems. J. Opt. Tech., 86, 2019.

Y.E. Shelepin and N.N. Krasilnikov. Principle of least action, physiology of vision and conditioned reflex theory. Ross. Fiziol. Zh. im. I. M. Sechenova, 89(6).

Y. Shelepin, N. Krasilnikov, G. Trufanov, A. Harauzov, S. Pronin, and Foking A. The principle of least action and viusal perception. In Twenty-ninth European Conference on Visual Perception, volume 35, August 2006.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In CVPR 2001, volume 1, pages I–511–I–518 vol.1, 2001. doi:10.1109/CVPR.2001.990517.

Kevin P. Murphy. Machine learning : a probabilistic perspective. MIT Press, Cambridge, Mass. [u.a.]. ISBN

Mohammad Saberian Jian Yang Nuno Vasconcelos Mohammad Moghimi, Serge Belongie and Li-Jia Li. Boosted convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, Proceedings of the British Machine Vision Conference (BMVC), pages 24.1–24.13. BMVA Press, September 2016. ISBN 1 901725-59-6. doi:10.5244/C.30.24. URL https://dx.doi.org/10.5244/C.30.24.

Mohammad J. Saberian and Nuno Vasconcelos. Multiclass boosting: Theory and algorithms. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 24, pages 2124–2132. Curran Associates, Inc., 2011.

Alan Mosca and George Magoulas. Deep incremental boosting. 08 2017.

Chunsheng Liu, Shuang Li, Faliang Chang, and Wenhui Dong. Supplemental boosting and cascaded convnet based transfer learning structure for fast traffic sign detection in unknown application scenes. Sensors, 18:2386, 07 2018. doi:10.3390/s18072386.

Alex Graves. Adaptive computation time for recurrent neural networks. CoRR, abs/1603.08983, 2016. URLhttp://arxiv.org/abs/1603.08983.

Michael Figurnov, Maxwell D. Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry P. Vetrov, and Ruslan

Salakhutdinov. Spatially adaptive computation time for residual networks. CoRR, abs/1612.02297, 2016. URL http://arxiv.org/abs/1612.02297.

Mason McGill and Pietro Perona. Deciding how to decide: Dynamic routing in artificial neural networks. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Pro-

ceedings of Machine Learning Research, pages 2363–2372, International Convention Centre,

Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/mcgill17a.html.

K. Neshatpour, F. Behnia, H. Homayoun, and A. Sasan. Icnn: An iterative implementation of convolutional neural networks to enable energy and computational complexity aware dynamic approximation. In 2018 Design, Automation Test in Europe Conference Exhibition (DATE), pages 551–556, 2018.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2204–2212. Curran Associates, Inc., 2014.

Miriam Bellver, Xavier Giro-i Nieto, Ferran Marques, and Jordi Torres. Hierarchical object detection with deep reinforcement learning. In Deep Reinforcement Learning Workshop, NIPS, December 2016.

Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In IEEE International Conference on Computer Vision, October 2017.

Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. 04 2018.

Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. arXiv preprint arXiv:1904.10343, 2019.

Chen Huang, Simon Lucey, and Deva Ramanan. Learning policies for adaptive tracking with deep feature cascades. pages 105–114, 10 2017. doi:10.1109/ICCV.2017.

Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. URL https://arxiv.org/abs/2010.11929.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

John R. Rice. The algorithm selection problem. Advances in Computers, 15:65–118, 1976. URL http://dblp.uni-trier.de/db/journals/ac/ac15.html#Rice76.

Andre Biedenkapp1, H. Furkan Bozkurt1andTheresa Eimer, Frank Hutter1, and Marius Lindauer. Dynamic algorithm configuration:foundation of a new meta-algorithmic framework. volume 163, 2020.

D Wolpert. Stacked generalization. Neural Networks, 1992.

Ji Zhu, Saharon Rosset, Hui Zou, and Trevor Hastie. Multi-class adaboost. 2009.

R Malashin. Extraction of object hierarchy data from trained deep-learning neural networks via analysis of the confusion matrix. J. Opt. Tech., 83, 2016.

Zhichao Li, Yi Yang, Xiao Liu, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. CoRR, abs/1703.10332, 2017. URL http://arxiv.org/abs/1703.10332.

Приложения

A BoostCNN

Целью повышения является решение следующей задачи оптимизации:

$$f^* = \min_f R(f) = \min_f \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)), \quad (9)$$

где L (y, yˆ) – некоторая функция потерь, (xi, yi), i ∈ N – обучающие выборки, а комитет f имеет вид (1). Поскольку задача подбора составной функции сложна, бустинг решает проблему последовательно:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + v\beta_m \phi(\mathbf{x}; \theta_m), \quad (10)$$

где θm – параметры модели, β – вес, минимизирующий (9) и 0 <v <1 – параметр усадки. Мы реализовали BoostCNN [Mohammad Moghimi and Li, 2016], который выполняет оптимизацию градиентным спуском в функциональном пространстве с подходом GD-MC. В этом случае функция потерь имеет вид:

$$L(z_i, f(x_i)) = \sum_{j=1, j\neq z} \exp \frac{1}{2}[< y_{z_i}, f(x_i) > - < y_j, f(x_i) >], \quad (11)$$

где zi ∈ 1... M – метка класса, а y – код метки. В соответствии с методами повышения градиента CNN учится воспроизводить градиенты целевой функции в функциональном пространстве с функцией потерь MSE; Коэффициент β находится линейным перебором, минимизирующим (11) вдоль направления fm. Следуя [Mohammad Moghimi and Li, 2016], мы заменили линейный поиск двоичным поиском δR. Мы обнаружили, что влияние линейного поиска на процесс повышения неоднозначно. Мы проиллюстрируем это на наборе данных cifar-10. Для первых экспериментов мы повторно реализовали результаты из [Mohammad Moghimi and Li, 2016] с помощью сети cifar-quick, состоящей из трех сверточных слоев с объединением и активацией RELU, за которыми следуют два полностью связанных слоя. На рисунке 7 показана динамика тренировки. Когда использование линейного поиска оказало существенное влияние, большое значение v может привести к нестабильному обучению сетей, а иногда процесс может отклоняться из-за чрезмерных потерь на сильно взвешенных примерах. Как можно видеть, линейный поиск увеличивает скорость обучения в первых пяти шагах ускорения, но после этого приводит к переобучению. Согласно [Mohammad Moghimi and Li, 2016] GD-MC предпочтительнее пакетного обучения

для ансамблевого обучения с CNN, и мы смогли воспроизвести их результаты с той же сетевой архитектурой, которая использовалась в качестве базового обучающегося. Однако, согласно нашим экспериментам, можно утверждать, что преимущество BoostCNN над бэггингом в их экспериментах было достигнуто исключительно за счет недостаточной адаптации отдельных сетей во время одной итерации бэггинга. Мы оптимизировали некоторые параметры Bagging: использовали большой мешок для выборки с заменой (такой же, как и количество обучающих примеров), увеличили количество эпох на шаг повышения, добавили перенос веса (как [Mohammad Moghimi and Li, 2016] сделал для GD - MC) и сравнил результаты с BoostCNN и неоптимизированным Bagging. Результаты изображены на Рисунке 8a. Видно, что пакетирование превосходит BoostCNN более чем на 1% за двадцать итераций (и почти на 3% лучше, чем результат, полученный в [Mohammad Moghimi and Li, 2016]). Настройка усадки может немного улучшить BoostCNN, но все же в наших экспериментах она подходит для десятой итерации, в то время как упаковка в пакеты улучшается. Вывод верен для разных сетевых архитектур. На рисунках 8b и 8c показаны кривые обучения, когда Resnet-18 используется в качестве слабого классификатора. Бэггинг показывает гораздо меньшую тенденцию к переобучению. Одна интересная находка заключается в том, что (a) потеря MSE в кодовых словах и (б) классическая кросс-энтропийная потеря с однократным горячим кодированием обеспечивает очень похожую динамику обучения для отдельных сетей. Например, resnet-18, обученный в течение 100 эпох, обеспечивал частоту ошибок 7,5% с увеличением изображения и около 14% без увеличения, независимо от того, какую функцию потерь мы использовали. Бэггинг с MSE и кодовыми словами дает немного лучшие результаты, чем бэггинг с кросс-энтропией и однократным кодированием.

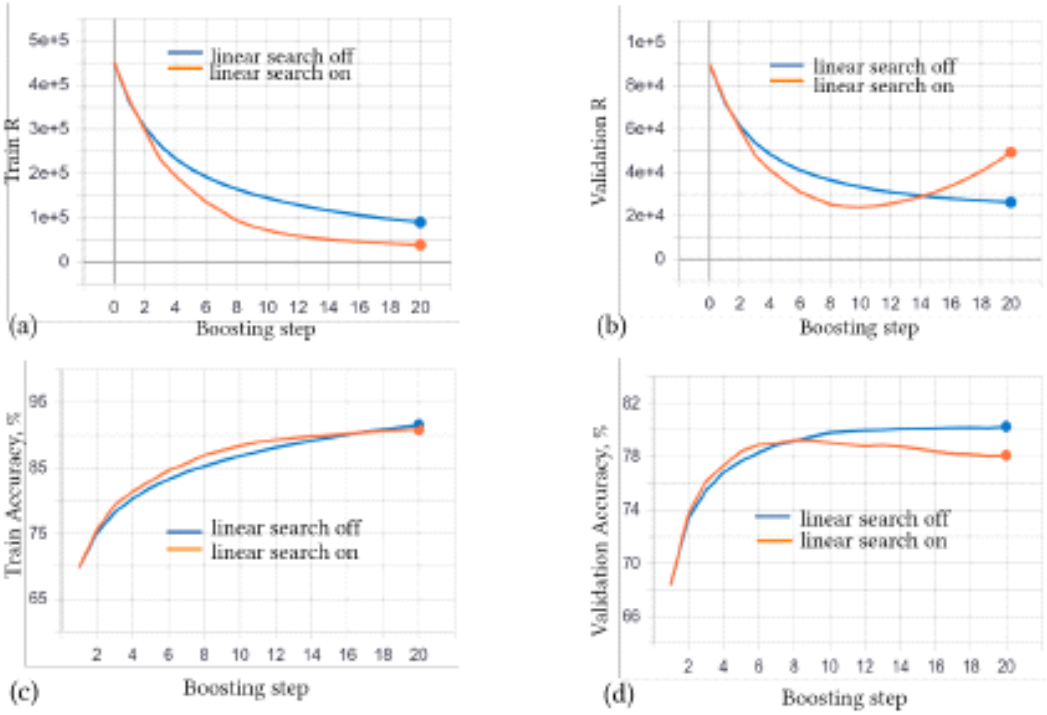


Рис. 7: Кривая обучения BoostCNN с линейным поиском и без него. (a), (b) – целевая функция на наборах поездов и валидации, внизу (c), (d) – точность обучения и валидации

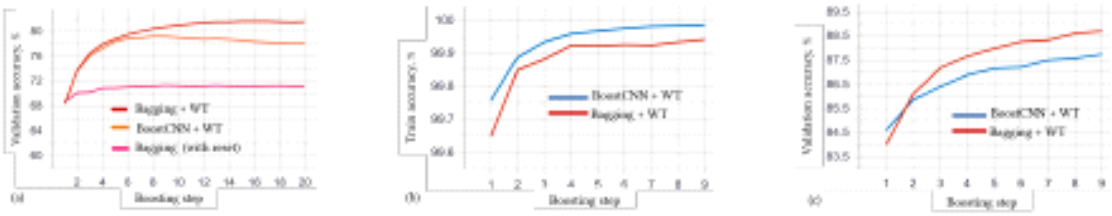


Рисунок 8: Различные методы обучения ансамбля на наборе данных CIFAR с простой CNN в качестве базового обучающегося. (a) точность проверки комитетом BoostCNN и упаковки (со сбросом) воспроизведена с параметрами из [Mohammad Moghimi and Li, 2016] и сравнение с упаковкой с переносом веса. (b) и (c) – кривые обучения для BoostCNN и упаковки с переносом веса, архитектура

В Базовый уровень

Таблица 4 является слабым классификатором, во время обучения не используются никакие дополнения. в суммировании ответов всех классификаторов из таблицы 1.

Таблица 4: Объединение классификаторов пула 1 с разными методами

Method	Test accuracy, %
Adaboost + Decisions trees	0.619
SVM + rbf kernel	0.782
SVM + linear kernel	0.79
RandomForests	0.772
ExtraTrees	0.756
Decision Tree	0.677
5-KNN	0.734
15-KNN	0.755
3-layer MLP	0.794
5-layer MLP	**0.795**

Метод Точность теста,% Adaboost + Деревья решений 0,619 SVM + ядро rbf 0,782 SVM + линейное ядро 0,79 RandomForests 0,772 ExtraTrees 0,756 Дерево решений 0,677 5-KNN 0,734 15-KNN 0,755 3-слойный MLP 0,794 5-слойный MLP 0,795

Сеть с пятью полносвязными слоями и активациями RELU дала лучший результат. Из соображений обобщения в экспериментах мы предпочли более мелкую трехслойную МЛП, которая давала почти такую же точность и в то же время почти в два раза меньше параметров.

НЕЙРОМОРФНАЯ КРЕМНИЕВАЯ ФОТОНИКА

Тейт А.Н., Чжоу Э., Феррейра де Лима Т., Ву А.Х., Нахмиас М.А., Шастри Б.Дж., Прукнал П.Р.

Аннотация

Мы сообщаем о первых наблюдениях интегрированной аналоговой фотонной сети, в которой соединения конфигурируются весовыми банками микрокольца, а также о первом использовании электрооптических модуляторов в качестве фотонных нейронов. Математический изоморфизм между кремниевой фотонной схемой и непрерывной нейронной моделью демонстрируется посредством анализа динамической бифуркации. Используя этот изоморфизм, существующие инструменты нейронной инженерии можно адаптировать к кремниевым фотонным системам обработки информации. Кремниевая фотонная нейронная сеть с 49 узлами, запрограммированная с использованием «нейронного компилятора», смоделирована и, согласно прогнозам, превзойдет традиционный подход в 1960 раз в задаче эмуляции игрушечной дифференциальной системы. Фотонные нейронные сети, использующие кремниевые фотонные платформы, могут получить доступ к новым режимам сверхбыстрой обработки информации для радио, управления и научных вычислений.

ВВЕДЕНИЕ

Свет формирует глобальную основу передачи информации, но редко используется для преобразования информации, хотя и не из-за отсутствия попыток [1–3]. Цифровая оптическая логика сталкивается с фундаментальными физическими проблемами [4]. Аналоговые оптические сопроцессоры столкнулись с двумя основными экономическими проблемами: оптические системы никогда не обеспечивали конкурентоспособность в производстве и не удовлетворяли в достаточной мере общие потребности в обработке. Начавшиеся изменения в спросе и предложении на фотонику могут вызвать возрождение оптической обработки информации. Развивающаяся индустрия кремниевой фотонной интеграции обещает обеспечить производственную экономию, обычно предназначенную для микроэлектроники. Несмотря на прочный спрос на приемопередатчики для центров обработки данных [5], индустриализация фотоники повлияет на другие области приложений [6]. Промышленные экосистемы микротехнологий продвигают разработку дорожных карт [7], стандартизацию библиотек [8, 9] и расширенную доступность [10], все это может открыть принципиально новые направления исследований

в области крупномасштабных фотонных систем. Реализованы крупномасштабные устройства управления лучом [11] и предусмотрены внутрикристальные сети связи [12, 13]; однако возможности масштабируемых кремниевых систем обработки фотонной информации остаются в значительной степени неизученными. Одновременно с этим фотонные устройства нашли ниши аналоговой обработки сигналов, в которых электроника больше не может удовлетворять требованиям к полосе пропускания и реконфигурируемости. Примером такой ситуации является радиочастотная (RF) обработка, при которой внешние интерфейсы ограничиваются радиочастотной электроникой, аналого-цифровыми преобразователями (АЦП) и процессорами цифровых сигналов (DSP) [14, 15]. В ответ RF-фотоника предложила соответствующие решения для настраиваемых RF-фильтров [16, 17], самого АЦП [18] и простых задач обработки, которые могут быть перенесены из DSP в аналоговую подсистему [19–22]. Радиочастотные фотонные схемы, которые могут быть преобразованы из волокна в кремний, вероятно, принесут экономическую выгоду крупносерийного производства. Более того, сочетание высокопроизводительной аналоговой фотоники с беспрецедентной возможностью для крупномасштабной системной интеграции может привести к принципиально новым концепциям, выходящим за рамки того, что можно рассматривать в волоконно-оптических системах. Масштабируемая аналоговая обработка требует математической основы, которая обеспечивает правила программирования параметров устройства для получения желаемого поведения системы. Из моделей, которые могут восполнить этот пробел, одни из самых хорошо изученных и широко распространенных – это модели нейронных сетей. Системы, которые математически изоморфны моделям нейронных сетей (т.е. нейроморфные системы), могут раскрыть это богатство существующих алгоритмов [23, 24], доказательств [25, 26] и инструментов [27, 28]. Эта стратегия пережила недавнее возрождение в нетрадиционных областях вычислений [29–31] и машинного обучения [32–35], опора которых на существующую теорию позволяет сосредоточиться на энергоэффективных архитектурах и приложениях для работы с большими данными соответственно. Противоположный подход, резервуарные вычисления, недавно вызвал интерес сообщества фотоники [36–39]. Резервуарные методы основаны на различении желаемого поведения из большого количества немоделированных сложных динамик, черпая вдохновение из определенных свойств мозга (например, аналоговых, распределенных), вместо использования строгого изоморфизма с моделью. В [40] была предложена совместимая с кремнием архитектура фотонных нейронных сетей, называемая «широковещательная передача и вес». В этой архитектуре, показанной на рисунке 1, выходу каждого узла назначается несущая с

уникальной длиной волны, которая мультиплексируется с разделением по длине волны (WDM) и транслируется другим узлам. Входящие сигналы WDM взвешиваются с помощью реконфигурируемых фильтров с непрерывным значением, называемых весовыми банками микрокольца (MRR) [41–43], а затем суммируются с помощью определения общей мощности. Эта электрическая взвешенная сумма затем модулирует соответствующий канал WDM. Нелинейная электрооптическая передаточная функция, такая как лазер на пороге или, в этой работе, насыщенный модулятор, обеспечивает нелинейность, необходимую для функциональности нейрона. Здесь мы сообщаем о первой экспериментальной демонстрации интегрированной фотонной нейронной сети. Аналоговые межкомпонентные соединения WDM в сети реконфигурируются кремниевыми банками весов MRR, чтобы вызвать качественный поведенческий переход участков (то есть бифуркаций), которые служат наблюдаемыми «отпечатками пальцев» лежащей в основе динамики [44]. Воспроизведение нейроморфных бифуркаций в результате переконфигурирования весов MRR указывает на математический изоморфизм между изготовленным образцом и двухузловой моделью рекуррентной нейронной сети с непрерывным временем (CTRNN) [45]. Этот результат предполагает, что инструменты программирования для CTRNN могут быть применены к более крупным кремниевым фотонным нейронным сетям. Кремниевые фотонные сети CTRNN могут предоставить новые возможности обработки в реальном времени в ряде областей приложений. Нейронные модели с непрерывным временем применялись для спектрального анализа [46], оценки канала с расширенным спектром [47] и управления антенной решеткой [48], и существует настоятельная потребность в реализации этих функций в более широких полосах частот с меньшим энергопотреблением. Кроме того, методики, разработанные для аудиоприложений, такие как подавление шума [24], вероятно, могут быть использованы для радиочастотных проблем, если они будут реализованы на сверхбыстром оборудовании. Подмножество CTRNN, сети Хопфельда [49], широко используются в задачах математического программирования и оптимизации [23]. Аппаратные фотонные ускорители могут решить эти и другие проблемы в научных вычислениях. В качестве примера адаптации методологии проектирования CTRNN к нейроморфной кремниевой фотонике мы моделируем сеть с 49 модуляторами, запрограммированную с использованием «нейронного компилятора» [27]. Игрушечная задача эмуляции дифференциального уравнения используется для сравнения фотонного подхода с подходом обычного ЦП, прогнозируя hTahridswwaroerkacaclseloerparteisoenntosf t1h, 9e6f0 r × st. исследование фотонных нейронов с помощью модуляторов, в отличие от активных лазерных устройств. За последние несколько лет интерес к ла-

зерным устройствам с нейроноподобным поведением всплеска возрос [50, 51], но экспериментальные работы пока сосредоточены на изолированных нейронах [52–54], линейной цепи возбуждаемых MRR [55]., и сеть без возможности реконфигурирования [56]. Этот сетевой разрыв можно объяснить проблемами реализации компактных и настраиваемых спектральных фильтров с низкими потерями в активных платформах III / V, необходимых для нейронов лазерного класса. Модуляторы совместимы с кремниевыми фотонными платформами, на которых также могут размещаться весовые банки MRR. Таким образом, в то время как нейроны лазерного класса предоставляют более широкие возможности обработки за счет динамики импульсов, нейроны класса модулятора может быть проще в изготовлении, при этом они все еще обладают огромным набором функций CTRNN.

МЕТОДЫ

Экспериментальная установка и изображение весовой сети MRR показаны на рис. 2. Образцы были изготовлены на пластинах кремний-на-изоляторе (КНИ) на заводе нанотехнологий в Вашингтоне с помощью группы быстрого прототипирования SiEPIC Ebeam [10]. Толщина кремния составляет 220 нм, а толщина скрытого оксида (BOX) составляет 3 мкм. WG шириной 500 нм были сформированы литографией Ebeam и полностью протравлены до BOX [57]. После оксидной пленки W D M MRR весовой ряд BPD * LD MZM MRR весовой ряд BPD * LD MZM слои Ti / W и Al. Омический нагрев в нитях Ti / W вызывает термооптические резонансные сдвиги длины волны в весах MRR.

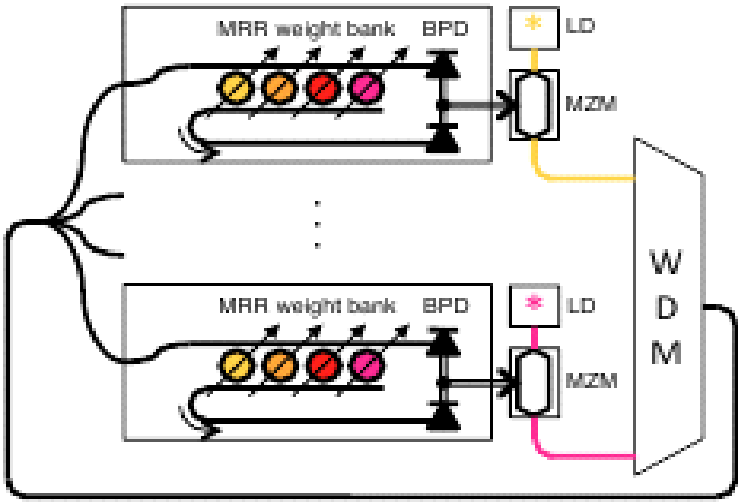


Рисунок 1. Концепция сети вещания и веса STAR с модуляторами, используемыми в качестве нейронов. MRR: микрокольцевый резонатор, BPD: симметричный фотодиод, LD: лазерный диод, MZM: модулятор Маха-Цендера, WDM: мультиплексор с разделением по длине волны. (3 мкм),

Образец устанавливается на столике юстировки со стабилизированной температурой и соединяется с 9-волоконной антенной решеткой с помощью фокусирующих решетчатых элементов связи субволнового диапазона [58]. Сеть в конфигурации STAR для широковещательной передачи состоит из 2 банков весов MRR, каждый с весами MRR радиусом 4x 10 мкм. Каждый банк весов MRR калибруется с использованием метода, представленного в работе. [41, 42]: процедура измерения в рабочем состоянии выполняется для определения моделей термооптических перекрестных помех и пропускания через МРЗ-фильтр через край. Во время этой фазы калибровки электрические соединения обратной связи отключаются, и набор каналов длины волны переносит набор линейно разделяемых обучающих сигналов. После калибровки пользователь может указать желаемую матрицу весов, а модель управления вычисляет и применяет соответствующие электрические токи. Взвешенные выходные сигналы сети обнаруживаются на микросхеме, и электрические взвешенные суммы управляют волоконными модуляторами Маха-Цендера (MZM). Обнаруженные сигналы подвергаются фильтрации нижних частот на частоте 10 кГц, представленной символами конденсаторов на рис. 2. Фильтрация нижних частот используется для нарушения динамических характеристик с временной задержкой, которые возникают, когда задержка обратной связи намного больше постоянной времени состояния [59]. В этой установке с сетью на кристалле и нейронами-модуляторами вне кристалла, динамика с задержкой по волокну будет мешать динамическому анализу CTRNN [60].

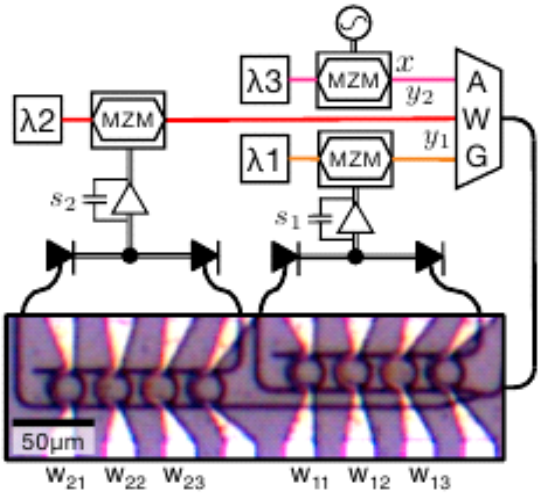


Рисунок. 2. Экспериментальная установка с двумя нейронами MZM и одним внешним входом, мультиплексированными по длинам волн в решетке из массивов волноводов (AWG) и подключенными к встроенной сети широковещательной передачи и взвешивания (на фото). Рекуррентная сеть 2x2 конфигурируется весами MRR, w11, w12 и т. д. Состояние нейрона представлено напряжениями s1 и s2 на трансимпедансных усилителях с фильтром нижних частот.

MZM, модулирующие различные длины волн λ1 и λ2 с выходными сигналами нейрона y1 (t) и y2 (t), соответственно. Электрооптическая передаточная функция MZM служит насыщающей нелинейностью y = σẎ (s), связанной с нейроном непрерывного времени. Третья длина волны λ3 несет внешний входной сигнал x (t), полученный от генератора сигналов. Все оптические сигналы (x, y1 и y2) мультиплексируются по длине волны в решетчатой решетке волноводов (AWG), а затем возвращаются во встроенную трансляцию STAR, состоящую из расщепляющихся Y-переходов [61].

БИФУРКАЦИЯ

CUSP Модель CTRNN описывается набором обычных дифференциальных уравнений, связанных через матрицу реконфигурируемых весов W.

d□s (t) dt = W∘y (t) – □ s (t) τ + wℊ inx (t) (1)

∘y (t) = σ ൿ [□ s (t)] (2)

где s (t) – переменные состояния с постоянными времени τ, y (t) – выходы нейронов, w□in – входные веса, а x (t) – внешний вход. σ ⅄ – насыщающая передаточная функция, связанная с каждым нейроном. В этом случае передаточная функция нейрона ввода-вывода соответствует синусоидальной электрооптической передаточной функции MZM, которая принимает электронно-взвешенную сумму и выдает новый оптический сигнал. Для анализа эта синусоидальная передаточная функция была снижена, и Тейлор расширил ее до ее первого нелинейного члена.

σ ⁐ (s) ≈ σ (s) = αs – κs3 (3)

где σ (s) – приближение со сглаживанием, а α и κ – положительные коэффициенты. Математический анализ динамических систем начинается с изучения фиксированных состояний (где s˙ = 0) и влияния параметров на их поведение. Изменения числа или стабильности фиксированных точек, называемые бифуркациями, могут использоваться в качестве наблюдаемых идентификаторов основной динамики. Простейший случай (1) – одиночный узел с собственной обратной связью:

$$0 = W_F \sigma(s^*) - \frac{s^*}{\tau} + w_{in} x^* \qquad (4)$$

$$= \kappa W_F s^{*3} - \left(\alpha W_F - \tau^{-1}\right) s^* - w_{in} x^* \qquad (5)$$

nd x ∗ – установившиеся скалярные значения состояния и входа нейрона соответственно.

$$\kappa W_F s^{*2} = \alpha W_F - \tau^{-1} \qquad (6)$$

$$s^*_{(1)} = 0; \qquad s^*_{(2,3)} = \pm\sqrt{\frac{\alpha}{\kappa}\frac{W_F - W_B}{W_F}} \qquad (7)$$

Одиночный узел с весом с обратной связью, WF, может демонстрировать несколько бифуркаций между моностабильным и бистабильным режимами, которые здесь выводятся. Когда вход равен нулю, стационарные решения имеют вид

Полученная в результате знакомая S-образная бистабильная кривая нанесена на ось x-y на рис. 3. Три корня s ∗ существуют, когда вес обратной связи зафиксирован выше значения бифуркации вил. Ребра этого бистабильного режима называются седло-узловыми точками, потому что неустойчивое среднее седло и один из стабильных узлов аннигилируют друг друга. Находятся седло-узловые точки s ∗ SN, в которых производная x ∗ в (8) равна нулю по s ∗.

$$s^*_{SN} = \pm\sqrt{\frac{\alpha}{3\kappa}\frac{W_F - W_B}{W_F}} \qquad (9)$$

Заменив это в (8), приходим к уравнению для каспа

$$x^*_{SN} = \pm\frac{2}{3}\frac{\alpha^{3/2}}{w_{in}\sqrt{3\kappa W_F}}(W_F - W_B)^{3/2} \qquad (10)$$

который проецируется на ось WF -x на рис. 3. Бифуркация, иначе говоря, складка, более информативна, чем бифуркации вил или седло-узел, потому что она описывается только со ссылкой на два параметра, в то время как другие бифуркации могут происходить в системах одного параметра.

Результаты

Теоретическая модель каспа экспериментально наблюдается на установке, описанной в гл. II. Внешний генератор сигналов вводит треугольную волну с коэффициентом заполнения 50% на частоте 3 кГц, они нанесены друг на друга на рис. 3. Поскольку эта поверхность многозначна, она разделена на нижнюю (синюю) и верхнюю (красная) ветви, соответствующие восходящим и падающим данным соответственно. Затем данные согласуются с теоретической моделью стационарной поверхности возврата, описываемой формулой (5). Параметры τ, α и κ выбраны так, чтобы минимизировать общую среднеквадратичную ошибку между моделью и поверхностями данных. Построенные точки данных снова интерполируются с поверхности данных в соответствующей плоскости.

Модель наилучшего соответствия имеет пик на уровне WB = 0,54. На рис. 3 поверхности данных и подгонки модели срезаны при WF = 0,85 и нанесены в плоскости x-s, чтобы получить бистабильную кривую, описываемую формулой (8); нанесенные на график точки данных берутся из записанного сигнала при соответствующем весе.

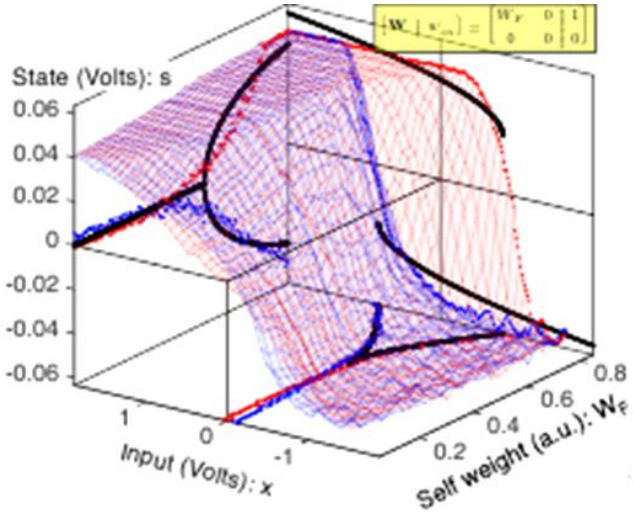


Рисунок. 3. Бифуркация бугорка в одном узле с весом обратной связи, WF, внешним входом, x и нейросинаптическим состоянием, s. Синяя / красная сетка: увеличение / уменьшение ввода. Синие и красные точки данных берутся на срезах поверхностей данных; толстые черные кривые – соответствующие срезы теоретической модели. Тонкими линиями показаны плоскости, на которых делаются срезы. и вес обратной связи узла 1 параметризован как w11 = WF. Когда весовой коэффициент обратной связи проходит через 500 точек от 0,05 до 0,85, осциллограф фиксирует входной сигнал x и нейросинаптический сигнал s1.

Поверхность снова разрезается в точке x = 0 и строится в плоскости s-WF, чтобы получить кривую вил, описываемую формулой (7); нанесенные на график точки данных интерполируются в соответствующей плоскости. Наконец, поверхность разрезается при s = 0 и наносится на плоскость x-WF, чтобы получить кривую возврата, описываемую (10). Экспериментальное воспроизведение бифуркаций вил, бистабил и каспов демонстрирует изоморфизм между одноузловой моделью и кремниевой фотонной системой. Открытие области между восходящей и нисходящей поверхностями данных характерно для бистабильности. Что еще более важно, переходы между моностабильным и бистабильным режимами медитируются конфигурациями банка весов MRR. Граница перехода точно повторяет форму возврата. Неидеальности в подгонке видны в вилах и бистабильных срезах, несмотря на их качественное воспроизведение количества и тенденций роста стабильных точек. Эти неидеальности могут быть приписаны жесткому насыщению электрического трансимпедансного усилителя, когда входное напряжение и вес

обратной связи высоки. Одноузловая бифуркация куспида не является демонстрацией многоузловой сети. Измерения стабильного каспа действительно служат контролем, демонстрирующим отсутствие паразитных запаздывающих колебаний, как это наблюдалось в [60]. В следующем разделе мы изучаем колебательную бифуркацию, которая может происходить только в многоузловых динамических системах.

БИФУРКАЦИЯ ХОПФА

Динамические системы способны колебаться, если существует замкнутая орбита (также известная как предельный цикл) в пространстве состояний, которая, следовательно, должна превышать одно измерение. Бифуркация Андронова-Хопфа (Хопфа) возникает, когда устойчивая фиксированная точка становится нестабильной, порождая устойчивый предельный цикл. Бифуркации Хопфа далее характеризуются колебаниями, которые приближаются к нулевой амплитуде и ненулевой частоте вблизи точки бифуркации [62]. Бифуркация Хопфа может возникнуть в двухузловой нейронной сети, описываемой (1), при определенных условиях. Мы фиксируем о ff-диагональные веса асимметрично, так что t − w12 = w21 = 1, и параметризуем диагонали так, что w11 = w22 = WF. Как и раньше, WF используется для обозначения веса собственной обратной связи. В этой формулировке всегда существует одно и только одно установившееся состояние при ♈s = 0. Чтобы исследовать его устойчивость, мы линеаризуем систему вокруг этой точки, чтобы получить матрицу Якоби, собственные значения которой указывают на устойчивость в неподвижной точке.

$$\mathbf{J} = \frac{d}{ds}\left(\frac{ds}{dt}\right) = \alpha \begin{bmatrix} W_F - (\alpha\tau)^{-1} & -1 \\ 1 & W_F - (\alpha\tau)^{-1} \end{bmatrix} \quad (11)$$

which has eigenvalues

$$\lambda = W_F - (\alpha\tau)^{-1} \pm i \quad (12)$$

Мнимая часть пары собственных значений указывает на колебательное поведение. Действительная часть собственного значения меняет знак при бифуркационном весе WB = (ατ) −1. В этом случае, когда единственное решение с фиксированной точкой становится неустойчивым, вместо новых устойчивых состояний возникает устойчивый предельный цикл. Вблизи порога мы можем принять круговую форму предельного цикла, чтобы смоделировать его ожидаемую амплитуду A и частоту ω

$$s_1(t) = A\sin(\omega t); \quad s_2(t) = A\cos(\omega t) \quad (13)$$

В точках, где ωt кратно 2π, производная s2 по времени равна нулю. Рассматривая уравнение sÀ2 из (1)

$$\left.\frac{ds_2}{dt}\right|_{\omega t=2\pi m} = 0 = \sigma(0) + W_F\sigma(A) - \tau^{-1}A \quad (14)$$

$$= \alpha(W_F - W_B)A - W_F\kappa A^3 \quad (15)$$

$$A = \sqrt{\frac{\alpha}{\kappa}\frac{W_F - W_B}{W_F}} \quad (16)$$

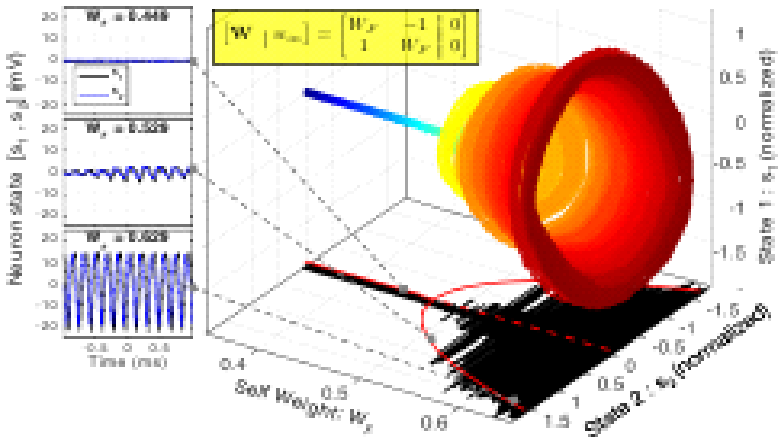


Рисунок. 4. Бифуркация Хопфа между устойчивым и колеблющимся состояниями. O ff- диагональные веса асимметричны, а диагональные собственные веса подметаются согласованно. Цвет представляет весовой параметр обратной связи, WF. Черная тень: средние экспериментальные амплитуды; сплошная красная кривая: соответствующая модель фитинга; пунктирная красная линия: нестабильное решение. На вставках показаны временные кривые ниже, рядом и над бифуркацией. где m – целое число. Амплитуда имеет форму, аналогичную разветвлению вил в (7). Уравнение для sÀ1 в этой же точке можно использовать для определения угловой частоты:

ds1 dt □к 剔 ゲラ ωt = 2πmω == α − A − ωκA = 2WFσ (0) – σ (A) ((1187)) = τ − 1 WF (19).

Таким образом, ожидаемая частота предельного цикла конечна в точке Хопфа. точка и обратно пропорциональна выше.

Результаты

Мы экспериментально воспроизводим предсказания модели колебаний, амплитуды и частоты, параметризируя веса MRR сети, как описано выше, и изменяя их вместе. На вставках к рис. 4 показаны временные кривые ниже, вблизи и выше порога колебаний. Вблизи порога переходные колебания с непостоянной огибающей могут быть вызваны шумом. Выше порога колебания происходят в диапазоне 1-10 кГц, что ограничивается электронными фильтрами нижних частот и задержкой обратной связи. На рис. 4 показан результат точной развертки весов самоотдачи в 2-узловой сети, демонстрирующий форму параболоида би-

фуркации Хопфа. WF проходит по 300 точкам от 0,35 до 0,65, в то время как веса с диагональю являются фиксированными и асимметричными. Напряжение нейрона 1 отображается в зависимости от напряжения нейрона 2 с цветом, соответствующим параметру WF. Пиковая амплитуда колебаний для каждого груза затем проецируется на плоскость WF – y2 черным цветом, и эти амплитуды соответствуют модели из (16) (красный). Бифуркация происходит при WB = 0,48 в подгоночной модели. На рис. 5 показана частота колебаний выше точки Хопфа. Частота определяется путем обнаружения положительных переходов через нуль в s1 (t) и вычисления на основе разницы во времени. Данные отбрасываются для WB <WF <0,53, поскольку колебания в чувствительной переходной области хаотичны. Затем данные о частоте согласуются с моделью (19). Ось частот масштабируется так, что 1,0 соответствует частоте модели на границе области, которая составляет 4,81 кГц. Бифуркация Хопфа происходит только в системах более чем одного измерения, таким образом, подтверждая наблюдение небольшой интегрированной фотонной нейронной сети. Значительно выше точки бифуркации экспериментальные амплитуда и частота колебаний полностью соответствуют предсказаниям модели. Расхождения между моделью и наблюдениями на рис. 4 и 5 видны в чувствительном переходном диапазоне WB <WF <0,53. Вблизи точки бифуркации время, необходимое для схождения к устойчивому предельному циклу, значительно больше, чем время, находящееся выше бифуркации. Если колебание не успевает стабилизироваться в пределах окна записи, его равновесная амплитуда недооценивается. Другой потенциальный источник несоответствия – шум, который не был включен в модель. Предельные циклы с амплитудами, сравнимыми с амплитудой шума, могут быть дестабилизированы из-за их близости к неустойчивой фиксированной точке в нуле. Этот эффект мог бы объяснить среднюю вставку рис. 4, на которой небольшое колебание нарастает, а затем сокращается. Наконец, частично это несоответствие можно объяснить неточностью веса. Два банка гирь MRR были откалиброваны независимо с использованием метода, описанного в работе. [42]; однако не учитывались перекрестные тепловые помехи между разными банками. Как видно на рис. 2, физическое расстояние между w12 (номинально –1) и w22 (номинально WF) составляет примерно 100 мкм. Хотя межбанковские перекрестные помехи не являются серьезным эффектом, w12 очень чувствительна, потому что вес –1 соответствует резонансу, а динамика особенно чувствительна к значениям веса около точки бифуркации. Этот источник неточности веса отсутствовал в бифуркации куспида в разд. III, потому что динамически менялся только один банк весов.

V. ПРИМЕР КОНСТРУКЦИИ СИСТЕМЫ

Основным результатом этой статьи является демонстрация динамического изоморфизма между кремниевой фотонной системой и моделью CTRNN из (1). Этот результат означает, что более крупные и более быстрые кремниевые фотонные системы аналогичной формы могут использовать существующие инструменты для моделирования нейронных сетей. В этом разделе мы представляем пример процесса проектирования приложения фотонной сети нейронов MZM-типа под управлением Neural Engineering Framework (NEF) [63]. Мы моделируем сеть из 49 MZM-нейронов, решая задаваемое пользователем обычное дифференциальное уравнение (ODE), а затем сравниваем ее производительность с обычным компьютером, решающим то же ODE. Системы ОДУ повсеместно используются в вычислительных задачах [64–66], и этот факт мотивировал их разработку.

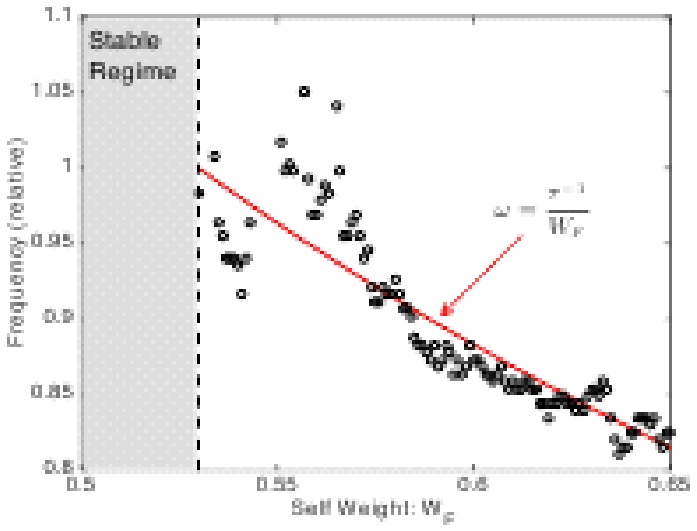


Рисунок. 5. Частота колебаний выше бифуркации Хопфа. Наблюдаемые данные (черные точки) сравниваются с ожидаемым трендом (19) (красная кривая). Частоты нормированы на пороговую частоту 4,81 кГц.

Установка специализированного оборудования [67]. Аппаратные эмуляторы, в отличие от программных симуляторов, обычно представляют собой аналоговые системы, которые можно сконфигурировать так, чтобы они демонстрировали ту же динамику, что и переменные в интересующем ОДУ. Для некоторых задач эмуляторы могут значительно улучшить скорость или потребление энергии по сравнению с обычным цифровым компьютером, выполняющим соответствующее моделирование [68]. Помимо производительности, в эмуляторах главное внимание уделяется широте проблем, которые можно эмулировать. Большая настраиваемость, предлагаемая весами нейронной сети, является преимуществом в этом отношении, но также представляет проблему при

определении того, как программировать веса. NEF предоставляет процедуру для получения произвольного ODE и возврата матрицы весов, которая приведет к его эмуляции с помощью CTRNN. Преимущество NEF перед другими нейронными структурами (например, слоистыми перцептронами) состоит в том, что он не полагается на адаптацию, а вместо этого гарантирует детерминированное решение для произвольных проблем определенных классов, включая эмуляцию ODE. Эти характеристики, присущие цифровым компиляторам, послужили основанием для обозначения «нейронного компилятора». Первоначально разработанный для оценки теорий познания, NEF был использован для решения инженерных задач [69] и использовался для программирования электронного нейроморфного оборудования [70]. Мы используем тестовую задачу эмуляции ODE, чтобы сравнить производительность фотонной CTRNN с производительностью обычного ЦП. В отличие от показателей уровня реализации, таких как тактовая частота или ток утечки транзисторов, тесты являются индикаторами уровня задач, хорошо подходящими для сравнения разнородных технологий. Различные нейроморфные электронные архитектуры протестированы в [71], а тестовые задачи для встроенных систем предложены в [72]. В этом случае фундаментальные различия в реализации цифровых компьютеров и нейронных сетей ограничивают значимость метрик на уровне устройств. Мы используем интерфейс компилятора NEF, чтобы установить тест, основанный на решении классического аттрактора Лоренца:

$$\begin{aligned} \dot{x}_0 &= \sigma(x_1 - x_0) \\ \dot{x}_1 &= -x_0 x_2 - x_1 \\ \dot{x}_2 &= x_0 x_1 - \beta(x_2 + \rho) - \rho \end{aligned} \qquad (20)$$

При параметрах по умолчанию: (σ, β, ρ) = (10, 8/3, 28) решения системы хаотичны.

Фотонный эмулятор CTRNN

NEF аппроксимирует функции f (x), используя линейную комбинацию заданных нейронных кривых настройки в рассматриваемой области значений ⓪x. Переменные моделирования x представлены линейными комбинациями реальных состояний сети s. Каждый нейрон в популяции имеет одинаковую кривую настройки σ, отличающуюся коэффициентом усиления g, вектором кодирующего устройства α и совокупностью b, так что yi = σ (gi ⁞ αi · s + bi). Введение повторных связей в популяции обеспечивает эффективную динамическую систему вида x˙ = f - (x).

Мы определяем настроечную кривую как синусоидальную электрооптическую передаточную характеристику MZM, и мы определяем интересующую систему в соответствии с трехпараметрической системой Лоренца из (20). Затем NEF предоставляет рекуррентную матрицу весов W, что приводит к эффективной эмуляции. Модификации стандартной процедуры NEF были сделаны, чтобы использовать связь функции передачи MZM с базисом Фурье, тем самым уменьшая количество необходимых нейронов. Вместо того, чтобы рисовать кодеры случайным образом, они были выбраны как вершины единичного куба α = [± 1, ± 1, ± 1].

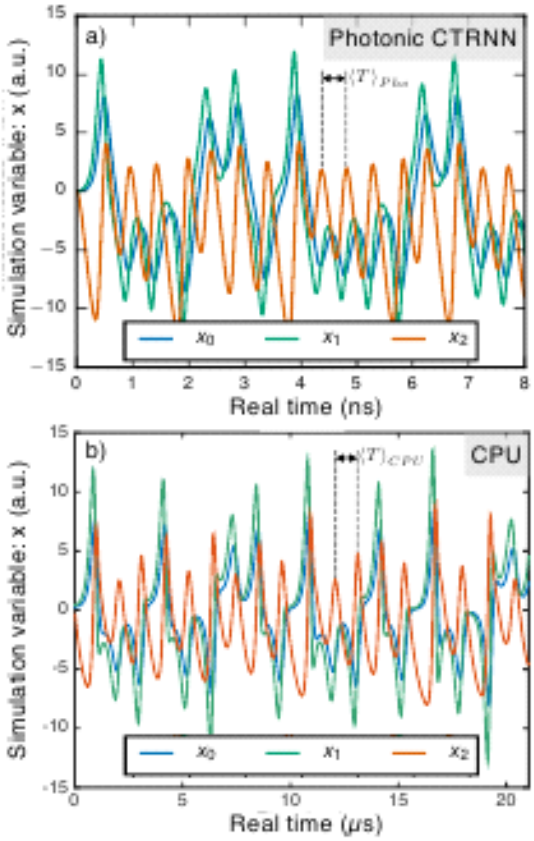


Рисунок. 6. а) Моделирование эмулятора кремниевой фотонной нейронной сети с непрерывным временем, запрограммированного с использованием NEF. б) Обычное моделирование ОДУ в дискретном времени. Временные окна охватывают равные интервалы времени эмуляции, измеряемые x2 интервалами пересечения, ħ T. Оси времени масштабируются по τ и Δt, соответственно, показывая ускорение в реальном времени 1,960 ×.

Усиления были выбраны так, чтобы соответствовать первым трем частотам Фурье области: eghal∈fp {esrπio / d2., SOπ, ff3 sseπts / 2w} e, rewchheorseensπtoisbethbe∈M {Z0, Msπf / u2n} c.tiAonnxtra нейрон с константой вывод добавлен для учета нулевой частоты разложения Фурье. Таким образом, t8o · t3al · n2u + m1be = r 4of9.neurons равно # α · #g · #b + 1 = Операционная скорость системы определяется синаптической постоянной времени τ, которая эквивалентна постоянной времени

кремниевого MZM. Кремниевые MZM с полосой пропускания 40 ГГц сейчас широко распространены [73, 74]; однако задержка обратной связи является ограничивающим фактором, поскольку она должна быть меньше синаптических постоянных времени. Предположим, что сеть имеет геометрию, показанную на рис. 1, самый длинный путь обратной связи проходит через порт приема последнего (розового) веса MRR первого (желтого) банка нейронов. Если число нейронов равно Nmu = m49f, eeadnbdatchkeleMngRthR ipsitLch = isNDD (≈120 + μ2m +, t3h), все три msuaxmi – mand соответственно соответствуют первому проходу через банк, траекторию падающего волновода и тракту волновода обратной связи. В этом примере задержка обратной связи nL / c составляет 69 пс, что означает, что схема драйвера модулятора должна быть отфильтрована нижними частотами. При моделировании аппаратной нейронной сети мы выбираем τ = 100ps.

Б. Симулятор ЦП

Обычные процессоры используют приближение дискретного времени для моделирования непрерывных ОДУ, простейшим из которых является продолжение Эйлера:

$$\vec{x}[(n+1)\Delta t] = \vec{x}[n\Delta t] + \Delta t \vec{f}(\vec{x}[n\Delta t]) \qquad (21)$$

где Δt – интервал временного шага, который связан как со временем моделирования, так и с физическим реальным временем. Чтобы оценить значение физического времени Δt, мы разрабатываем и проверяем простую модель ЦП. Для каждого временного шага ЦП должен вычислить f (∴ x [nΔt]), как определено в (20), что приведет к 9 операциям с плавающей точкой (FLOP) и 12 операциям чтения из кэша операндов. Обновление Эйлера в (21) включает одно умножение, одно сложение и одно чтение / запись для каждой переменной состояния, в результате чего получается 6 FLOP и 6 обращений к кэшу. С задержкой FLOP, равной 1 тактовому циклу, задержкой кэша уровня 1 (L1) 4 цикла [75] и тактовой частотой 2,6 ГГц, эта модель предсказывает временной шаг Δt = 33 нс. Эта модель проверена эмпирически с использованием Intel Core i5-4288U [75]. Машинно-оптимизированная программа randominициализирует и выполняет цикл через 106 шагов Эйлера системы Лоренца, более 100 испытаний. Время процессора составило vΔt = 24,5 ± 1,5 нс. Мы отмечаем, что архитектура ЦП значительно канареечна, в том числе и способами, которые могут повысить производительность в этой конкретной задаче (например, за счет хранения операндов в регистрах), но означают, что эта модель служит количественной базой.

Бенчмаркинг

Моделирование в дискретном времени и эмуляция в непрерывном времени связаны с физическим реальным временем с помощью переменных Δt и τ, но должны тестироваться с использованием общей временной основы эмуляции / моделирования. Базис времени эмуляции / моделирования может быть установлен на основе переменной x2 с хорошим поведением, чей интервал перехода через нуль до сих пор определяется Rfseeixrmhriuebldiattteioodna

Мы выполнили серию эмуляции задач в neuromorphic кремния ФОТОГРАФИЕЙ тоник система. Из-за изоморфизма, продемонстрированного в п. III-IV, аналогичные процедуры могут быть разработаны с использованием других инструментов CTRNN. Это упражнение страдает от ограниченной актуальности игрушечных задач в дополнение к пренебрежению оптимизацией, возможной в цифровых технологиях (например, программируемыми полями вентильными матрицами). Тем не менее, общий подход к сравнительному анализу, основанный на программировании нейронных сетей на уровне задач, может быть значительно усовершенствован для оценки возможностей нейроморфной фотоники в конкретных областях применения.

ЗАКЛЮЧЕНИЕ

Мы продемонстрировали реконфигурируемую аналоговую нейронную сеть в кремниевой фотонной интегральной схеме, использующей модуляторы в качестве нейронных элементов. Опосредованные сетью бугорки и бифуркации Хопфа наблюдались как первое доказательство концепции интегрированной системы широковещания и взвешивания [40]. Абстракции нейронных сетей – мощные инструменты для преодоления разрыва между физической динамикой и полезными приложениями, а производство кремниевых фотонов открывает возможности для крупномасштабных фотонных систем. Моделирование нейронных сетей с 49 модуляторами, выполняющих функции нейронные сети, может быть применено к неадресуемым вычислительным областям, требующим сверхбыстрого, реконфигурируемого оборудования и процессоров. Более того, кремниевые фотонные нейронные сети могут стать первыми попытками освоить более широкий класс кремниевых фотонных систем для масштабируемой обработки информации.

БЛАГОДАРНОСТИ

Эта работа поддержана программой Национального научного фонда (NSF) по расширению доступа к радиочастотному спектру (EARS) (Премия 1642991). Поддержка изготовления была предоставлена через про-

грамму Совета по естественным наукам и инженерным исследованиям Канады (NSERC) по кремниевым электронно-фотонным интегральным схемам (SiEPIC). Устройства были изготовлены Ричардом Бойко на заводе нанофабрикатов Вашингтонского университета, входящем в Национальную сеть нанотехнологической инфраструктуры NSF (NNIN).

REFERENCES

[1] O. A. Reimann and W. F. Kosonocky, “Progress in optical computer research,” IEEE Spectrum 2, 181–195 (1965).

[2] F. B. McCormick, T. J. Cloonan, F. A. P. Tooley, A. L.Lentine, J. M. Sasian, J. L. Brubaker, R. L. Morrison,S. L. Walker, R. J. Crisci, R. A. Novotny, S. J. Hinterlong, H. S. Hinton, and E. Kerbis, “Six-stage digital free-space optical switching network using symmetric self-electro-optic-effect devices,” Appl. Opt. 32, 5153–5171 (1993).

[3] S. Jutamulia and F. Yu, “Overview of hybrid optical neural networks,” Optics & Laser Technology 28, 59 – 72 (1996).

[4] R. W. Keyes, “Optical logic-in the light of computer technology,” Optica Acta: International Journal of Optics 32, 525–535 (1985).

[5] Y. Vlasov, “Silicon CMOS-integrated nano-photonics for computer and data communications beyond 100G,” IEEE Commun. Mag. 50, s67–s72 (2012).

[6] M. Hochberg, N. C. Harris, R. Ding, Y. Zhang, A. Novack, Z. Xuan, and T. Baehr-Jones, “Silicon photonics: The next fabless semiconductor industry,” IEEE SolidState Circuits Magazine 5, 48–58 (2013).

[7] D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J.-M. Fe´de´li, J.-M. Hartmann, J. H. Schmid, D.-X. Xu, F. Boeuf, P. O’Brien, G. Z. Mashanovich, and M. Nedeljkovic, “Roadmap on silicon photonics,” Journal of Optics 18, 073003 (2016).

[8] A.-J. Lim, J. Song, Q. Fang, C. Li, X. Tu, N. Duan, K. K. Chen, R.-C. Tern, and T.-Y. Liow, “Review of silicon photonics foundry efforts,” IEEE J. Sel. Top. Quantum Electron. 20, 405–416 (2014).

[9] J. S. Orcutt, B. Moss, C. Sun, J. Leu, M. Georgas, J. Shainline, E. Zgraggen, H. Li, J. Sun, M. Weaver, S. Urosˇevic´, M. Popovic´, R. J. Ram, and V. Stojanovic´, “Open foundry platform for high-performance electronic-photonic integration,” Opt. Express 20, 12222–12232 (2012).

[10] L. Chrostowski and M. Hochberg, Silicon Photonics Design: From Devices to Systems (Cambridge University Press, 2015).

[11] J. Sun, E. Timurdogan, A. Yaacobi, Z. Su, E. Hosseini, D. Cole, and M. Watts, “Large-scale silicon hotonic circuits for optical phased arrays,” Selected Topics in Quantum Electronics, IEEE Journal of 20, 264–278 (2014).

[12] R. G. Beausoleil, “Large-scale integrated photonics for high-performance interconnects,” J. Emerg. Technol. Comput. Syst. 7, 6:1–6:54 (2011).

[13] S. Le Beux, J. Trajkovic, I. O’Connor, G. Nicolescu, G. Bois, and P. Paulin, “Optical ring network-on-chip (ORNoC): Architecture and design methodology,” in “Design, Automation Test in Europe Conference Exhibition (DATE), 2011,” (2011), pp. 1–6.

[14] J. Capmany, J. Mora, I. Gasulla, J. Sancho, J. Lloret, and S. Sales, “Microwave photonic signal processing,” Journal of Lightwave Technology 31, 571–586 (2013).

[15] A. Farsaei, Y. Wang, R. Molavi, H. Jayatilleka, M. Caverley, M. Beikahmadi, A. H. M. Shirazi, N. Jaeger,L. Chrostowski, and S. Mirabbasi, “A review of wireless photonic systems: Design methodologies and topologies, constraints, challenges, and innovations in electronics and photonics,” Optics Communications pp. – (2016).

[16] N.-N. Feng, P. Dong, D. Feng, W. Qian, H. Liang, D. C. Lee, J. B. Luff, A. Agarwal, T. Banwell, R. Menendez, P. Toliver, T. K. Woodward, and M. Asghari, “Thermally-efficient reconfigurable narrowband rfphotonic filter,” Opt. Express 18, 24648–24653 (2010).

[17] L. Zhuang, C. G. H. Roeloffzen, M. Hoekman, K.-J. Boller, and A. J. Lowery, “Programmable photonic signal processor chip for radiofrequency applications,” Optica 2, 854–859 (2015).

[18] G. C. Valley, “Photonic analog-to-digital converters,” Opt. Express 15, 1955–1982 (2007).

[19] M. H. Khan, H. Shen, Y. Xuan, L. Zhao, S. Xiao, D. E. Leaird, A. M. Weiner, and M. Qi, “Ultrabroad-

bandwidth arbitrary radiofrequency waveform generation with a silicon photonic chip-based spectral shaper,” Nature: Photonics 4, 117–122 (2010).

[20] J. Chang, J. Meister, and P. R. Prucnal, “Implementing a novel highly scalable adaptive photonic beamformer using “blind” guided accelerated random search,” Journal of Lightwave Technology 32, 3623–3629 (2014).

[21] T. Ferreira de Lima, A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Scalable wideband principal component analysis via microwave photonics,” IEEE Photonics Journal 8, 1–9 (2016).

[22] G.-k. Chang and L. Cheng, “The benefits of convergence,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20140442 (2016).

[23] U.-P. Wen, K.-M. Lan, and H.-S. Shih, “A review of hopfield neural networks for solving mathematical programming problems,” European Journal of Operational Research 198, 675 – 687 (2009).

[24] T. Lee and F. Theunissen, “A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal fea-

tures," Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 471 (2015).

[25] D. J. C. MacKay, "A practical Baysian framework for backpropagation networks," Neural Computation 4, 448–472 (1992).

[26] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks 2, 359–366 (1989).

[27] C. Eliasmith and C. H. Anderson, Neural engineering: Computation, representation, and dynamics in neurobiological systems (MIT Press, 2004).

[28] F. Donnarumma, R. Prevete, A. de Giorgio, G. Montone, and G. Pezzulo, "Learning programs is better than learning dynamics: A programmable neural network hierarchical architecture in a multi-task scenario," Adaptive Behavior 24, 27–51 (2016).

[29] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking neuron integrated circuit with a scalable communication network and interface," Science 345, 668–673 (2014).

[30] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. Kuang, R. Manohar, W. Risk, B. Jackson, and D. Modha, Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 34, 1537–1557 (2015).

[31] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," arXiv:1506.03264 (2015).

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recogni-

tion," Proceedings of the IEEE 86, 2278–2324 (1998).

[33] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation 18, 1527–1554 (2006).

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature 521, 436–444 (2015).

[35] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre,. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," Nature 529, 484–489 (2016).

[36] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," Nat Commun 4, 1364 (2013).

[37] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat Commun 5 (2014).

[38] M. C. Soriano, D. Brunner, M. Escalona-Mora´n, C. R. Mirasso, and I. Fischer, "Minimal approach to neuro-inspired information processing," Frontiers in Computational Neuroscience 9, 68 (2015).

[39] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," Scientific Reports 6, 22381 EP – (2016).

[40] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," J. Lightwave Technol. 32, 3427–3439 (2014).

[41] A. Tait, T. Ferreira de Lima, M. Nahmias, B. Shastri, and P. Prucnal, "Continuous calibration of microring weights for analog optical networks," Photonics Technol. Lett. 28, 887–890 (2016).

[42] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Multi-channel control for microring weight banks," Opt. Express 24, 8895–8906 (2016).

[43] A. N. Tait, A. X. Wu, T. Ferreira de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Mcroring weight banks," IEEE Journal of Selected Topics in Quantum Electronics PP, 1–1 (2016).

[44] A. Tait, A. Wu, E. Zhou, T. Ferreira de Lima, M. Nahmias, B. Shastri, and P. Prucnal, "Demonstration of a silicon photonic neural network," in "Summer Topicals [1] O. A. Reimann and W. F. Kosonocky, "Progress in optical computer research," IEEE Spectrum 2, 181–195 (1965).

[45] R. D. Beer, "On the dynamics of small continuous-time recurrent neural networks," Adaptive Behavior 3, 469–509 (1995).

[46] V. K. Tumuluru, P. Wang, and D. Niyato, "A neural network based spectrum prediction scheme for cognitive radio," in "Communications (ICC), 2010 IEEE International Conference on," (2010), pp. 1–5.

[47] U. Mitra and H. V. Poor, "Neural network techniques for adaptive multiuser demodulation," IEEE Journal on Selected Areas in Communications 12, 1460–1470 (1994).

[48] K.-L. Du, A. Lai, K. Cheng, and M. Swamy, "Neural methods for antenna array signal processing: a review," Signal Processing 82, 547 – 561 (2002).

[49] J. J. Hopfield and D. W. Tank, ""neural" computation of decisions in optimization problems," Biological Cybernetics 52, 141–152 (1985).

[50] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," IEEE J. Sel. Top. Quantum Electron. 19, 1–12 (2013).

[51] P. R. Prucnal, B. J. Shastri, T. Ferreira de Lima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," Adv. Opt. Photon. 8, 228–299 (2016).

[52] F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Relative refractory period in an excitable semiconductor laser," Phys. Rev. Lett. 112, 183902 (2014).

[53] B. Romeira, R. Avo´, J. L. Figueiredo, S. Barland, and J. Javaloyes, "Regenerative memory in time-delayed neuromorphic photonic resonators," Scientific Reports 6, 19510 EP – (2016).

[54] M. A. Nahmias, A. N. Tait, L. Tolias, M. P. Chang, T. Ferreira de Lima, B. J. Shastri, and P. R. Prucnal, "An integrated analog O/E/O link for multi-channel laser neurons," Applied Physics Letters 108, 151106 (2016).

[55] T. V. Vaerenbergh, M. Fiers, P. Mechet, T. Spuesens, R. Kumar, G. Morthier, B. Schrauwen, J. Dambre, and P. Bienstman, "Cascadable excitability in microrings," Opt. Express 20, 20292–20308 (2012).

[56] B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," Sci. Rep. 5, 19126 (2015).

[57] R. J. Bojko, J. Li, L. He, T. Baehr-Jones, M. Hochberg, and Y. Aida, "Electron beam lithography writing strategies for low loss, high confinement silicon optical waveguides," J. Vac. Sci. Technol., B 29 (2011).

[58] Y. Wang, X. Wang, J. Flueckiger, H. Yun, W. Shi, R. Bojko, N. A. Jaeger, and L. Chrostowski, "Focusing subwavelength grating couplers with low back reflections for rapid prototyping of silicon photonic circuits," Opt. Express 22, 20652–20662 (2014).

[59] B. Romeira, F. Kong, W. Li, J. M. Figueiredo, J. Javaloyes, and J. Yao, "Broadband chaotic signals and breather oscillations in an optoelectronic oscillator incorporating a microwave photonic filter," Lightwave Technology, Journal of 32, 3933–3942 (2014).

[60] E. Zhou, A. Tait, A. Wu, T. Ferreira de Lima, M. Nahmias, B. Shastri, and P. Prucnal, "Silicon photonic weight bank control of integrated analog network dynamics," in "Optical Interconnects Conference, 2016 IEEE," (IEEE, 2016), p. TuP9.

[61] Y. Zhang, S. Yang, A. E.-J. Lim, G.-Q. Lo, C. Galland, T. Baehr-Jones, and M. Hochberg, "A compact and low loss y-junction for submicron silicon waveguide," Opt. Express 21, 1310–1316 (2013).

[62] E. Izhikevich, Dynamical systems in neuroscience: the geometry of excitability and bursting (MIT press, 2006).

[63] T. C. Stewart and C. Eliasmith, "Large-scale synthesis of functional spiking neural circuits," Proceedings of the IEEE 102, 881–898 (2014).

[64] J. Hoffman and C. Johnson, “A new approach to computational turbulence modeling,” Computer Methods in Applied Mechanics and Engineering 195, 2865 – 2880 (2006).

[65] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings, “A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modelling of complex fluids: Part ii: Transient simulation using space-time separated representations,” Journal of Non-Newtonian Fluid Mechanics 144, 98 –121 (2007).

[66] H. Yoshino and M. Shibata, “Chapter 9 higher dimensional numerical relativity: Current status,” Progress of Theoretical Physics Supplement 189, 269–309 (2011).

[67] G. Cowan, R. Melville, and Y. Tsividis, “A vlsi analog computer/digital computer accelerator,” Solid-State Circuits, IEEE Journal of 41, 42–53 (2006).

[68] N. Ratier, “Analog computing of partial differential equations,” in “Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on,” (2012), pp. 275–282.

[69] K. E. Friedl, A. R. Voelker, A. Peer, and C. Eliasmith, “Human-inspired neurorobotic system for classifying surface textures by touch,” IEEE Robotics and Automation Letters 1, 516–523 (2016).

[70] A. Mundy, J. Knight, T. Stewart, and S. Furber, “An efficient spinnaker implementation of the neural engineering framework,” in “Neural Networks (IJCNN), 2015 International Joint Conference on,” (2015), pp. 1–8.

[71] A. Diamond, T. Nowotny, and M. Schmuker, “Comparing neuromorphic solutions in action: implementing a bio-inspired solution to a benchmark classification task on three parallel-computing platforms,” Frontiers in Neuroscience 9 (2016).

[72] T. C. Stewart, T. DeWolf, A. Kleinhans, and C. Eliasmith, “Closed-loop neuromorphic benchmarks,” Frontiers in Neuroscience 9 (2015).

[73] G. T. Reed, G. Mashanovich, F. Y. Gardes, and D. J. Thomson, “Silicon optical modulators,” Nat Photon 4, 518–526 (2010).

[74] D. Patel, S. Ghosh, M. Chagnon, A. Samani, V. Veerasubramanian, M. Osman, and D. V. Plant, “Design, analysis, and transmission system performance of a 41 ghz silicon photonic modulator,” Optics express 23, 14263—14287 (2015).

[75] Intel Corporation, Intel 64 and IA-32 Architectures Optimization Reference Manual (2016). Meeting Series (SUM), 2016,” (IEEE, 2016).

НЕЙРОМОРФНЫЙ ФОТОННЫЙ ПРОЦЕССОР TERAMAC

Нахмиас М.А., Пэн С.-Т., Феррейра де Лима Т., Хуан Ч., Тейт А.Н., Шастри Б.Д., Прукнал П.Р.

Аннотация.

Мы показываем, что интегрированный лазерный нейрон может демонстрировать исключительно низкую задержку (<1 нс) и скорость (~ 1 × 1012 MAC / с на устройство) по сравнению с современными процессорами в цифровой электронике. Мы экспериментально демонстрируем положительные (возбуждающие) и отрицательные (подавляющие) входы с 8 каналами с длиной волны и эффективностью (<1 пДж / MAK) во время работы с обратной связью.

I. ВВЕДЕНИЕ

Недавний спрос на оборудование для глубокого обучения привел к огромным коммуникационным и вычислительным требованиям. Специализированные процессоры, такие как Tensor Processing Unit (TPU), по популярности и производительности обгоняют универсальные вычислительные процессоры, такие как CPU или GPU [1]. Ключевой метрикой в таком оборудовании является потребление энергии, которое включает в себя как перемещение данных, так и вычисления, последние из которых состоят в основном из матричных вычислений. Составляющими компонентами матричных вычислений являются операции умножения с накоплением (MAC), то есть операции вида

$a = a + w \times x$.

У фотоники есть потенциал для устранения узких мест, как в связи, так и в вычислениях: (1) оптические межсоединения могут обеспечивать каналы связи с низким энергопотреблением, не ограниченные расстоянием [2], в то время как (2) операции фотонной матрицы выгодно масштабируются, поскольку потребление энергии пропорционально количеству каналов, а не количество MAC [3]. Здесь мы тестируем и исследуем свойства процессора лазерных нейронов, созданного на платформе фотонной интегральной схемы (PIC) из фосфида индия. Этот лазерный нейрон демонстрирует биологически значимое свойство пиков в гораздо более быстром масштабе времени, чем биологические (> 108) или электрические (> 103) системы. Кроме того, он демонстрирует необычайную производительность по сравнению с новейшим оборудованием в области глубокого обучения и нейроморфного электронного оборудования:

одно устройство может обрабатывать ~1 × 1012 MAC / с в секунду с энергоэффективностью ~270 фДж на MAC. операция.

II. МНОГОЗАВИСИМАЯ ФУНКЦИОНАЛЬНОСТЬ

Процессор разработан для обеспечения совместимости с сетевым протоколом с длиной волны, называемым Broadcast-and-Weight (B&W) [4]. Этот протокол использует мультиплексирование с разделением по длине волны (WDM), чтобы обеспечить масштабируемые межсоединения между процессорами лазерных нейронов. Большие сети (> 100) могут быть созданы на кристалле, для чего требуется лишь небольшое количество волноводов без каких-либо пересечений волноводов [5]. Мы разработали композитную конструкцию устройства, изготовленную на стандартной платформе PIC из фосфида индия в Институте Генриха Герца. Структура представлена на рис. 1a. В нашей экспериментальной демонстрации использовалось всего 8 каналов с длиной волны с независимыми пиковыми сигналами: 4 входа (a) (b)

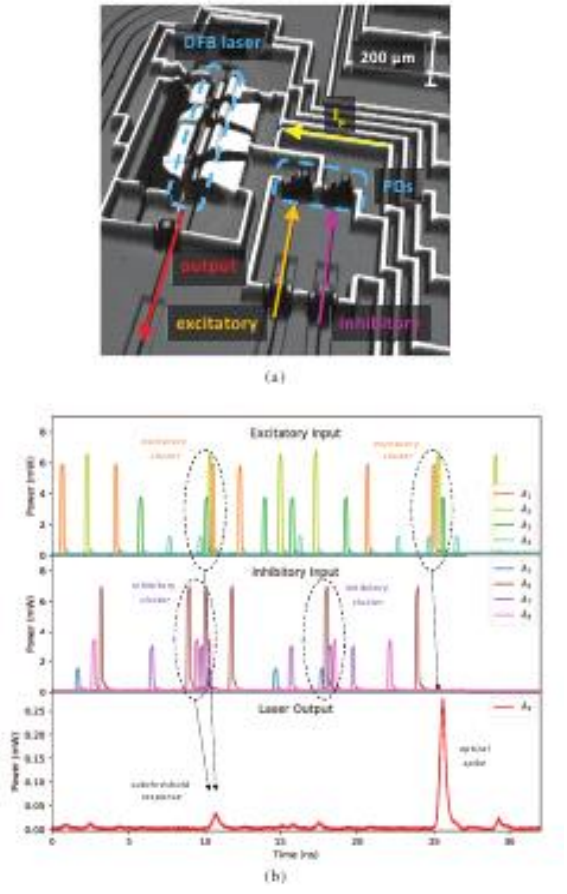


Рис. 1: (a) Топографическая микрофотография интегрированного лазерного нейрона. Возбуждающий и тормозной входы поступают на сбалансированную пару фотодетекторов (PD), которая управляет лазером с распределенной обратной связью (DFB), смещенным током Ip. (b) Устройство протестировано с 8 независимыми длинами волн каналов λi... λ8. Все длины волн находятся в телекоммуникационном C-диапазоне.

были связаны с каждым фотодетектором. Лазер работает чуть ниже порога генерации s.t. потеря превышает общую прибыль. Запрещающий и возбуждающий фотодетектор обеспечивает отрицательный и положительный вход через двухтактную конфигурацию. Только когда большой кластер возбуждающих импульсов прибывает близко во времени без торможения, лазер испускает оптический импульс (шириной ~ 300 пс).

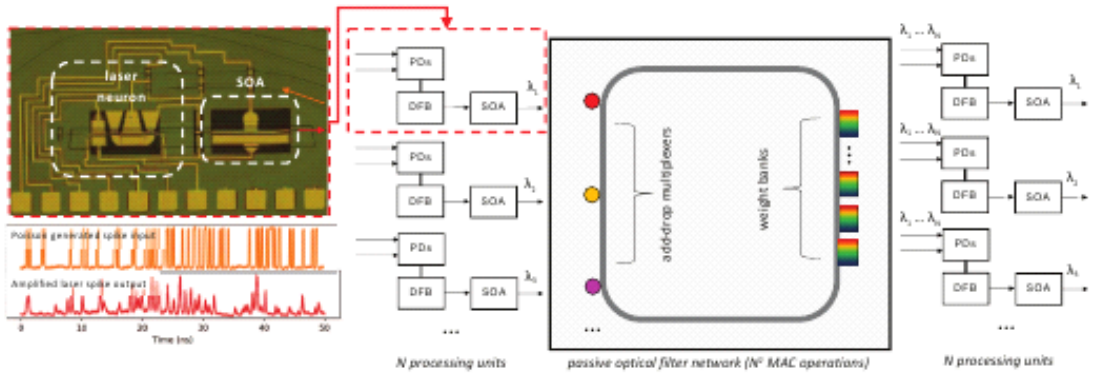


Рис. 2: Слева: микрофотография лазерного нейрона с полупроводниковым оптическим усилителем. Слева внизу: измеренная входная кривая, сгенерированная с использованием процесса Пуассона, и измеренный выходной сигнал при усилении с обратной связью. Справа: схема гипотетической системы с пассивной оптической сетью на кристалле. N2 MAC операции выполняются, но расход энергии. масштабируется с номером процессора N.

Блокирующий кластер предотвращает высвобождение импульса. Эта базовая функциональность, показанная на рис. 1b, имитирует нейрон Leaky Integrate-and-Fire, модель полного всплеска по Тьюрингу, хорошо известную в области вычислительной нейробиологии [6]. Система может масштабироваться за пределы протестированных 8 каналов, чтобы охватить более 100 каналов, видимых в плотном WDM. Недавние работы по разработке фильтров с микрокольцом показали, что возможно до ~ 200 каналов [7]. Если лазерные устройства сопрягаются с пассивными фильтрами с использованием промежуточной технологии или гибридной кремниевой / III-V платформы, мы можем масштабировать систему с помощью множества интегрированных компонентов. При N ~ 200 и рефрактерном периоде ~ 0,2 нс один лазерный нейрон может выполнять примерно ~ 1 × 1012 MAK / c.

III. ПОТРЕБЛЕНИЕ ЭНЕРГИИ

Для желаемой средней точности на канал, потребление мощности масштабируется с числом каналов N, а не с числом операций MAC N2. Это связано с тем, что пассивные схемы в принципе не потребляют энергию. Как показано на рис. 2, серия из N лазерных нейронов вводит

сигналы в пассивную сеть, которая перераспределяет эти сигналы (т.е. умножает их на матрицу весов) перед следующим этапом. Пока каждый набор из N процессоров может усиливать свой выходной сигнал и компенсировать пассивные потери – обычно <3 дБ на кристалле – система может продолжать регенеративную обработку информации. Для каскадной работы требуется соблюдение условия усиления с обратной связью: выходная мощность должна превышать входную. Для достижения этого условия мы разработали лазерный нейрон с полупроводниковым оптическим усилителем (ПОУ), прикрепленным к выходу лазера, микрофотография которого показана на рис. 2. Для измерения энергии, необходимой для накачки ПОУ, мы сгенерировали вводят импульсы через процесс Пуассона в возбуждающие частичные разряды с длительностью импульса 0,2 нс и частотой Пуассона λp = 1 ГГц. Мы скорректировали ток SOA для соответствия условию усиления замкнутого контура. что произошло при ISOA = 105 мА при напряжении 2,50 В. Поскольку лазер и фотодетектор вместе потребляли <10 мВт мощности, потребление энергии SOA доминирует. Тем не менее, с вычислительной мощностью TeraMAC это составляет около 270 фДж на MAC. Еще одно ключевое преимущество заключается в том, что, поскольку по умолчанию сигналы кодируются на свет, дополнительная энергия не расходуется на перемещение данных из одного места в другое – обычно это основной источник потерь мощности в электронике.

IV. ЗАКЛЮЧЕНИЕ

Мы изготовили и протестировали лазерный нейрон, созданный на платформе фотонных интегральных схем. Он демонстрирует множество полезных свойств, которые позволяют обойти многие ограничения, присущие цифровой электронике. Обратите внимание, что – в качестве доказательства концепции – есть много способов улучшить производительность. Например, более сильная динамика поглощения может привести к более коротким импульсам; потери микроволн между фотодетектором и лазером можно уменьшить за счет увеличения импеданса входной линии; устройства меньшего размера могут привести к более высокой эффективности преобразования, а использование новых материалов, таких как графен, может привести к гораздо более быстрой (> 100 TMACs / s) работе [8]. Дальнейшая оптимизация может в конечном итоге сделать усилитель ненужным, проложив путь к революционной энергоэффективности (<1 фДж / MAC) и производительности для систем глубокого обучения следующего поколения.

REFERENCES

[1] N. P. Jouppi et al., “In-datacenter performance analysis of a tensor processing unit,” in Proceedings. ACM, 2017, pp. 1–12.

[2] D. A. Miller, “Attojoule optoelectronics for low-energy information processing and communications,” JLT, vol. 35, no. 3, pp. 346–396, 2017.

[3] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” Nature Photonics, vol. 11, pp. 441, 2017.

[4] A. N. Tait et al., “Broadcast and weight: an integrated network for scalable photonic spike processing,” JLT, vol. 32, no. 21, pp. 3427–3439, 2014.

[5] P. R. Prucnal et al., Neuromorphic photonics. CRC Press, 2017.

[6] M. A. Nahmias et al., “A leaky integrate-and-fire laser neuron for ultrafast cognitive computing,” IEEE JSTQE, vol. 19, no. 5, 2013.

[7] A. N. Tait et al., “Two-pole microring weight banks,” Optics Letters (accepted), 2018.

[8] B. J. Shastri et al., “Spike processing with a graphene excitable laser,” Scientific Reports, vol. 6, pp. 19 126, 2016.

ИССЛЕДОВАНИЕ КРЕМНИЕВОЙ ФОТОНИКИ ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ

Санни Ф.П., Тахери Э., Никдаст М., Пасрича С.

Глубокое обучение привело к беспрецедентным успехам в решении некоторых очень сложных проблем в таких областях, как компьютерное зрение, обработка естественного языка и общее распознавание образов.

Эти достижения являются кульминацией многолетних исследований лучших методов обучения и более глубоких моделей нейронных сетей, а также усовершенствований аппаратных платформ, которые используются для обучения и выполнения глубоких моделей нейронных сетей. Многие аппаратные ускорители на интегральных схемах для конкретных приложений (ASIC) для глубокого обучения вызвали интерес в последние годы из-за их улучшенной производительности и энергоэффективности по сравнению с традиционными архитектурами ЦП и ГП. Однако эти ускорители ограничены фундаментальными узкими местами из-за:

1) замедления масштабирования CMOS, которое ограничивает вычислительные возможности и производительность на ватт появляющихся электронных процессоров

2) использование металлических межсоединений для перемещения данных, которые не работают. хорошо масштабируются и являются основной причиной неэффективности полосы пропускания, задержек и энергопотребления практически в каждом современном процессоре.

Кремниевая фотоника стала многообещающей КМОП-совместимой альтернативой для реализации нового поколения ускорителей глубокого обучения, которые могут использовать свет как для связи, так и для вычислений. В этой статье рассматривается ландшафт кремниевой фотоники для ускорения глубокого обучения с описанием разработок в рамках абстракций дизайна снизу вверх, чтобы передать как возможности, так и ограничения парадигмы кремниевой фотоники в контексте ускорения глубокого обучения.

КОНЦЕПЦИИ CCS • Аппаратное обеспечение → Новые технологии → Новые оптические и фотонные технологии • Вычислительные методологии → Машинное обучение • Аппаратное обеспечение → Очень крупномасштабная интеграция.

Дополнительные ключевые слова и фразы: кремниевая фотоника, глубокое обучение, нейроморфные вычисления.

1 ВВЕДЕНИЕ

Глубокое обучение, которое является подразделом. Область искусственного интеллекта (ИИ) была в центре многих беспрецедентных успехов в последние годы в решении очень сложных проблем в областях компьютерного зрения, обработки естественного языка, прогнозирования временных рядов и понимания больших данных. Такое развитие событий примечательно, учитывая, что большинство исследователей отказались от идеи использования глубокого обучения в 1990-х годах из-за трудностей с обучением таких моделей. Но основополагающая работа Хинтона и др. в 2006 году показали, как можно обучить глубокую нейронную сеть распознавать рукописные цифры с современной точностью (> 98%) [1]. Они назвали свою технику «Глубокое обучение». Научное сообщество быстро обратило на это внимание, и в последующие годы многие исследователи показали, что глубокое обучение не только возможно, но и способно обеспечить выдающуюся производительность для решения многих проблем, с которыми не могут сравниться никакие другие методы машинного обучения. Действительно, сегодня модели глубокого обучения лежат в основе интеллектуальных технологических решений, которые мы все регулярно используем, таких как поисковые системы, механизмы рекомендаций по музыке и видео, распознавание речи в виртуальных помощниках и обнаружение объектов в Интернете вещей (IoT). камеры. Многие новые приложения, такие как беспилотные автомобили [2], автономная робототехника [3], обнаружение фейковых новостей [4], прогнозирование роста пандемии и тенденций [5], обнаружение сетевых аномалий [6] и языковой перевод в реальном времени [7]] опираются на все более сложные модели глубокого - обучения.

Магия глубокого обучения во многом обязана архитектуре нашего мозга. Еще в 1943 году нейрофизиолог Уоррен МакКаллох и математик Уолтер Питтс представили упрощенную модель того, как биологические нейроны работают вместе в мозгу животных для выполнения сложных вычислений [8]. Это была первая архитектура искусственной нейронной сети (ИНС), которая вдохновила гонку на создание интеллектуальных машин, которые могли бы конкурировать и в конечном итоге превосходить возможности человеческого мозга. Введение перцептрона в 1957 году Фрэнком Розенблаттом стало еще одной вехой, показавшей, как простую ИНС можно обучить для решения задач классификации [9]. Однако ограниченные возможности оборудования для запуска даже умеренно сложных ИНС заставили исследователей отказаться от изучения ИНС в конце 1960-х годов. Несмотря на то, что в 1980-х и начале 1990-х годов появились новые архитектуры и лучшие методы обучения,

прогресс был ограничен из-за нескольких факторов, важнейшим из которых было отсутствие мощных машин для обучения и работы этих моделей. К счастью, за последнее десятилетие постоянно улучшающиеся возможности технологии изготовления дополнительных металлооксидных полупроводников (CMOS) позволили создать чрезвычайно мощные графические процессоры (GPU) класса TFLOP и процессорные чипы ЦП с миллиардами транзисторов в малых форм-факторах, которые сделали это возможным. для обучения и использования архитектуры глубокой ИНС (т. е. многослойного персептрона (MLP)) своевременно и с минимальными затратами. В сочетании с доступностью больших наборов данных в эпоху Интернета вещей и больших данных, теоретическими достижениями в алгоритмах обучения и появлением новых архитектур глубоких ИНС, таких как сверточные нейронные сети (CNN), глубокое обучение теперь установило свое доминирование над другими моделями машинного обучения. для многих проблем, представляющих интерес в областях компьютерного зрения, обработки естественного языка и общего распознавания образов. В связи с тем, что исследователи создают более глубокие и сложные архитектуры MLP и CNN, чтобы поднять уровень производительности глубокого обучения на новый уровень, базовая аппаратная платформа должна постоянно обеспечивать более высокие уровни производительности, а также удовлетворять строгим ограничениям на рассеивание мощности. Это стремление к достижению более высокой производительности на ватт побудило архитекторов оборудования разрабатывать ускорители интегральных схем (ASIC) для конкретных приложений для глубокого обучения, которые имеют гораздо более высокую производительность на ватт, чем обычные процессоры общего назначения и графические процессоры. Чип IBM TrueNorth с 4096 ядрами, выпущенный в 2014 году, был одним из первых широко известных ускорителей глубокого обучения ASIC [10]. С тех пор стали доступны многие другие ускорители, в том числе Intel Loihi [11] и Tensor Processing Units (TPU) Google [12]. Некоторые академические усилия также привели к разработке новых типов ускорителей глубокого обучения на базе ASIC и FPGA [13] – [17]. Даже обычные графические процессоры и процессоры эволюционировали для ускорения выполнения модели глубокого обучения, например, графические процессоры Nvidia теперь включают тензорные ядра [18], а процессоры поддерживают все более продвинутые векторные инструкции [19], оба из которых предназначены для ускорения общих матричных и векторных операций. в обработке глубокого обучения. Помимо решений в цифровой области, были также предложены ускорители, которые работают в аналоговой области [20] – [22] или в области аналогово-цифровых смешанных сигналов [23] – [25].

К сожалению, эти архитектуры электронных ускорителей начинают сталкиваться с фундаментальными ограничениями в эпоху после принятия закона Мура, когда возможности обработки больше не улучшаются, как это было в последние несколько десятилетий [26]. В частности, электронное перемещение данных по металлическим проводам в этих ускорителях является основным узким местом в полосе пропускания и энергии [27]. Фотонные межсоединения предлагают одно из самых многообещающих решений для преодоления этих проблем с перемещением данных. Фотонные связи уже заменили металлические для передачи информации со скоростью света почти на каждом уровне иерархии вычислений, и теперь рассматриваются возможности интеграции на уровне чипа [28]. Появление кремниевой фотоники, которая обеспечила экономичную интеграцию оптических компонентов на основе производства КМОП-электроники, стало одним из основных катализаторов создания фотонных межсоединений в масштабе кристалла [29].

Еще более примечателен тот факт, что различные вычисления, необходимые для глубокого обучения, такие как умножение матрицы на вектор, могут выполняться полностью в оптической области [30]. Таким образом, мы близки к тому моменту, когда станет возможным реализовать ускорители глубокого обучения, использующие кремниевую фотонику как для связи, так и для вычислений. Такие ускорители глубокого обучения на основе кремниевой фотоники могут обеспечить беспрецедентный уровень энергоэффективности и параллелизма. Например, с помощью операций умножения и накопления (MAC), которые доминируют в вычислениях с глубоким обучением, ускорители на основе фотоники могут достичь эффективности использования энергии (определяемой как (MAC / s / mm2) / (joules / MAC)), которая почти в 1000 раз лучше по сравнению с до самых энергоэффективных электронных ускорителей на сегодняшний день [31]. Более того, рабочая полоса пропускания фотонных MAC может приближаться к скорости фотодетектирования, обычно в диапазоне сотен ГГц. Это намного превосходит современные электронные системы, работающие с тактовой частотой несколько ГГц [32].

В этой статье мы исследуем ландшафт кремниевой фотоники для ускорения обучения и вывода моделей глубокого обучения. Предыдущие исследования по родственной теме были сосредоточены либо на изучении производительности и энергетических аспектов определенного типа архитектуры фотонной нейронной сети (например, архитектуры резервуарных вычислений [33] – [35] и архитектуры широковещательной передачи (B&W)) [31], [36] – [38]), или создали упрощенную классификацию на основе реализованных моделей нейронных сетей (например, MLP, CNN) [39]. В отличие от этого, в этой статье мы предлагаем другое

и более полное руководство по развитию ускорения глубокого обучения на основе кремниевой фотоники с восходящей классификацией по абстракциям уровня проектирования: от альтернатив и устройств производства более низкого уровня до спектра нейронов, микроархитектур, и охватывающих множество архитектур интегрированных нейронных сетей на системном уровне. Наша цель – предоставить обзор множества вариантов дизайна, доступных для кремниевой фотоники, для реализации ускорителей фотонного глубокого обучения, а также обсуждение их преимуществ и ограничений. Возможность использовать CMOS-совместимые материалы, такие как германий (Ge) и нитрид кремния (SiN), позволила создать новые варианты фотодиодов, модуляторов, ответвителей и лазеров с очень интересным компромиссом между характеристиками, энергетической надежностью и надежностью. Эти компромиссы также существуют для различных основных типов устройств, таких как интерферометры Маха – Цендера (MZI) и микрокольцевые резонаторы (MR), которые могут использоваться в качестве строительных блоков фотонных искусственных нейронов. Было предложено множество различных типов микроархитектур искусственных нейронов на основе фотоники, таких как некогерентная черно-белая архитектура [40] и когерентный искусственный линейный нейрон (COLN) [36]. Такие нейроны могут быть объединены в каскад, соблюдая профили потерь фотонного сигнала и целевые показатели отношения сигнал / шум (SNR), для создания более крупных структур нейронных сетей на основе фотоники. Мы считаем, что такая классификация по абстракциям дизайна снизу вверх обеспечивает интуитивно понятный и полезный способ понять возможности и ограничения парадигмы кремниевой фотоники в контексте ускорения глубокого обучения.

Остальная часть этой статьи организована следующим образом. Раздел 2 начинается с краткого обсуждения моделей глубокого обучения. В разделе 3 представлен обзор фундаментальных кремниевых фотонных устройств, которые широко используются в фотонных нейронных сетях и актуальны для ускорения моделей глубокого обучения. Раздел 4 описывает различные типы архитектур искусственных нейронов, разработанные с использованием кремниевых фотонных компонентов. Эти нейронные архитектуры образуют строительные блоки архитектур фотонных нейронных сетей, которые обсуждаются в разделе 5. Наконец. Раздел 6 завершается обсуждением выдающихся проблем и возможностей кремниевой фотоники для ускорения глубокого обучения.

2 ОБЗОР ГЛУБОКОГО ОБУЧЕНИЯ

Глубокое обучение – это подмножество машинного обучения (ML), которое само по себе является подмножеством более широкой области ИИ. Глубокое обучение направлено на имитацию глубинной архитектуры человеческого мозга, в котором миллиарды взаимосвязанных нейронов действуют как вычислительные единицы. Человеческий мозг также работает иерархически, начиная с более простых концепций, а затем комбинируя их для изучения более абстрактных идей. Этот способ обучения отражен в моделях глубокого обучения, которые разбивают входные данные на функции, а затем рекомбинируют их для выполнения поставленной задачи (например, обнаружения, классификации). После того, как соответствующие функции были изучены с помощью модели глубокого обучения на этапе обучения, модель может быть применена к задачам аналогичного характера без вмешательства человека. Как упоминалось ранее, в последние годы глубокому обучению уделяется много внимания. Но концепция не нова. Идея сделать машины такими же умными, как люди, лежит в основе аналитического механизма, задуманного Чарльзом Бэббиджем в 1837 году. Область искусственного интеллекта и исследования по созданию машин, способных думать, как люди, начались еще в середине 20 века с вычислительной моделью нейронных сетей и работы нейронов, разработанной Уорреном МакКаллоком и Уолтером Питтсом в 1943 году [8]. Алгоритм персептрона был изобретен психологом Фрэнком Розенблаттом в его основополагающей статье 1957 года [9]. Используя этот алгоритм, Розенблатт создал первый однослойный персептрон (см. Рисунок 1 (a)), который представляет собой электронное вычислительное устройство, соблюдающее биологические принципы функционирования человеческого мозга.

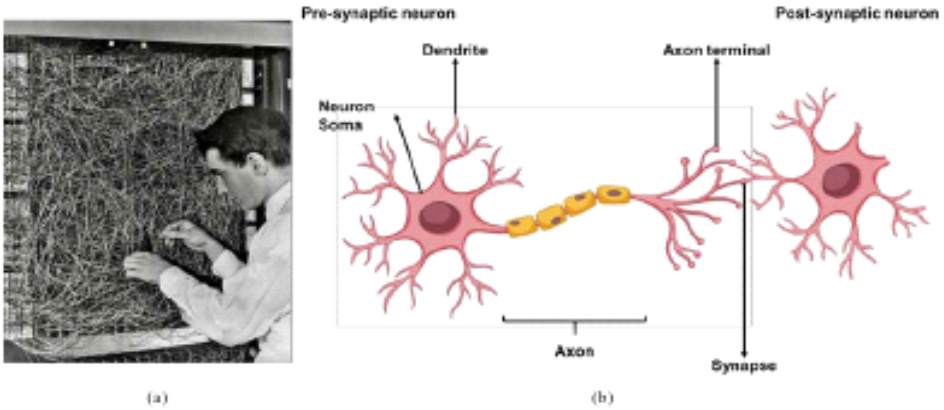


Рисунок 1: (a) Франк Розенблатт со своим однослойным перцептроном Mark-1; (б) Изображение нейрона и синаптической связи с другим нейроном. Это упрощенное изображение нейрона, показывающее только самые основные компоненты. (a) (b)

2.1. Модели нейронов.

Человеческий мозг состоит из около 86 миллиардов нейронов [41], каждый из которых соединен между собой дендритами и аксонными связями (см. рисунок 1 (b)). Биологический нейрон, который является основным обрабатывающим компонентом мозга, состоит из сомы, дендрита, аксона и синапса. Сома или клеточное тело нейрона содержит ядро и другие структуры, общие для живых клеток. Эти структуры поддерживают химический процесс внутри нейрона. Дендриты являются продолжением сомы нейрона и действуют как приемники или входы в нейрон. Аксоны образуют «хвосты» нейрона и несут сигналы от сомы. Аксон может далее разделиться на ветви для достижения невероятной взаимосвязи. В зависимости от типа нейрона эта взаимосвязь может достигать до 100 000 разветвленных соединений, чего сегодня немыслимо достичь с помощью логических вентилей КМОП. Связь между нейронами через их конечности происходит в точках контакта, называемых синапсами (рисунок 1 (b)). Нейронные сигналы передаются в виде электрических импульсов по этим взаимосвязям, состоящим из дендритов, аксонов и синапсов. Эти связи между нейронами вдоль синапсов могут со временем усиливаться или ослабляться в зависимости от активности синапсов. Это называется синаптической пластичностью. Также предполагается, что синаптическая пластичность является ключевым компонентом кодирования воспоминаний в мозге [42]. Модель Маккаллоха-Питтса представляет собой очень упрощенную модель этого биологического нейрона [8]. В его состав входят: esoMwfi tah seuamchm iantipount uansits iagnnde dt hae nw ae itghhrte svhaollude.g aTthee, apsr osdhuocwtsn ionf ftihgeu rien p2.u tTs hae nsdu tmhemira ticoonsensor, ruesn эта сумма передается на пороговый вентиль. Если суммированный сигнал xcCeueldlosc thh-eP ittthsr esmhoodledl, tahdea pgtaetde ga elinneeraatr etsh rae ssihgonlda l faorn dt htehire tnheruersohno lgd egnaetrea, t esso (tohr efi rnee) в зависимости от того, какой продукт производит расширение, производится это двоичный выходной нейрон. В более современных терминах этот линейный порог в модели называется функцией активации модели.

Нейрон срабатывает только тогда, когда значение суммы пересекает пороговое значение T, что делает модель Мак-Каллока-Питтса двоичным выходным нейроном. Эта бинарная модель является мощным инструментом и может использоваться для решения простых задач бинарной классификации. Но для более сложных задач необходимы более сложные функции активации и модели нейронов. Существуют и другие модели нейронов, имитирующие биофизические характеристики нейрона, такие как модель Ходжкина-Хаксли [43] и многие другие [44] – [46].

Использование таких моделей требует сложных вычислений их биологических взаимодействий, что может потребовать больших вычислительных затрат. Чтобы обойти это, предпочтительны эффективные с вычислительной точки зрения нейроны с интеграцией и запуском (IF). Между прочим, ПФ нейроны являются одной из самых старых моделей нейронов, появившихся в литературе [47]. Нейрон Leaky Integrate-and-Fire (LIF) [48] – еще одна чрезвычайно популярная модель нейрона из-за ее простоты, позволяющей реализовать сложные функции в глубоких нейронных сетях. Другая модель нейрона, которая имитирует биофизические характеристики нейрона, во многом похожую на модель Ходжкина-Хаксли, но с меньшей вычислительной сложностью, – это модель нейрона с импульсами Ижикевича [49].

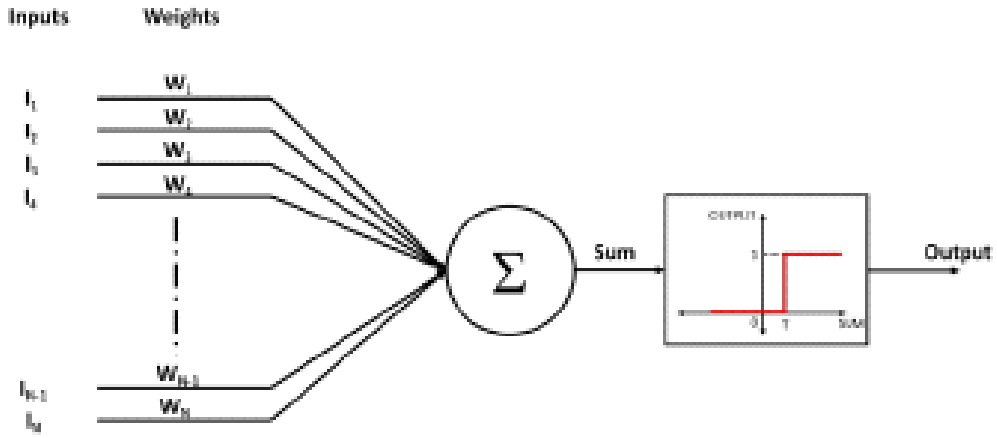


Рисунок 2: Вычислительная модель биологического нейрона Мак-Каллока-Питтса [8]. В модели используется линейная функция активации.

Все эти модели нейронов следуют одному и тому же основному принципу: нейроны принимают входные сигналы от нескольких синапсов, суммируют их и запускают соответствующий выходной сигнал, если порог превышен. Различия между ними возникают в том, как моделируются пороговые и биофизические взаимодействия. Обсуждаемые здесь модели нейронов помогают имитировать биологическую работу мозга и, следовательно, являются неотъемлемой частью модели нейронных сетей, называемой Spiking Neural Networks (SNN), которую мы обсудим дальше. Эту междисциплинарную концепцию имитации мозга с использованием продвинутых моделей нейронов и реализации нейронных систем часто называют нейроморфной инженерией или нейроморфными вычислениями.

2.2 Пиковые нейронные сети (SNN)

Идея, лежащая в основе Spiking Neural Networks (SNN), состоит в том, чтобы как можно точнее имитировать человеческий мозг. Мозг демонстрирует низкое энергопотребление, быстрый вывод, обработку со-

бытий, непрерывное обучение и массивный параллелизм. Он также основан на вычислении на основе событий, когда информация кодируется в пиках [50]. Действительно, SNN были введены в 1997 году для имитации этого метода вычислений, основанного на шипах [48]. SNN используют асинхронную обработку, управляемую событиями, для реализации нейронных сетей. Входы в нейрон SNN называются потенциалами действия или пиками (см. Рисунок 3), которые нейрон получает от своего пресинаптического нейрона. Эти двоичные пики могут переносить информацию через сеть либо посредством кодирования скорости, либо посредством временного кодирования. Кодирование скорости, также называемое частотным кодированием, представляет собой модель возбуждения нейронов, которая предполагает, что информация о стимуле, вызвавшем срабатывание нейрона, может быть закодирована с частотой, с которой нейрон срабатывает. Таким образом, этот способ кодирования информации требует точного расчета скорострельности. Временное кодирование использует временное разрешение или время между последовательными всплесками для передачи информации. Для обоих типов кодирования связь между нейронами представлена синаптическими весами, которые влияют на входные пики, чтобы создать взвешенную серию пиков на входе каждого нейрона. Взвешенные всплески входного сигнала влияют на мембранный потенциал нейрона, который относится к интенсивности активации нейрона. Как только мембранный потенциал превышает пороговое значение, нейрон генерирует импульс (то есть запускает потенциал действия) в свой постсинаптический нейрон. Эта деятельность проиллюстрирована на рисунке 3. В последнее десятилетие возрос интерес к реализации мозговых вычислений [51], чтобы преодолеть ограничения, установленные обычными архитектурами фон-Неймана. Суперкомпьютеры сегодня могут достигать сотен пета FLOPS (операций с плавающей запятой в секунду) при обработке данных, но за счет десятков миллионов ватт [52], тогда как человеческий мозг достигает этого за счет всего 20 ватт [53]. Реализации SNN надеются достичь такого замечательного уровня энергоэффективности, который демонстрирует человеческий мозг. Для реализации этой цели активно исследуются многие технологии, включая КМОП [54] – [56], новые типы транзисторов [57] – [59] и энергонезависимую память [60] – [63]. Используя такие технологические достижения, были реализованы различные ускорители SNN. Например, SpiNNaker [64] из Манчестерского университета был построен с использованием процессоров ARM и реализует модель нейрона Ижикевича для повышения вычислительной эффективности. Он использует глобально асинхронную, локально синхронную (GALS) систему связи между ядрами обработки и синхронной динамической памятью произвольного доступа (SDRAM) для хранения

значений синаптических весов. TrueNorth [65] от IBM содержал 5,4 миллиарда транзисторов, при этом потребляя всего 70 мВт для работы. Процессор состоит из массивов нейросинаптических процессоров с низким энергопотреблением.

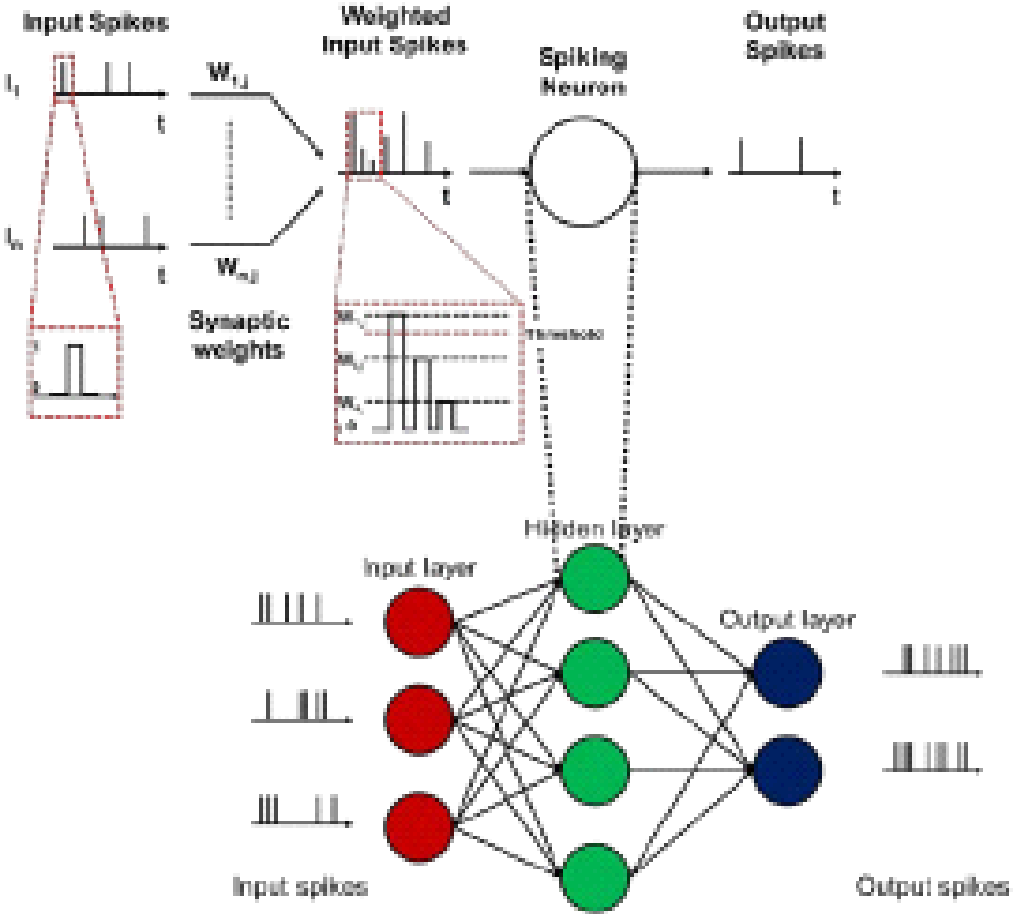


Рисунок 3: Простое представление о том, как работает нейрон с импульсами, основная единица SNN. Показанный здесь импульсный нейрон может быть любой из моделей, упомянутых в разделе 2.1, для реализации SNN.

каждый из которых содержит память, 6 процессоров и подсистемы связи для имитации нейронных функций. TrueNorth реализует модель нейрона LIF в своем SNN. Loihi [66] от Intel со 128 нейроморфными ядрами и 130 000 нейронов – еще одна такая реализация, которая продемонстрировала 1000-кратную скорость и 10 000-кратную энергоэффективность по сравнению с ЦП [67]. Loihi реализует вариант нейрона LIF, который называется «нейрон LIF на основе текущего синапса (CUBA)».

2.3. Искусственные нейронные сети (ИНС)

Появление вычислительной модели для представления нейронной активности проложило путь для искусственных нейронных сетей (ИНС). По сравнению с SNN, ANN заметно абстрактны в своем подходе к реализации функций мозга. Веса, которые представляют синаптиче-

скую пластичность, представляют собой простые скаляры. Используемые нейроны также намного проще, и им поручено накапливать входные весовые продукты с последующей передачей результирующих выходных данных через нелинейную функцию. ИНС имитируют активность мозга, моделируя набор взаимосвязанных нейронов, расположенных слоями. В простейшем представлении будет три слоя: входной, выходной и скрытый между этими двумя слоями (рисунок 4 (a)). Входной уровень принимает данные извне ИНС; скрытый слой – это место, где происходят вычисления; а выходной слой – это то, где мы можем получить результаты от нейронной сети. Функции активации также играют важную роль в моделировании интеллекта. Математически, без соответствующих функций активации, модель нейрона представляет собой простую линейную модель, которая умножает и накапливает продукты входного веса. Чтобы внести в сеть нелинейность и дать модели возможность аппроксимировать более сложные функции, нам необходимо использовать соответствующие нелинейные функции активации, такие как сигмоид, выпрямленный линейный блок (ReLu) и tanh, чтобы перечислить некоторые из них. Используя эти функции, ИНС могут изучать очень сложные нелинейные отношения между входными функциями.

Здесь следует провести важное различие между ИНС и традиционными алгоритмами машинного обучения, такими как вспомогательные векторные машины (SVM), K-ближайших соседей (KNN), случайные леса (RF) и т. Д. ИНС отличает их способность обрабатывать большие количество данных с минимальным вмешательством человека. Традиционные алгоритмы машинного обучения обычно требуют, чтобы человек-эксперт предоставил необходимый набор правил, по которым они работают. Часто также требуется помощь для извлечения функций из данных, например, для выбора ядра в SVM. Глубокие нейронные сети (DNN) – это ИНС с несколькими скрытыми слоями (см. Рисунок 4 (b)), которые могут использовать сложную взаимосвязь между нейронами для вычисления и эффективного представления очень сложных нелинейных отношений после обучения. Во время фазы обучения активации входа проходят прямой путь от входа к скрытым слоям и, наконец, к слою вывода.

Ошибка (часто называемая потерей) между выходом DNN 7 и ожидаемым выходом распространяется по модели в обратном направлении для обновления весов и смещений нейронов таким образом, чтобы уменьшить потери. Этот процесс итеративно повторяется до тех пор, пока выходные данные модели (например, прогноз класса изображения) не станут как можно ближе к ожидаемым выходным данным, то есть потери не будут минимизированы. После обучения модель может делать прогнозы на основе входных данных, что называется фазой вывода. Фа-

за обучения DNN – это процесс, требующий больших затрат времени и ресурсов по сравнению с фазой вывода. Известные архитектуры обучения, в которых используются DNN, включают в себя многоуровневые персептроны (MLP), рекуррентные нейронные сети (RNN), машины глубокого Больцмана (DBM), составные автокодеры (SAE) и сверточные нейронные сети (CNN). MLP включают только слои с прямым соединением (FC), как показано на рисунке 4 (b), где каждый нейрон в слое подключен к каждому нейрону в предыдущем и последующем слоях. Некоторые архитектуры моделей могут демонстрировать временное динамическое поведение и иметь внутреннее состояние или память из-за своей сетевой структуры. Их в широком смысле можно назвать рекуррентными нейронными сетями (RNN). Внутренняя память в их структуре делает их идеальными для задач распознавания, таких как распознавание образов, почерка и распознавания речи, обработка естественного языка и т. д. Исследования RNN начались с Дэвида Румехальта в 1986 году [68]. На сегодняшний день популярны многие варианты RNN, в том числе сети с долгосрочной памятью (LSTM), стробированные рекуррентные блоки (GRU), RNN с непрерывным временем (CTRNN) и т. Д.

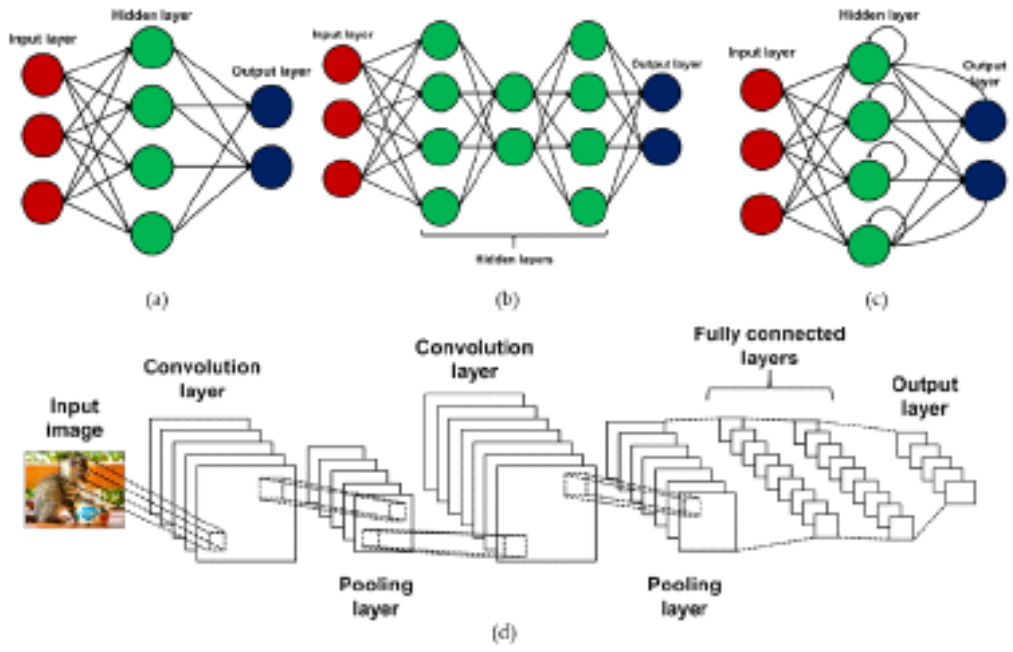


Рисунок 4: (a) Многоуровневая архитектура ИНС; на этом рисунке показана неглубокая ИНС с одним скрытым слоем. (b) По мере увеличения количества скрытых слоев в ИНС мы получаем DNN. Обратите внимание, что слои ИНС заполнены нейронами с соответствующими весами и смещениями. (c) Полностью подключенная RNN, которая показывает обратные связи в своем скрытом слое и моделирует память или сохраненные состояния в этой архитектуре. (d) Представление различных уровней и операций в CNN.

Упрощенная RNN показана на рисунке 4 (c). CNN нацелены на обработку двумерных или более многомерных функций вместо одномерных в MLP. Они широко используются для задач классификации при

обработке изображений и видео. Структура CNN изображена на рисунке 4 (d). Типичная CNN содержит три типа уровней: сверточный (Conv), объединенный (Pool) и полностью связанный (FC). В слоях Conv и Pool есть несколько каналов (называемых картами функций), которые извлекают различные локальные функции из входных данных. Эти слои объединяют функции более низкого уровня из нескольких каналов предыдущего слоя в функции более высокого уровня, которые передаются на следующий уровень, до последнего слоя классификации, на котором создается прогноз вывода. Слои Conv имеют гораздо меньше параметров, чем слои FC, но требуют больших вычислительных ресурсов из-за множества операций свертки, которые требуются между весовыми коэффициентами фильтра и активациями ввода по всем их каналам. Слои объединения генерируют активацию выходных данных только на основе локального воспринимающего поля в соответствующей карте входных характеристик (например, один «объединенный» выход из группы входов 2x2). Два широко используемых варианта объединения уровней – это максимальное и среднее объединение, и они производят максимальное или среднее значение каждого воспринимающего поля соответственно. Наконец, слои FC следуют за слоями Conv и Pool и действуют как классификатор с извлеченными функциями, подобно тому, как эти слои используются в MLP. DNN начинают широко использоваться в реальных приложениях, таких как автономное вождение, робототехника и обработка данных в Интернете вещей. Ресурсоемкость, необходимая для обучения моделей DNN, была удовлетворена благодаря появлению графических процессоров, которые используются для значительного сокращения времени обучения DNN из-за большего параллелизма на уровне данных и потоков, поддерживаемого в GPU, чем в CPU. Как и в случае с SNN, растет интерес к разработке энергоэффективных ускорителей ASIC для DNN. Такие ускорители DNN, например, Neural Processing Unit (NPU) [69], предназначены для ускорения фазы вывода, хотя некоторые ускорители также нацелены на повышение эффективности обучения. Примером ускорителя DNN, который оказался очень успешным для ускорения обучения и вывода с помощью DNN, является TPU [12] от Google. TPU имеет специальные блоки матричного умножения и управление распределенной памятью, что делает его идеальным для выполнения тяжелой работы, необходимой для обучения моделей DNN, а также для задач вывода. TPU широко используются в центрах обработки данных Google. В новых архитектурах GPU также используются аналогичные ядра Tensor для ускорения DNN [70]. Исследователи также предложили использовать технологию энергонезависимой памяти и обработку в памяти (PIM) для ускорителей DNN. PRIME [71] и ISAAC [20] являются примерами таких ускорителей, которые используют рези-

стивную память с произвольным доступом (ReRAM) и PIM для ускорения выполнения DNN.

2.4 Резервные вычисления (RC)

RC – менее популярная модель нейронной сети, чем ANN и SNN, но здесь кратко рассматривается из-за ее пригодности для реализации на основе фотоники и рассмотрения в предшествующих проектах на основе фотоники. RC можно рассматривать как тип RNN, в котором обучаются только параметры последнего неповторяющегося выходного уровня (называемого слоем считывания), в то время как все остальные параметры инициализируются случайным образом при соблюдении некоторого условия, которое по существу предотвращает хаотическое поведение, и тогда они остаются необученными. Таким образом, RC представляет собой тип частично адаптивной RNN, который контрастирует с полностью адаптивным подходом традиционных ANN и SNN. Резервуар состоит из связанных нелинейных узлов и представляет собой фиксированную рекуррентную сеть, как показано на рисунке 5. Этот резервуар выполняет множество нелинейных операций, и выходные данные из них объединяются в линейные комбинации для выполнения задачи. У пользователя мало прямого доступа к резервуару, а манипуляции с выходными данными ограничены слоем считывания. Чтобы достичь желаемого поведения, обученные линейные классификаторы на уровне считывания используются в среде обучения с учителем. Преимущество такой фиксированной случайной сети становится очевидным с некоторыми аппаратными платформами (в частности, на основе фотоники), где невозможно установить все внутренние параметры.

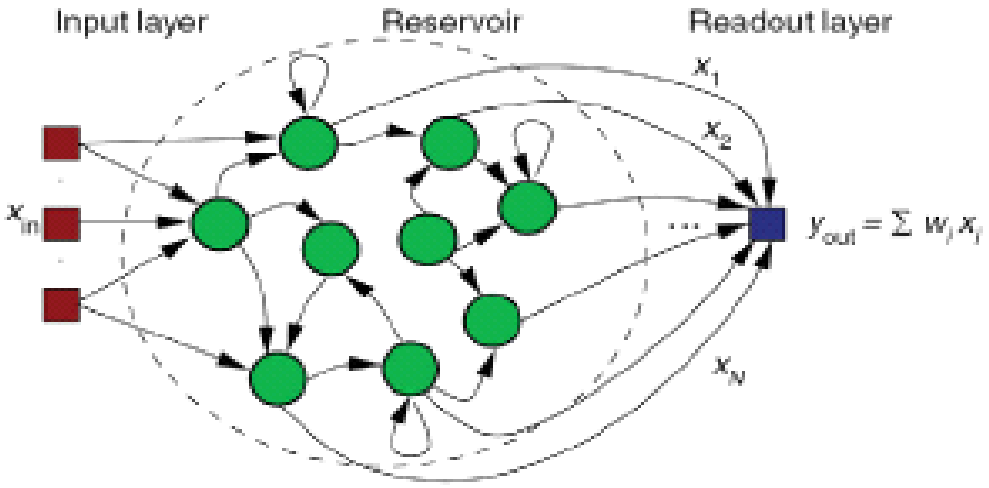


Рисунок 5: Стандартная схема архитектуры резервуарных вычислений (RC) с входным слоем красным цветом, резервуаром зеленым (со случайными, но фиксированными соединениями) и слоем считывания синим цветом, где выходные данные резервуара объединяются в желаемые выход.

RC может использоваться для имитации поведения обычных ИНС из-за его внутренней параллельности. Как и нейронная сеть, резервуар часто состоит из большого количества взаимосвязанных нелинейных узлов. Следовательно, существующие аппаратные реализации нейронных сетей могут использоваться и использовались в качестве резервуаров, как в [72]. Однако, в отличие от традиционных нейронных сетей, веса межсоединений не обязательно должны быть адаптируемыми или даже точно управляемыми. Фактически, для изменения веса в RC требуется только глобальное масштабирование усиления. Это делает требования к реализациям резервуаров более мягкими и позволяет исследовать технологии, которые могут быть менее подходящими для реализации традиционных, полностью обучаемых нейронных сетей. Таким образом, RC был популярной целью для ранней реализации полностью оптических вычислений, и было много настольных моделей, которые продемонстрировали, как могут быть достигнуты полностью оптические вычисления резервуара [73] – [81]. Эти реализации часто создавались с использованием телекоммуникационного оборудования (например, волоконно-оптических контуров, MZI, лазеров, фотодетекторов и решетчатых волноводов (AWG)) и служили доказательством проверки концепции эффективности вычислений оптического резервуара. Для реализации RC и других моделей DNN на вычислительном чипе кремниевая фотоника является многообещающим кандидатом на новые технологии. Теперь мы сделаем обзор технологии кремниевой фотоники (раздел 3), а затем подробно обсудим микроархитектуру нейронов, реализованную с использованием этой технологии (раздел 4), и различные архитектуры глубокого обучения, построенные с использованием фотонных нейронов (раздел 5).

3 ОБЗОР КРЕМНИЕВОЙ ФОТОНИКИ

Оптическая связь широко используется в сетях связи, где требуется недорогая и широкополосная связь при низком энергопотреблении и на больших расстояниях, например, в сетях дальней связи. В последние годы кремниевая фотоника сделала возможным использование CMOS-совместимой интегрированной фотоники и получила широкое распространение в коммерческих предложениях для недорогих оптических межсоединений в центрах обработки данных. В настоящее время активно рассматриваются оптические межсоединения в гораздо меньших масштабах, чтобы соединить несколько процессоров на уровне платы и даже для соединения ядер в одном вычислительном чипе. Как следует из названия, в кремниевой фотонике для связи используется свет, который направляется через кремниевую (Si) среду на кристалле CMOS. В платформе для изготовления кремний-на-изоляторе (КНИ) высокий контраст

показателя преломления между сердцевиной волновода (кремний) и оболочкой волновода и подложкой (например, диоксидом кремния) приводит к распространению направленного оптического сигнала за счет полного внутреннего отражения. Один волновод может использоваться для одновременного переноса нескольких длин волн света, каждый из которых может передавать данные с высокой скоростью и высокой частотой без каких-либо помех. Это возможно с помощью метода, называемого мультиплексированием с разделением по длине волны (WDM). Количество длин волн в волноводе называется степенью WDM волновода. Степень WDM может быть увеличена до 64 и выше, после чего мультиплексирование часто называют плотным мультиплексированием с разделением по длине волны (DWDM)

Для связи в масштабе чипа с кремниевой фотоникой цифровые данные от электронных компонентов (например, процессора, памяти) могут быть закодированы в оптический сигнал с использованием электронного преобразования в оптическое (E / O) с помощью таких устройств, как модуляторы микрокольцевого резонатора (MR), впоследствии передается по волноводу с несколькими длинами волн несущей, а затем обнаруживается на приемнике, где оптическое преобразование в электронное (O / E) выполняется с помощью таких устройств, как фотодетекторы (PD). Растет интерес к использованию кремниевой фотоники не только для связи. В частности, кремниевые фотонные устройства также могут использоваться для выполнения вычислений в оптической области. Вместе такие легкие коммуникации и вычисления могут значительно ускорить выполнение рабочих нагрузок глубокого обучения. 9

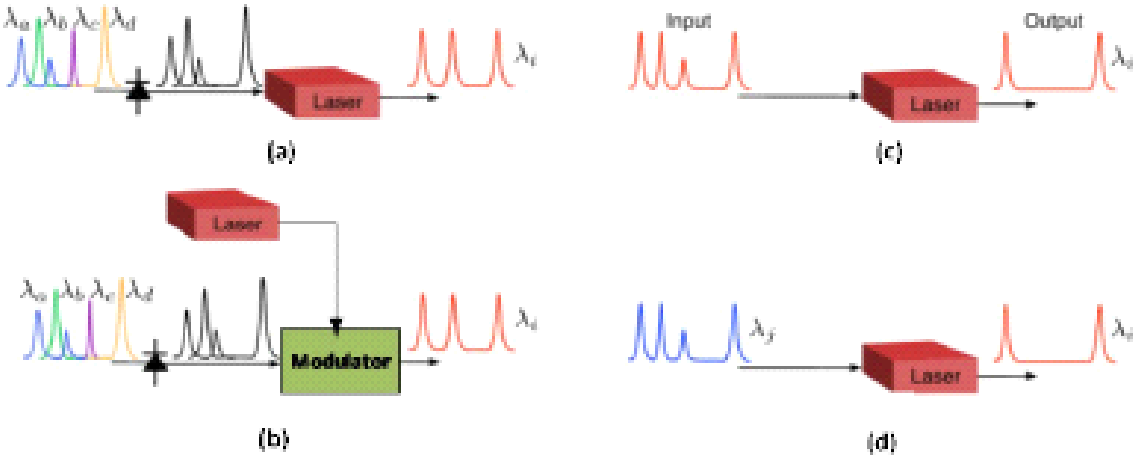


Рисунок 6: Лазеры, используемые в фотонных нейронных сетях [83], [84]: (a) лазер с прямой модуляцией; (б) лазер, подключенный к модулятору [83], [85]; c) когерентный лазер с одинаковой длиной волны на входе и выходе; и (d) некогерентный лазер с разными длинами волн на входе и выходе.

Хотя кремниевые фотонные устройства сталкиваются с рядом проблем, связанных с надежными вычислениями и связью на уровне кри-

сталла (например, они чувствительны к изменениям температуры и производственного процесса [82]), они также предлагают несколько преимуществ (например, высокую скорость, широкую полосу пропускания, и низкое энергопотребление) для поддержки межнейронных коммуникаций и реализации различных нейронных функций, необходимых в фотонных нейронных сетях. Такие нейронные функции и их реализации обсуждаются в следующем разделе. В этом разделе мы рассмотрим некоторые из фундаментальных кремниевых фотонных устройств, которые используются для реализации ИНС и СНС на основе фотоники. Обратите внимание, что форма сигнала, обозначенная черным цветом, является электрическим сигналом, а формы сигналов других цветов – оптическими. Разные цвета представляют разные длины волн.

3.1. Лазеры

Лазер – ключевое требование в оптических схемах и нейронных сетях, служащий источником света для поддержки оптической связи и вычислений. Лазеры могут быть как внутри кристалла, так и вне его. Хотя внешние лазеры обеспечивают лучшую эффективность излучения света, они требуют использования ответвителей для передачи внешнего оптического сигнала на кристалл, где такие ответвители вызывают большие потери оптической мощности. С другой стороны, встроенные в кристалл лазеры обеспечивают лучшую плотность интеграции и более низкие оптические потери, так как нет необходимости связывать свет от внешнего источника. Однако встроенные лазеры страдают низкой эффективностью излучения и нестабильностью по отношению к тепловым изменениям [86]. Лазеры используются в фотонных нейронных сетях для реализации различных нейронных функций и требований в таких системах. В лазерах с прямой модуляцией (см. Рисунок 6 (a) [87]) лазер сам модулирует данные в виде оптического сигнала, в то время как в другом устройстве, как показано на рисунке 6 (b), выходной сигнал лазера может модулироваться модулятором. который отвечает за модуляцию данных на оптический сигнал. Действительно, использование лазера в сочетании с модуляторами является обычным явлением в оптических сетях межсоединений [88], [89]. В фотонных нейронных сетях эта лазерная конфигурация может использоваться для разработки масштабируемой нейронной сети [83], [85], где внешний лазерный источник в сочетании с модуляторами может поддерживать множественные внутрикристальные межнейронные коммуникации. В дополнение к модуляции, лазеры также могут использоваться для реализации функций нейронной активации [90], как обсуждалось в разделе 4, потому что лазеры показали потенциал для имитации функций нейронной активации [91] – [94], где оптический стимул на входе лазера может привести к оптическому выходу на основе функции активации (см. рисунки 6

(c) и 6 (d)). Таким образом, лазеры используются во всех фотонных нейронных сетях не только для поддержки межнейронной связи, но также, в некоторых случаях, для реализации различных нейронных функций в оптической области. Лазер можно реализовать по-разному. Лазер с вертикальным резонатором, излучающий поверхность (VCSEL), представляет собой полупроводниковый лазерный диод с излучением лазерного луча перпендикулярно поверхности кристалла, как показано на рисунке 7 (a). Такая функция позволяет размещать несколько VCSEL в массиве для питания большого количества оптических нейронов и, следовательно, создавать масштабируемые нейронные сети [91], [92], [97]. В дополнение к преимуществу масштабируемости нейронные сети на основе VCSEL могут быть реализованы с использованием как внешних, так и внутрикристальных VCSEL [98]. Более того, VCSEL показали поведение нейронов по возбудимости [99], при котором лазер излучает свет, когда комбинация входных сигналов достигает порогового значения. В результате VCSEL предлагают масштабируемость, эффективность и несколько функций, необходимых для проектирования фотонных нейронных сетей. Лазеры на микродисках, показанные на рисунке 7 (b), представляют собой другой тип лазеров, в которых кольцевой резонатор образован последовательными полными внутренними отражениями внутри волновода круглой формы [101]. По сравнению с VCSEL лазеры на микродисках более эффективны по площади (радиус лазерного устройства – радиус микрополости – несколько микрон) и обеспечивают более низкий пороговый ток и максимальную оптическую мощность на кристалле [102]. Более того, лазеры на микродисках предлагают низкие оптические потери [100], и, подобно массивам VCSEL, они могут быть помещены в массив лазеров [101], что позволяет реализовать масштабируемые реализации фотонных нейронных сетей. Кроме того, лазеры на микродисках продемонстрировали динамику возбудимости [103] для поддержки реализации импульсных нейронов. (a) (b)

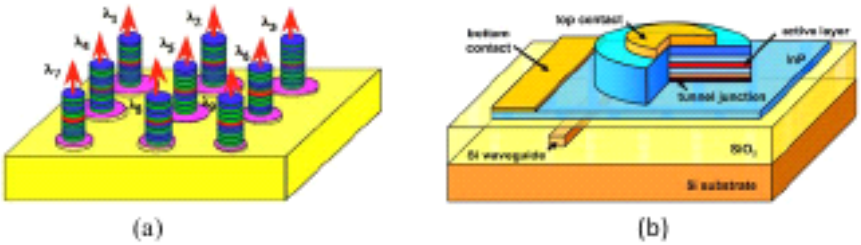


Рисунок 7: (а) Массив VCSEL [95]. (б) Лазер на микродиске [96].

3.2 Волноводы

Кремниевый фотонный волновод аналогичен металлическому проводу, что позволяет передавать и маршрутизировать оптические сигналы в фотонных нейронных сетях. Как показано на рисунке 8, волноводы

можно разделить на гребневые и полосовые. Гребневые волноводы часто используются в активных устройствах и сетях, поскольку они позволяют выполнять электрические соединения с волноводом (например, через PN-переходы), где характеристики оптического сигнала можно активно контролировать и изменять с помощью электрооптических или термооптических эффекты в кремнии [28]. С другой стороны, полосковые волноводы обычно используются в пассивных устройствах и сетях для пассивной маршрутизации оптических сигналов [28]. Как обсуждалось ранее, один волновод может поддерживать одновременную передачу нескольких длин оптических волн без помех (с использованием WDM). Это обеспечивает связь со сверхвысокой пропускной способностью, что представляет большой интерес в дизайне нейронных сетей для поддержки требовательной межнейронной связи. Когда оптический сигнал проходит через волновод, он испытывает некоторые оптические потери (т. Е. Потери при распространении, часто измеряемые в дБ / см), вызванные, например, некоторыми недостатками в структуре волновода (например, шероховатостью боковой стенки волновода). Сведение к минимуму таких оптических потерь в кремниевых фотонных волноводах имеет важное значение, поскольку оно ограничивает масштабируемость фотонных нейронных сетей и существенно снижает мощность и энергоэффективность в таких сетях. Было предпринято много усилий для минимизации потерь распространения в кремниевых фотонных волноводах и волноводах с КНИ с потерями на распространение всего 0,026 дБ / см [104]. В общем, эти потери на распространение в волноводах зависят от точной настройки геометрии в этих устройствах, и, следовательно, любое искажение формы в волноводе (например, угловые боковые стенки) снижает его эффективность передачи (то есть увеличивает потери на распространение). Помимо потерь при распространении, изгибы волновода создают потери на оптических изгибах, когда оптический сигнал будет ослабляться из-за рассогласования мод и потерь излучения в изгибах волновода. Эти потери на изгибе пропорциональны радиусу изгиба волновода.

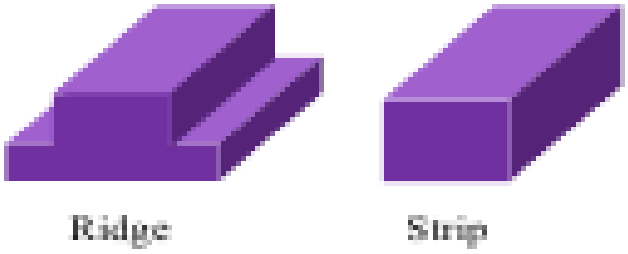


Рисунок 8: Кремниевый фотонный волновод [2e8]: гребенчатый волновод и полосовой волновод.

3.3 Ответвители

Кремниевые фотонные ответвители, как показано на рисунке 9, используются для передачи оптического сигнала от оптического волокна (например, подключенного к внешнему лазерному источнику) во встроенный волновод из-за значительного несоответствия между перекрестными сечение оптических волокон (десятки микрон) и кремниевых фотонных волноводов (сотни нанометров). Такое рассогласование обычно приводит к некоторым оптическим потерям (т. е. Потерям связи), которые рассматриваются как значительная часть оптических потерь в оптических сетях, использующих лазеры вне кристалла. Двумя основными решениями для соединения являются соединение поверхностной решетки и соединение краев. Соединители с поверхностной решеткой, показанные на рисунке 9 (a), имеют преимущество с точки зрения более простого и недорогого процесса изготовления, но за счет низкой эффективности соединения, в то время как краевые соединители, показанные на рисунке 9 (b), обеспечивают лучшее эффективность соединения, но требует более сложного процесса изготовления и упаковки [28]. В краевых ответвителях, как показано на рис. 9 (a), используется конический волновод для передачи света от волокна к микросхеме. Элементы связи с поверхностной решеткой передают входящий свет из волокна в волновод с помощью дифракционных решеток, в которых периодическая структура разделяет и дифрагирует свет и в конечном итоге направляет свет в волновод. Дифракционная связь является обычным средством оптической связи между VCSEL из-за ее простой реализации [106], а также полезна для реализации вычислений фотонного резервуара [107].

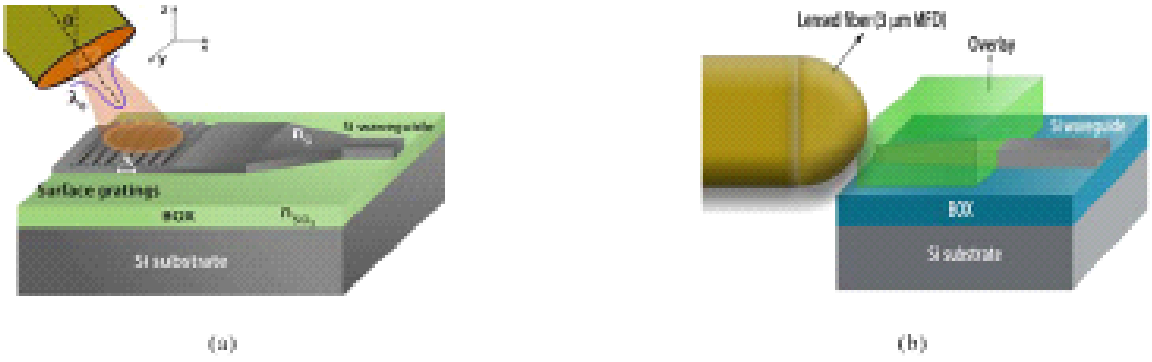


Рис. 9: Элементы связи [105]: (a) элемент связи с поверхностной решеткой и (b) элемент связи с краем.

3.4 Модуляторы, фильтры и переключатели

Микрокольцевые резонаторы (MR) широко используются для разработки модуляторов, переключателей и оптических фильтров в оптических сетях межсоединений [108], [109]. Помимо таких приложений в се-

тях межсоединений, они являются многообещающими устройствами для реализации искусственных нейронных синапсов [90], [110], [111] и функции возбуждения нейронов [112], [113], которые дополнительно обсуждаются в разделе 4. MR, как показано на рисунке 10 (a), сделаны с кольцевым волноводом рядом с входным волноводом и капельным волноводом (также известным как фильтр добавления-капли). Когда капельный волновод отсутствует (например, в модуляторах и некоторых фильтрах), MR является широкополосным фильтром (см. Рисунок 10 (b)). MR может находиться в двух различных состояниях: включенном или выключенном резонансе, в зависимости от которых оптический сигнал может переключаться на разные порты. Как показано на рисунке 10 (a), когда MR находится в выключенном состоянии, входной сигнал направляется в сквозной порт, поскольку кольцо не находится в резонансе с входным оптическим сигналом. С другой стороны, когда MR находится во включенном состоянии, кольцо передает входной оптический сигнал и отбрасывает его на порт вывода. Резонансная длина волны MR может быть настроена для реализации различных функций, необходимых для разработки оптических модуляторов, переключателей и фильтров. Здесь под настройкой понимается изменение резонансной длины волны MR за счет использования электрооптических или термооптических эффектов кремния, которые могут изменять характеристики оптического сигнала и, следовательно, резонансную длину волны в случае MR. По сравнению с механизмами настройки, основанными на электрооптических эффектах, механизмы, основанные на термооптических эффектах, медленнее (несколько микросекунд против десятков наносекунд в методах настройки, основанных на электрооптических эффектах), но более энергоэффективны. На рисунке 10 (c) показан пример модулятора на основе MR, который отвечает за модуляцию электронных данных в оптический сигнал. Модулятор может модулировать электронные данные на определенной длине оптической волны, и этот модулированный оптический сигнал может быть отфильтрован с помощью селективного по длине волны MP-фильтра в приемнике (см. Рисунок 10 (b)), а затем обнаружен и преобразован в электронные данные через фотоприемник. (a) ON drop add OFF in through (b) MR лазерномодулированный оптический сигнал MR Модулятор Данные решетки соединителя clk (c) Драйвер модулятора (e) Направленные ответвители Данные фотодетектора clk (f) add (d) падение через вход фазовращателя Выходной сигнал По сравнению с MR, микродисковый резонатор (см. Рис. 10 (d)), в котором используется диск вместо кольцевой структуры, предлагает лучшее оптическое ограничение, чтобы обеспечить меньший размер диска и потенциально более низкое энергопотребление [114].

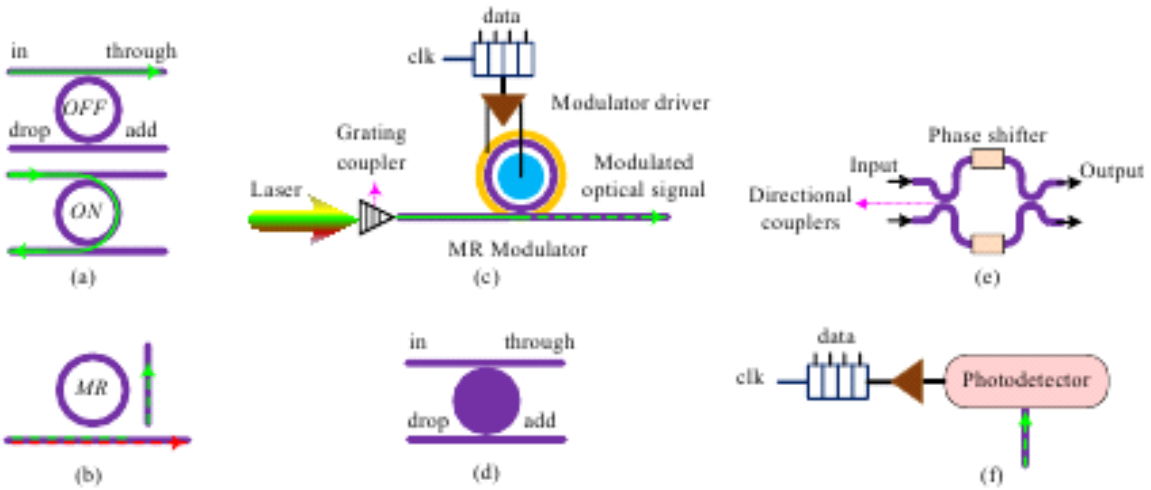


Рис. 10: Кремниевые фотонные коммутационные устройства [28]: (а) добавочно-капельный фильтр / переключатель MR; (б) MP фильтр на все проходы; (с) модулятор MR; (г) Микродисковый резонатор; е) MZI; и f) фотодетектор.

ИМЦ, как показано на рисунке 10 (е), состоят из двух волноводов с направленными ответвителями и 12 фазовращателями. Фазовращатели, реализованные с использованием электрооптической или термооптической настройки, изменяют оптическую фазу в одном или обоих плечах ИМЦ, создавая на выходе конструктивные или деструктивные помехи для переключения оптического сигнала между выходными портами. Подобно MR, MZI применялись при разработке оптических модуляторов, переключателей и фильтров. По сравнению с MZI, MR имеют меньшую площадь основания и меньшее энергопотребление. С другой стороны, MZI обеспечивают широкую полосу пропускания и лучшую устойчивость к температурным колебаниям. Таким образом, MR, микродиски и MZI широко используются для разработки модуляторов, переключателей и фильтров. В фотонных нейронных сетях набор оптических фильтров, в каждом из которых можно регулировать оптическую передачу, может быть сгруппирован в банк весов для поддержки взвешивания сигналов активации как части фотонного нейрона [40].

3.5 Фотодетекторы

Фотодетекторы (ФД), как показано на рисунке 10 (f), могут использоваться для обнаружения оптического сигнала и преобразования его в электрический. Небольшой фотодетектор предлагает широкую полосу пропускания за счет низкой энергоэффективности. Эффективный фотодетектор обеспечивает требуемый электрический выход с небольшим оптическим сигналом на входе. Однако этот слабый оптический сигнал на входе фотодетектора может привести к низкой пропускной способности фотодетектора. Мощность оптического сигнала на входе фотодетектора должна быть больше, чем чувствительность фотодетектора, которая

определяется как электрический выход на оптический вход. Это означает, что мощность лазерного источника в оптической линии связи должна быть достаточно большой, чтобы правильно управлять фотодетектором с учетом суммы различных оптических потерь в линии связи. В конструкции фотонных нейронных сетей фотодетекторы не только преобразуют оптические сигналы в электрические, но также объединяют (т. е. Суммируют величины) нескольких оптических сигналов на разных длинах волн [40], [85], [111], что является полезной функцией в разработка кремниевого фотонного нейрона.

3.6. Устройства на основе материалов с фазовым переходом.

Устройства, в которых для настройки используются материалы с фазовым переходом (PCM), представляют большой интерес в кремниевых фотонных схемах для разработки модуляторов [115], переключателей на основе MZI [116] и фазовращателей с малыми потерями [117]. Основным принципом этих фотонных устройств на основе ИКМ является использование ИКМ (например, GST: Ge2Sb2Te5, использованного в [115]) для эффективного индуцирования высоких изменений показателя преломления для эффективной настройки фазы. В отличие от электро- и термооптической настройки, настройка на основе ПКМ энергонезависима: она требует энергии только для перехода между аморфным и кристаллическим состояниями в ПКМ [117]. Это может позволить снизить накладные расходы на настройку кремниевых фотонных устройств, например, от резонанса до нерезонанса в MR на основе PCM. В некоторых фотонных нейронных сетях устройства на основе PCM [118] – [120] предлагаются как часть дизайна нейронов. Например, в конструкции синапса, предложенной в [118], несколько модулей ИКМ размещаются на волноводе для управления оптической передачей в волноводе и реализации функции синапса. Более того, ИКМ полезны для реализации суммирования [119] и весовых функций [118], [119], [121].

3.7 Другие устройства

Полупроводниковый оптический усилитель (SOA) – это устройство, в котором полупроводник используется для добавления усиления к оптическому сигналу без электрооптических или оптоэлектрических преобразований. SOA в основном используются для компенсации оптических потерь в оптических системах связи. В фотонных нейронных сетях SOA можно использовать для реализации функций обучения [122], [123]. Однако SOA страдают от плохой эффективности связи с оптическими волокнами и чувствительны к поляризации из-за своей плоской формы [124]. Полупроводниковые оптические усилители с вертикаль-

ным резонатором (VCSOA) обеспечивают лучшую эффективность связи и более низкую чувствительность к поляризации, а также их можно интегрировать в двумерные массивы [113]. Более того, на основе предложенной реализации функции обучения в [125], VCSOA могут предложить низкое энергопотребление для реализации функций обучения в нейронных сетях. Пространственный модулятор света (SLM) – это устройство, которое может использоваться для изменения амплитуды, поляризации и фазы по пространственной протяженности светового луча [126]. Интегрированная пространственная модуляция света в кремниевой фотонике может обеспечить полностью оптические реконфигурируемые устройства с возможными применениями при тестировании оптических схем и реконфигурируемых многопортовых оптических фильтров, сплиттеров и модуляторов для передачи данных [127]. SLM могут использоваться в вычислительных архитектурах резервуаров, как описано в разделе 5.3. Рассеиватель в виде колонны – это тип устройства, которое можно использовать для реализации пластовых вычислений [34]. Эти устройства могут помочь ускорить классификацию биологических клеток [128], [129]. Например, [129] представил доказательство концепции, основанной на моделировании во временной области с конечной разностью (FDTD), интегрированного фотонного приложения Extreme-Learning Machine (ELM) для быстрой и безметочной классификации биологических клеток. В этом приложении пассивный оптический столик, содержащий набор столбчатых рассеивателей, встроенных в оболочку из нитрида кремния, используется для обработки света, рассеянного вперед ячейкой при освещении через зеленый монохроматический источник. Оптический компаратор – обычное устройство в конструкции аналого-цифровых преобразователей [130]. Оптические компараторы могут быть изготовлены с использованием MR, SOA и лазеров [131]. Полностью оптические компараторы предпочтительнее оптоэлектронных, поскольку они могут обеспечить более высокую скорость и более низкое энергопотребление за счет исключения электрооптических преобразований [131]. Оптический компаратор также полезен для реализации уровней максимального объединения в CNN [132].

Наконец, кремниевые фотонные решетки с решетчатыми волноводами (AWG) обычно используются в качестве оптических (де) мультиплексоров в системах WDM. Эти устройства способны объединять множество длин волн в одно оптическое волокно, что значительно увеличивает пропускную способность оптических сетей. AWG использовались для реализации умножения матриц [133] и CNN [134].

4 МИКРОАРХИТЕКТУРЫ КРЕМНИЕВЫХ ФОТОННЫХ НЕЙРОНОВ

В этом разделе мы рассмотрим различные реализации кремниевых фотонных нейронов, которые образуют строительные блоки фотонных нейронных сетей. В подразделе 4.1 обсуждается, как различные функции в отдельном нейроне реализуются с помощью кремниевых фотонных устройств. В подразделе 4.2 описаны два подхода к классификации реализации микроархитектур фотонных нейронов.

4.1. Реализация внутринейронных функций с помощью кремниевых фотонных устройств.

Искусственные нейроны предназначены для имитации различных функций биологических нейронов и могут быть объединены для создания масштабируемой, энергоэффективной и высокопроизводительной нейронной сети. Ожидается, что высокопроизводительный фотонный нейрон обеспечит адекватную надежность [40], [85], масштабируемость [90] и каскадность [111], [135], [136]. Надежность нейрона можно повысить, либо уменьшив шум на выходе нейрона, либо увеличив мощность желаемого оптического сигнала (т. е. Сигнала, несущего данные, которыми обмениваются через межнейронную связь), чтобы гарантировать, что нейрон не возбуждается нежелательным шумом и возбуждается только желаемым сигналом. Масштабируемый нейрон поддерживает достаточное количество входов разветвления для создания крупномасштабных сетей. Действительно, одним из основных факторов, влияющих на энергоэффективность вычислений нейронной сети, по сравнению с традиционными вычислениями фон Неймана, является высокая связность, вдохновленная мозгом млекопитающих. Каскадность, которая напрямую влияет на надежность нейронной сети, является еще одним важным фактором, влияющим на производительность нейрона. Каскадность конструкции нейрона определяется на основе мощности оптического сигнала нейрона для управления другими нейронами. Вместе надежность, масштабируемость и каскадность являются важными показателями при оценке производительности фотонного нейрона. В фотонных нейронных сетях широко используются два типа нейронов: обычные (без пиков) и с пиками, как описано в разделе 2. В целом фотонный нейрон выполняет четыре основные функции: взвешивание, суммирование, активация и обучение. Когда дело доходит до обучающей функции, между двумя типами нейронов есть существенные различия. В моделях пикового нейрона функция обучения реализуется на уровне нейрон-микроархитектуры: например, обучение без учителя (например, пиковая зависимая от времени пластичность (STDP)) часто реализуется как часть

пикового нейрона, чтобы точно имитировать функциональность биологического нейрон. Для обычных нейронов функция обучения не является частью модели нейрона и вместо этого реализуется на уровне архитектуры нейронной сети (например, с обучением с обратным распространением), а не на уровне нейрон-микроархитектуры. Мы обсудим четыре основные функции нейрона в следующих подразделах.

4.1.1 Весовая функция

В биологических нейронных сетях синапсы имеют большое значение, потому что синапс – это память для процесса обучения. Синапсы обеспечивают взвешенные связи между нейронами, где изменение весов является основной функцией процесса обучения (обсуждается в подразделе 4.1.4). Из-за адаптивного взвешивания в синапсах весом синапса манипулируют (в процессе обучения), чтобы изменить эффект каждого ввода. Чтобы имитировать такое динамическое взвешивание соединений, можно использовать кремниевые фотонные устройства, такие как MR, для управления оптической передачей между двумя нейронами. В результате такие устройства могут реализовывать весовые функции, при которых передача может управляться посредством

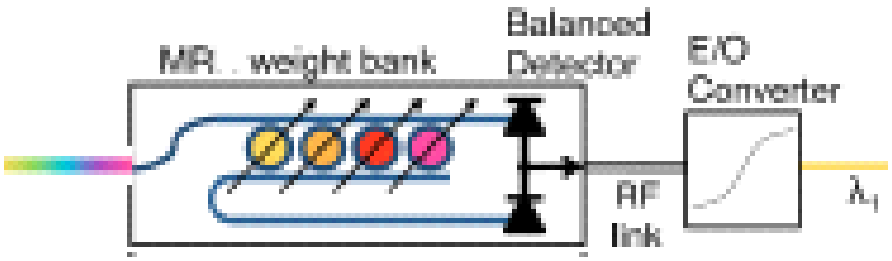


Рисунок 11: Модель нейрона, использованная в [40].

Здесь несколько длин волн используют общий волновод, чтобы обеспечить широкую полосу пропускания. Входные данные по разным длинам волн попадают в банк весов MR. Затем взвешенные входные данные суммируются с помощью сбалансированного фотодетектора, и лазер на последнем этапе преобразует электрический суммирующий сигнал в оптические пики (преобразователь E / O может быть модулятором [83] вместо лазера). MR – одно из ключевых устройств, используемых при разработке весовой функции. MR могут быть размещены в массивах, чтобы предложить банк динамических фильтров (см. Банк весов MR на рисунке 11) на входных соединениях постсинаптического нейрона [90], [110], [111]. Каждому MR в банке весов MR присвоено значение веса. Когда оптический сигнал 14 проходит MR в банке весов MR, MR может изменять мощность оптического сигнала пропорционально своему весовому значению. Затем взвешенные оптические сигналы отправляются на фотодетектор для выполнения функции суммиро-

вания, которая обсуждается в подразделе 4.1.2. В банке весов MR с парадигмой WDM можно использовать несколько длин волн для поддержки связи с высокой пропускной способностью и обеспечения большой масштабируемости [110], [111]. Однако количество длин волн, которые могут использоваться в банке весов MR, ограничено штрафом за перекрестный вес [40], где разнос каналов – частотный интервал между двумя последовательными оптическими каналами / длинами волн – должен гарантировать желаемый настроенный вес для каждого синапса.. Уменьшение разноса каналов (путем увеличения степени WDM) в банке весов MR увеличивает нежелательные эффекты (например, межканальные перекрестные помехи) на отношение пиковых шумов к шуму, что может привести к нежелательной настройке веса. Такие эффекты поперечного веса могут быть улучшены за счет увеличения мощности оптического сигнала для улучшения отношения пик / шум и, следовательно, надежности нейронов. В [40] авторы предложили аналитическую модель для разработки банка весов MR с учетом разноса каналов и энергоэффективности. Устройства, которые используют материалы фазового изменения (PCM) для настройки, также могут использоваться для реализации функции взвешивания. В [118] Ge2Sb2Te5 (GST), который представляет собой ИКМ, используется для эффективного изменения оптической передачи волновода для реализации фотонного синапса (весовой функции) в нейроне с импульсами. В этой конструкции, показанной на рисунке 12, на волноводе размещено несколько элементов / легирующих добавок из ИКМ, которые называются островком из ИКМ.

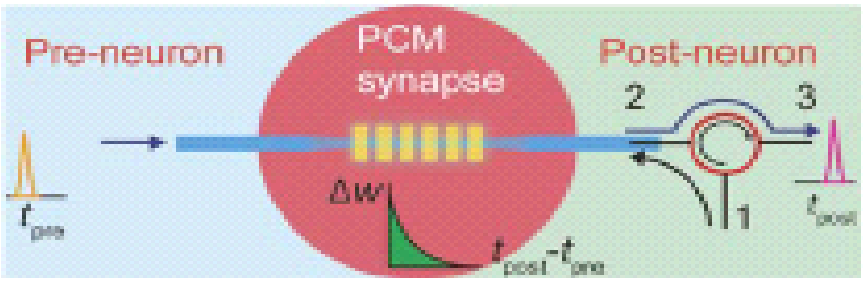


Рисунок 12: Синапс на основе PCM, предложенный в [118]. Синапс основан на размещении нескольких островков ИКМ (желтый) на сужающемся волноводе (синий) для управления оптической передачей между пресинаптическими и постсинаптическими нейронами. Импульсы взвешивания для изменения веса синапса оптически подаются через порт 1.

По сравнению с использованием одного островка ИКМ, эта конструкция улучшена за счет использования нескольких островков ИКМ на волноводе для каждого синапса, что помогает реализовать более эффективное изменение оптической передачи в волноводе. Результаты также показывают, что использование конструкции сужающегося волновода в сочетании с островками ИКМ более эффективно, чем использо-

вание стандартного не сужающегося волновода. Эксперименты в [118] подтверждают, что каждый вес синапса на основе ИКМ может быть получен с использованием заранее определенного количества входных оптических импульсов. Таким образом, точная и полностью оптическая настройка веса может быть достигнута за счет использования нескольких островков ИКМ с коническим волноводом. Однако такой полностью оптический синапс страдает низкой скоростью работы из-за процесса фото-структурного преобразования, которое влияет на движение атомов или ионов, поскольку скорость атомов и ионов намного ниже, чем скорость фотона [137]. Возбуждающие и тормозящие функции: помимо веса синапсов, тип нейромедиаторов играет важную роль в биологическом нейроне. Нейротрансмиттер – это химическое вещество, передающее информацию через синапс рецепторам постсинаптического нейрона [138]. В зависимости от типа нейротрансмиттеров взвешенные входные сигналы могут увеличивать или уменьшать мембранный потенциал постсинаптического нейрона. Если взвешенный вход увеличивает мембранный потенциал, взвешенный вход является возбуждающим (т. е. Побуждает нейрон возбуждаться). С другой стороны, тормозящий взвешенный вход снижает мембранный потенциал (т. е. Препятствует возбуждению нейрона). В искусственном нейроне это соответствует положительному или отрицательному весу для каждого синапса. Следовательно, взвешенный вход может увеличивать или уменьшать мембранный потенциал постсинаптического нейрона для поддержки возбуждающих и тормозных функций соответственно. Несколько кремниевых фотонных устройств были исследованы для реализации возбуждающих и тормозных функций, которые необходимы в конструкции нейронных сетей. Изучена возбуждающая функция лазеров с распределенной обратной связью (РОС) [93], MR [112] и VCSEL [97]. Более того, [133] проанализировали ингибирующую функцию VSCEL. Чтобы создать эффективную фотонную нейронную сеть, в конструкции нейрона требуются как возбуждающие, так и тормозящие функции. Например, [91] предложила модель нейрона, в которой VCSEL используется для реализации как возбуждающих, так и тормозных функций нейрона путем одновременного введения ортогонально поляризованных и параллельно поляризованных полей. В дополнение к возбуждающим и тормозным функциям, введение двух полей делает нейрон более надежным в присутствии шума и обеспечивает более быстрый ответ VCSEL. Предложенная в [90] конструкция нейрона, представляющая собой оптоэлектронный нейрон для поддержки WDM, также обеспечивает как возбуждающие, так и тормозные функции. Нейрон использует два фильтра на основе MR для представления положительных и отрицательных весов возбуждающей и тормозной функций соответственно. Кроме того, [140]

предлагает экспериментальную реализацию и анализ суммирования с возбуждающими и тормозными функциями в оптоэлектронном нейроне, предложенном в [90]. Однако в этом оптоэлектронном нейроне необходимы две длины волны, чтобы задействовать возбуждающую и тормозную функции при суммировании. Предложенный нейрон в [141] использует метод модуляции, который основан на использовании двух двухтактных ИМЦ и фазовращателя, чтобы реализовать как положительные, так и отрицательные веса на одной длине волны. Однако в описанных выше возбуждающих и тормозных функциях на основе лазера используются активные устройства. Фотонные нейроны, использующие активные устройства, в которых свет генерируется самим устройством, страдают от проблем интеграции, поскольку изготовление их с использованием стандартного процесса CMOS является дорогостоящим. С этой целью нейрон, предложенный в [113], обеспечивает как тормозные, так и возбуждающие функции с помощью MR, которые являются пассивными устройствами.

4.1.2 Функция суммирования

Как описано в разделе 2, в биологическом нейроне сома или тело нейрона отвечает за объединение (то есть суммирование) входных сигналов нейрона, так что постсинаптический нейрон может быть возбужден совокупным входным сигналом. шипы. Точно так же в традиционной модели искусственного нейрона (без пиков) функция суммирования объединяет все входные данные нейрона и направляет результат в функцию активации. Суммирование по входам очень важно, потому что оно напрямую влияет на масштабируемость нейрона. Масштабируемый нейрон может эффективно интегрировать большое количество входных данных, чтобы обеспечить большой разветвление и, следовательно, крупномасштабную фотонную нейронную сеть. Чтобы разработать такой масштабируемый нейрон, нейроны, предложенные в [90] и [111], комбинируют входные данные на разных длинах волн, чтобы обеспечить совместимость с WDM. В таких реализациях нейрон использует фотодетектор для объединения входных сигналов, передаваемых на нескольких длинах волн (см. Фотодетектор на рисунке 11). В результате существует потребность в преобразовании сигнала (например, из оптического в электрический и из электрического в оптический) в таких оптоэлектронных нейронах, где такое преобразование приводит к некоторым потерям мощности и, следовательно, ухудшению характеристик нейрона. В качестве альтернативы, функция суммирования может быть реализована в фотонном нейроне с использованием полностью оптического подхода. Например, в [142] исследуется функция суммирования в лазерах на микростолбах-полупроводниках на основе интегрального насыщающегося поглотителя. Результаты показывают, что микропил-

лярный лазер способен комбинировать импульсные стимулы и вызывать активационную функцию. Более того, DFB-лазеры [93] и VCSEL [91], [97] также используются для реализации функции суммирования в фотонных нейронах. Предложенный нейрон в [112] использует MR для реализации функции суммирования. Тем не менее, полностью оптические подходы не поддерживают WDM, следовательно, не могут обеспечить высокую взаимосвязь для реализации масштабируемых фотонных нейронных сетей.

4.1.3 Функции активации

Функция активации нейрона может быть линейной или нелинейной. В [141] представлена линейная модель нейрона для поддержки линейных сложений и вычитаний как с положительными, так и с отрицательными весами для реализации возбуждающих и тормозных функций. Предложенный линейный нейрон в [141] может поддерживать нелинейные сигмовидные и функции активации ReLU, добавляемые к базовому линейному нейрону. В [113], чтобы реализовать нелинейную функцию активации, эффекты нелинейности в MR используются для реализации нейрона малой мощности. Существуют различия в функциях активации в моделях спайковых и обычных нейронов. В пиках нейронов функция активации определяет время пика нейрона на основе агрегированных пиков входного сигнала. Следовательно, время распространения оптического сигнала зависит от функции активации в этом подходе, управляемом событиями.

Некоторые недавние попытки были направлены на устранение этих ограничений. Авторы в [37] предложили новую архитектуру, которая использует параллельное расположение MR вместо их каскадирования, хотя они используют неэффективные MZI для хранения векторов. Параллельное «безударное» расположение MR использовалось для уменьшения тепловых перекрестных помех между соседними MR и для достижения лучшего разрешения по весу. В других работах было предложено повторно использовать реализованные слои, например [156], а также методы для снижения энергопотребления и увеличения скорости работы за счет уменьшения использования электронных компонентов [160]. Также были предложения использовать микродиски вместо MR для дальнейшего увеличения плотности интеграции на чипах [161]. Но, как отмечено в [202], все варианты некогерентных архитектур могут страдать от низкой пропускной способности, поскольку электронные компоненты, такие как память, могут не работать с такой высокой частотой, как их фотонные компоненты. Для преодоления этих ограничений некогерентных архитектур необходимы дальнейшие исследования.

В качестве альтернативы, в обычных нейронах, используемых в ИНС, оптический сигнал распространяется от входа к выходу за заранее

определенное время. В [143] SOA используются для эмуляции сигмовидной функции активации обычных нейронов. Функция активации также может быть реализована с помощью фотонного лазера с электрическими управляющими сигналами в оптоэлектронных нейронах (см. Рисунок 11). В нейроне, использованном в [90], который называется Broadcast-and-Weight (B&W), возбуждаемый лазер и фотодетектор используются для имитации функции возбуждения искусственных нейронов. В частности, функция активации реализуется с использованием возбуждаемого лазера и может срабатывать, когда сигнал суммирования, выдаваемый фотодетектором, достигает порогового значения. Решение о подаче импульса, реализуемое функцией активации, вызывает в лазере импульсный оптический сигнал. Подход B&W также использовался в обычных нейронах. Например, в [144] используется нейрон на основе MZI, в котором функция активации является бинарной функцией +1 или -1 (функция символического решения). VSCEL также показали большой потенциал для реализации функции активации из-за их относительной небольшая занимаемая площадь [145], низкое энергопотребление [135], возможность 2D или 3D интеграции в массивы [99], низкие производственные затраты [135], [145] и эффективность подключения к оптическим волокнам [135], [147]]. В [97] исследуется импульсное поведение нейронов на основе VCSEL, которое управляется электрически. В [146] исследуется пиковое поведение VCSEL как с параллельными, так и с ортогонально поляризованными оптическими стимулами, и результаты показывают, что VCSEL способны создавать управляемые пики, необходимые для создания сверхбыстрых оптических нейронных сетей. В [135] исследуется поведение возбуждения VCSEL. Для исследования каскадируемости VCSEL с учетом их поведения возбуждения в [135] рассматриваются два VCSEL – передатчик VSCEL и приемник VCSEL. Результаты показывают, что контролируемые выбросы (с использованием внешнего управляющего сигнала) распространяются от первого VCSEL до 16 второго, подтверждая каскадируемость VCSEL и что они могут использоваться в качестве устройства возбуждения в фотонных нейронных сетях. Однако после запуска спайка с использованием VCSEL могут возникнуть собственные релаксационные колебания, которые ухудшают надежность и скорость фотонных нейронов [91], [148]. Большинство фотонных нейронов на основе VCSEL [99], [139], [146], [149] не поддерживают одновременно возбуждающую и тормозную активацию. Однако в [91] и [148] как возбуждающая, так и тормозная функции реализуются в VCSEL за счет использования инъекций с двойной поляризацией, т. е. Инъекций с ортогональной и параллельной поляризацией. Кремниевые фотонные устройства часто проектируются для работы с одной поляризацией, поэтому использование двухполяри-

зованных лазеров VCSEL является проблемой. PCM могут также использоваться при разработке фотонных устройств для реализации функций активации. Полностью оптический нейрон, предложенный в [120], показанный на рисунке 13, использует ИКМ в структуре MP для реализации поведения возбуждения нейронов с импульсами. На рис. 13 (a) показана схема предлагаемого нейрона, на рис. 13 (б) показаны его основные компоненты, а на рис. 13 (c) показана схема предлагаемого нейрона. Как показано на рисунках, PCM не только используется для реализации функции взвешивания, но также помещается на MR для имитации функции активации. Ответ передачи PCM на MR (показанный на рисунке 13 (b) -IV) использовался для эмуляции функции активации ReLU. (a) (b) (c)

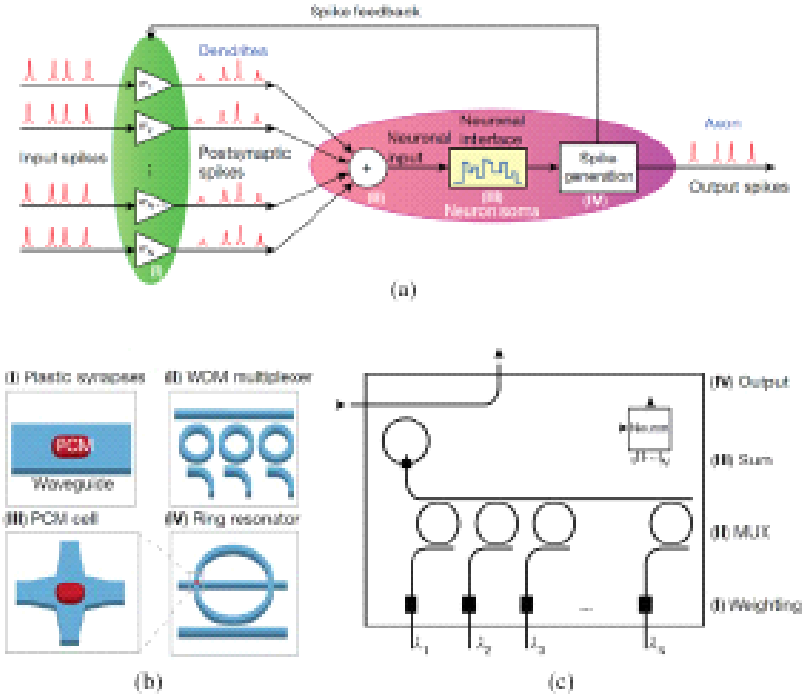


Рисунок 13: нейрон на основе PCM, предложенный в [120]: (a) схема модели нейрона, (b) основные компоненты нейрона, и (c) фотонная цепь нейрона.

4.1.4 Функция обучения STDP

В нейронах с импульсными сигналами обучение пластичности, зависящей от времени (STDP), обычно используется для точного имитации биологического нейрона, в то время как в обычных искусственных нейронах функция обучения реализуется на уровне архитектуры нейронной сети и обычно в электронный домен. В нейронах с пиками функция обучения STDP обновляет веса на основе пресинаптических и постсинаптических пиков, чтобы помочь постепенно уменьшить ошибку нейронной сети, которая соответствует разнице между желаемым и фактическим выходом. В процессе обучения STDP сила связей (т. Е. Синапсов) регулируется на основе времени всплеска пресинаптических и постсинаптических нейронов. Изменения в весовом синапсе зависят от

времени выброса на выходе и на входе. В зависимости от времени импульса пресинаптического нейрона и постсинаптического нейрона вес может быть увеличен или уменьшен для реализации функции обучения фотонного нейрона. Вес синапса увеличивается, что называется «потенцированием», если пресинаптический спайк возникает прямо перед постсинаптическим спайком. С другой стороны, вес синапса уменьшается, когда пресинаптический спайк пропускает возбуждение постсинаптического нейрона, то есть постсинаптический нейрон срабатывает в результате спайков, полученных от других пресинаптических нейронов.

В [122] STDP реализован с использованием SOA и модулятора электропоглощения (EAM), который может быть развернут в высокоскоростных (пикосекундных временных масштабах) вычислениях нейронной сети. Кроме того, [123] обсуждает фотонную реализацию STDP и его применение как в обучении с учителем, так и без него. Использование одного устройства SOA для реализации обучения STDP улучшает масштабируемость нейронов для реализации контролируемых и 17 неконтролируемых алгоритмов обучения в крупномасштабных нейронных сетях. Однако использование SOA для реализации STDP требует большого энергопотребления (например, по сравнению с использованием пассивных устройств). В [150] алгоритм обучения STDP реализован на основе использования пассивных MR, что подходит для проектирования крупномасштабных фотонных нейронных сетей с низким энергопотреблением. В [125] VCSOA используется для улучшения энергопотребления в реализациях обучения STDP на основе SOA. 4.1.5 Резюме Современные нейронные функции и их реализация с использованием кремниевых фотонных устройств, обсуждаемых в этом разделе, сведены в Таблицу 1. 4.2. Классификация реализаций кремниево-фотонной нейронной микроархитектуры Реализации нейронов с кремниевой фотоникой можно классифицировать двумя способами: 1) полностью оптические нейроны по сравнению с оптоэлектронными и 2) когерентные и некогерентные нейроны. В следующих подразделах мы подробно обсудим эти два подхода к классификации реализации нейронов.

4.2.1. Полностью оптические нейроны в сравнении с оптоэлектронными

В конструкции оптоэлектронных нейронов существует потребность в преобразованиях из электрического в оптический и из оптического в электрический. В таких нейронах взвешенные входные данные обычно суммируются / комбинируются с использованием фотодетектора для управления лазером [90]. Следовательно, оптические входы должны быть преобразованы в электрические сигналы, а электрический выход фотодетектора затем должен быть преобразован в оптический сигнал с помощью лазера (см. Рисунок 11). Из-за таких преобразований из опти-

ческих в электрические и из электрических в оптические области нейрон также называют нейроном O / E / O. По сравнению с полностью оптическими нейронами, нейроны O / E / O неэффективны по мощности из-за потерь мощности

Таблица 1: Сводка нейронных функций и их реализации с использованием кремниевых фотонных устройств

		Implementation	Advantages	Disadvantages	References
Weighting function		MR filters	Compatible with WDM	Not all-optical (require signal conversions)	[40], [110]
		Several PCM islands on a tapered waveguide	All optical weight update	Low operation speed	[118]
Summation function		Photodiode	Compatible with WDM	Low scalability	[90]
		Micropillar laser	Both summation and excitation in laser	Does not support WDM and inhibitory function	[142]
Activation function	Support excitatory and inhibitory functions	VCSEL by double polarized injection	Support both excitatory and inhibitory functions; VCSEL advantages (low power, small footprint, implementation in arrays for large-scale designs)	Require double polarized injection	[91], [148]
		MR	Use passive devices compatible with standard CMOS technology; support both excitatory and inhibitory functions	Sensitive to temperature and fabrication-process variations; has a limited cascadability	[112], [113]
		VCSEL	Low power, small footprint, implementation in arrays for large-scale designs	Uses active laser devices that consume high power compared to passive ones [112], [113]	[97], [99], [146], [135], [147]
		PCM and MR	Power efficient due to the use passive devices	low cascadability	[120]
STDP Learning function		Semiconductor optical amplifier (SOA)	Scalable	Power inefficient	[122], [123]
		Vertical-cavity semiconductor optical amplifier (VCSOA)	Power efficient in comparison with SOA-based STDP learning	Employing active devices, which are power hungry	[125]
		MR	Power efficient because of implementation with passive device – does not require neurons to spike at different wavelengths	Only support unsupervised STDP learning	[150]

Преимущества реализации Недостатки Ссылки Весовая функция MR-фильтры Совместимы с WDM Не все- оптический (требуется преобразование сигнала) [40], [110] Несколько островков ИКМ на сужающемся волноводе Обновление всех оптических весов Низкая скорость работы [118] Функция суммирования Фотодиод Совместимость с WDM Низкая масштабируемость [90] Микропиллярный лазер Суммирование и возбуждение в лазере Имеет не поддерживает WDM и ингибирующую функцию [142] Функция активации S upportexcitatoryandin hibitory functions VCSEL посредством двойной поляризованной инъекции Поддержка как возбуждающих, так и тормозных функций; Преимущества VCSEL (низкое энергопотребление, малая занимаемая площадь, реализация в массивах для крупномасштабных проектов) Требуется инжекция с двойной поляризацией [91], [148] MR. Использование пассивных устройств, совместимых со стандартной технологией CMOS; поддержи-

вают как возбуждающую, так и тормозную функции. Чувствительность к изменениям температуры и технологического процесса; имеет ограниченную каскадность [112], [113] VCSEL Низкое энергопотребление, небольшие размеры, реализация в массивах для крупномасштабных проектов. Использует активные лазерные устройства, которые потребляют большую мощность по сравнению с пассивными [112], [113] [97], [99], [146], [135], [147] PCM и MR Энергопотребление за счет использования пассивных устройств с низкой каскадируемостью [120]

Функция обучения STDP Полупроводниковый оптический усилитель (SOA) Масштабируемая мощность неэффективен [122], [123] Вертикальный -резонаторный полупроводниковый оптический усилитель (VCSOA) Эффективность энергопотребления по сравнению с обучением STDP на основе SOA Использование активных устройств, которые потребляют много энергии [125] MR Эффективная мощность из-за реализации с пассивным устройством – не требует, чтобы нейроны вспыхивали на разных длинах волн Только поддержка неконтролируемое обучение STDP [150]

4.2. Классификация реализаций кремниево-фотонной нейронной микроархитектуры

Реализации нейронов с кремниевой фотоникой можно классифицировать двумя способами: 1) полностью оптические нейроны по сравнению с оптоэлектронными и 2) когерентные и некогерентные нейроны. В следующих подразделах мы подробно обсудим эти два подхода к классификации реализации нейронов.

4.2.1. Полностью оптические нейроны в сравнении с оптоэлектронными

В конструкции оптоэлектронных нейронов существует потребность в преобразованиях из электрического в оптический и из оптического в электрический. В таких нейронах взвешенные входные данные обычно суммируются / комбинируются с использованием фотодетектора для управления лазером [90]. Следовательно, оптические входы должны быть преобразованы в электрические сигналы, а электрический выход фотодетектора затем должен быть преобразован в оптический сигнал с помощью лазера (см. Рисунок 11). Из-за таких преобразований из оптических в электрические и из электрических в оптические области нейрон также называют нейроном O / E / O. По сравнению с полностью оптическими нейронами, нейроны O / E / O неэффективны по мощности из-за потерь мощности Таблица 1: Сводка нейронных функций и их реализации с использованием кремниевых фотонных устройств Преимущества реализации Недостатки Ссылки Весовая функция MR-фильтры Совместимы с WDM Не все- оптический (требуется преобразование

сигнала) [40], [110] Несколько островков ИКМ на сужающемся волноводе Обновление всех оптических весов Низкая скорость работы [118] Функция суммирования Фотодиод Совместимость с WDM Низкая масштабируемость [90] Микропиллярный лазер Суммирование и возбуждение в лазере Имеет не поддерживает WDM и ингибирующую функцию [142] Функция активации S upportexcitatoryandin hibitory functions VCSEL посредством инъекции с двойной поляризацией Поддерживает как возбуждающую, так и тормозную функции; Преимущества VCSEL (низкое энергопотребление, малая занимаемая площадь, реализация в массивах для крупномасштабных проектов) Требуется инжекция с двойной поляризацией [91], [148] MR Используйте пассивные устройства, совместимые со стандартной технологией CMOS; поддерживают как возбуждающую, так и тормозную функции. Чувствительность к изменениям температуры и технологического процесса; имеет ограниченную каскадность [112], [113] VCSEL Низкое энергопотребление, небольшие размеры, реализация в массивах для крупномасштабных проектов. Использует активные лазерные устройства, которые потребляют большую мощность по сравнению с пассивными [112], [113] [97], [99], [146], [135], [147] PCM и MR Энергопотребление за счет использования пассивных устройств с низкой каскадируемостью [120] Функция обучения STDP Полупроводниковый оптический усилитель (SOA) Масштабируемая мощность неэффективен [122], [123] Вертикальный - резонаторный полупроводниковый оптический усилитель (VCSOA) Эффективность энергопотребления по сравнению с обучением STDP на основе SOA Использование активных устройств, которые потребляют много энергии [125] MR Эффективная мощность из-за реализации с пассивным устройством – не требует, чтобы нейроны вспыхивали на разных длинах волн. Только поддержка неконтролируемое обучение STDP [150] 18 принудительно в необходимых преобразованиях. Более того, из-за аналоговой природы внутринейронной связи [85] как в электрической, так и в оптической областях фотодетектор и лазер-модулятор чувствительны к шуму. Анализ шума для оптоэлектронных нейронов представлен в [85]. Чтобы компенсировать шум, мощность модулятора или усиление электрического трансимпеданса следует увеличить путем добавления трансимпедансного усилителя (TIA) [85]. Однако увеличение мощности модулятора и добавление TIA приводят к накладным расходам энергии. Обычные нейроны O / E / O [40], [85], [111] часто используют непосредственно модулированный лазер для возбуждения и пиков, что, в свою очередь, требует размещения фотонных устройств и лазера на одном чипе. Следовательно, такие нейроны страдают от тепловых проблем и изменений, вызванных встроенным лазером. С этой целью Modulator Neuron [83], [85] использует модулятор вместо лазера с

прямой модуляцией. Следовательно, нейроны могут использовать внешний лазер в качестве источника света для компенсации тепловых проблем. В полностью оптических нейронах нет необходимости в электро-оптическом преобразовании во время внутринейронной коммуникации, то есть все устройства в искусственном нейроне поддерживают передачу оптических сигналов [113], [120]. Например, [120] предложил полностью оптический нейрон с импульсами, показанный на рисунке 13, включая реализацию обучения STDP. Более того, [113] предложил полностью оптический нейрон на основе пассивных устройств (в которых свет не генерируется самим устройством). Модели, представленные в [113], предполагают, что MR обеспечивают быстрые и энергоэффективные возбуждающие и тормозные функции в фотонных нейронах. Помимо тормозных и возбуждающих функций, предлагаемая модель демонстрирует рефрактерное поведение, которое является важной функцией в реализации нейронной сети. Более того, поскольку в предлагаемом нейроне используются пассивные MP-устройства, его можно легко реализовать с помощью стандартной технологии КМОП. Однако таким полностью оптическим нейронам не хватает высокой каскадности для поддержки большой нейронной сети. Кроме того, как мы обсуждали в разделе 4.1.1, все оптические синапсы (как часть полностью оптических нейронов) страдают от низкоскоростной операции для реализации весовых функций.

4.2.2. Когерентные и некогерентные нейроны.

В зависимости от длины волны работы нейронов реализации нейронов можно разделить на когерентные и некогерентные [151]. Когерентные нейроны манипулируют фазой и амплитудой электрического поля с помощью одной длины волны. Некогерентные нейроны, например те, которые используют конфигурацию черно-белых фотонных нейронов, о которой говорилось ранее, манипулируют мощностью оптического сигнала и полагаются на несколько длин волн. Когерентные нейроны, предложенные в [32] и [141], используют MZI и являются энергоэффективными, поскольку им требуется одна длина волны. Однако MZI накладывают большие накладные расходы на область, и поэтому дизайн не может быть расширен для поддержки крупномасштабных нейронных сетей. Более того, MZI в когерентных нейронах требуют фазовращателей, в которых ошибка настройки неизбежна. Эта ошибка настройки может распространяться и увеличиваться в нейронной сети, снижая надежность сети. Использование лазеров на микродисках было исследовано в [113] для создания когерентного нейрона, в котором возбуждающая и тормозная функции реализуются за счет управления оптическими фазами. Однако регулировка оптических фаз, которую можно выполнить с помощью микронагревателя [113], добавляет новую про-

блему. Помимо задач фазового контроля, когерентные нейроны работают на одной длине волны и не могут различать разные длины волн. Следовательно, нейронная сеть, основанная на когерентных нейронах, не поддерживает реконфигурируемость [84] и WDM, что приводит к низкой пропускной способности. С другой стороны, некогерентные нейроны могут работать с несколькими длинами волн и поддерживать WDM, в котором несколько длин волн используют общий волновод, чтобы обеспечить высокую степень связи с меньшим количеством волноводов. Однако зависимость между длинами волн на входе и выходе в фотонных ИНС, использующих некогерентные нейроны, требует преобразования длины волны [84]. Такие преобразования могут потребовать накладных расходов с высоким энергопотреблением [152]. Более того, некогерентные нейроны также страдают от межканальных перекрестных помех, которые могут снизить надежность [40].

4.2.3 Сводная таблица

В таблице 2 обобщены современные подходы к реализации полностью оптической, оптоэлектронной, когерентной и некогерентной микроархитектуры нейронов.

Таблица 2: Сводка некоторых предлагаемых микроархитектур нейронов с использованием кремниевых фотонных устройств

		Implementation	Advantages	Disadvantages	References
O/E/O neuron	Noncoherent	Use weight bank for synapse, photodetector for summation, and laser for spiking	Use WDM to offer high bandwidth	Not all optical – consumes power in electro-optical and opto-electrical conversions; requires wavelength conversions	[90]
		Use modulator instead of direct modulation by laser	Use WDM to offer high bandwidth; support on-chip neurons with off-chip lasers	Not all optical; high power consumption because of employing couplers	[83], [85]
All-optical neuron	Noncoherent	Use MR weight banks for synapse and PCM for excitation function	Supports both unsupervised and supervised learning; no need for electro-optical or opto-electrical conversion	Cross-weight penalty [40] for weight tuning; low cascadability	[120]
	Coherent	MR (require an all-optical synapse: synapse [118] is suggested by the paper)	Use passive devices; no need for electro-optical or opto-electrical conversion	Optical learning is not included; low cascadability	[113]
		Splitters and MZIs	Higher reliability and low power overhead in comparison with using several wavelengths	Low bandwidth because of using one wavelength; area inefficient; exact splitting ratios are hard to achieve after fabrication due to variations; susceptible to noise in phase and splitting ratios [153]	[141]
		MZIs	Higher reliability and low power overhead in comparison with using several wavelengths	Low bandwidth because of using one wavelength; area inefficient and hence not scalable; susceptible to phase noise	[32]

Преимущества реализации Недостатки Ссылки O / E / O нейрон N онкогерентный. Используйте банк весов для синапса, фотодетектор для суммирования и лазер для пиковых значений Используйте WDM для обеспечения высокой пропускной способности Не все оптические – по-

требляет энергию в электрооптических и оптоэлектрических преобразователях; требует преобразования длины волны [90] Используйте модулятор вместо прямой модуляции лазером. Используйте WDM для обеспечения высокой пропускной способности; поддержка нейронов на кристалле с помощью внешних лазеров. Не все оптические; высокое энергопотребление из-за использования ответвителей [83], [85] A l l – o p t i c a l n e u r o n n o n c o h e r e n t Использование весовых банков MR для синапсов и PCM для функции возбуждения Поддерживает как неконтролируемое, так и контролируемое обучение; отсутствие необходимости в электрооптическом или оптоэлектрическом преобразовании Штраф за поперечный вес [40] для настройки веса; низкая каскадируемость [120] MPT (требуется полностью оптический синапс: синапс [118] предлагается в статье) Используйте пассивные устройства; нет необходимости в электрооптическом или оптоэлектрическом преобразовании. Оптическое обучение не включено; низкая каскадность [113] Сплиттеры и MZI Более высокая надежность и низкие накладные расходы по мощности по сравнению с использованием нескольких длин волн Малая полоса пропускания из-за использования одной длины волны; площадь неэффективная; Точные коэффициенты разделения трудно достичь после изготовления из-за изменений; чувствительность к шуму по фазе и коэффициентам разделения [153] [141] MZI Более высокая надежность и низкие накладные расходы по сравнению с использованием нескольких длин волн Малая полоса пропускания из-за использования одной длины волны; область неэффективна и, следовательно, не масштабируется; восприимчив к фазовому шуму [32]

5 КРЕМНИЕВАЯ ФОТОННАЯ НЕЙРОННАЯ СЕТЬ АРХИТЕКТУРЫ

На уровне архитектуры предыдущие работы были сосредоточены на реализации различных типов моделей нейронных сетей (обсуждаемых в разделе 2): ИНС (MLP, CNN, RNN), SNN и RC. Всеобъемлющее новшество конкретной работы определяется базовыми оптическими устройствами, используемыми в архитектуре, и фундаментальными принципами, такими как оптический резонанс и оптическая интерференция, которые управляют этими устройствами. Эти принципы в конечном итоге оказывают наибольшее влияние на проектные решения, ориентированные на производительность, мощность и надежность, а также на неотъемлемые ограничения архитектуры, построенной с их использованием. Таким образом, устройства и лежащие в их основе принципы стимулируют инновации, необходимые для реализации архитектур с использованием технологии кремниевой фотоники. Следова-

тельно, наша классификация в этом разделе будет основана на основных фотонных принципах, используемых для построения архитектур нейронных сетей.

5.1 Сетевые архитектуры на основе оптического резонанса

Реализации нейронных сетей на основе оптического резонанса обычно зависят от длины волны MR или микродисков, что приводит к использованию реализаций на основе WDM, где в волноводе используются несколько длин волн. Эти архитектуры являются некогерентными архитектурами и используют некогерентные нейронные микроархитектуры, обсуждаемые в разделе 4.2. Большинство этих архитектур используют или основываются на протоколе B&W, показанном на рисунке 14, для установки и обновления весов, как было продемонстрировано в [154], чтобы иметь изоморфизм к непрерывным рекуррентным нейронным сетям во времени (CTRNN). Петли обратной связи, характерные для RNN, могут эмулироваться MR, когда они достигают оптической бистабильности. При благоприятных условиях, касающихся резонансного материала и интенсивности падающего пропускания, выходное пропускание резонатора может входить в цикл гистерезиса с двумя стабильными уровнями пропускания. Это называется оптической бистабильностью резонаторов. В работе [154] также предлагается использовать модуляторы Маха – Цендера (MZM) для генерации сигмоидной функции активации. Эта конкретная работа предоставила доказательство того, что архитектуры MR на основе черно-белой печати могут использоваться для реализации нейронных сетей и что они могут обеспечить лучшую производительность по сравнению с традиционными CTRNN на основе ЦП. Для сравнительного анализа они рассмотрели приложение для моделирования Lorenz Attractor [155] и сообщили о 294-кратном ускорении с их фотонной архитектурой по сравнению с моделированием на базе ЦП. Протокол B&W – это многоволновой аналоговый сетевой протокол, в котором несколько выходов полностью фотонных нейронов мультиплексируются и распределяются по входам всех нейронов. Эти архитектуры имеют тенденцию использовать параллелизм, присущий фотонным архитектурам, с использованием нескольких длин волн для параллельной передачи данных с использованием WDM. Различные длины волн в волноводе представляют входные сигналы для нейрона. Веса реконфигурируются путем настройки MR, так что характеристики конкретной длины волны изменяются. Банки весов MR состоят из настраиваемых MR, которые могут быть настроены для отвода энергии от их резонансной длины волны, так что интенсивность длин волн отражает веса или значения ядра. Изменение интенсивности счи-

тывается с помощью фотодетекторов (PD) и суммируется для получения выходных значений из банка весов. Этот процесс подробно описан в подразделах 4.1.1 и 4.1.2. Очевидным преимуществом этого подхода является использование хорошо изученной и зрелой технологии MR для реализации фотонных нейронных сетей, что упрощает аппаратную реализацию и интеграцию. Однако проблема, с которой может столкнуться этот протокол, – это количество MR, необходимое для его реализации в реальных приложениях, и в зависимости от карты функций и размера ядра для CNN это может стать непомерным. Исследование с использованием этого протокола пытается обойти эту проблему.

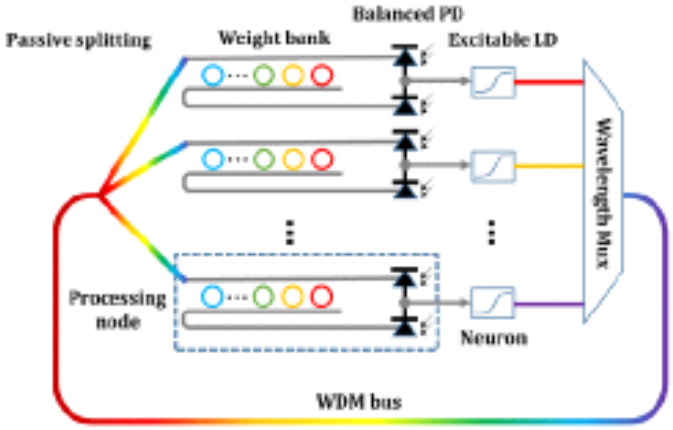


Рисунок 14: Протокол широковещательной передачи (B&W), как показано в [37].

Авторы в [156] используют протокол B&W с MZM для реализации сигмоидной нелинейности, как в [154], чтобы предложить ускоритель CNN, получивший название Photonic CNN Accelerator (PCNNA). PCNNA реализует один уровень CNN и последовательно повторно использует этот уровень с различными размерами ядра для реализации всей CNN. Карта входных функций и необходимые ядра загружаются из внешней памяти DRAM. Результаты выполнения отдельных слоев возвращаются в память. Это последовательное выполнение ядер с использованием оптического ядра, которое работает на более высокой тактовой частоте, чем его электрические компоненты. Упомянутое здесь оптическое ядро состоит из взвешивающих банков MR и цифро-аналоговых преобразователей (ЦАП), которые подают данные в банки MR и лазерные диоды (LD). Авторы утверждают, что, поскольку CNN используют ядра с одинаковыми размерами для каждого слоя, они разделяют одни и те же рецептивные поля, и, следовательно, вычисления свертки для разных ядер могут выполняться параллельно. Они продемонстрировали эффективность этой работы, реализовав AlexNet, которая представляет собой глубокую архитектуру CNN с восемью слоями, пятью слоями свертки и тремя полностью связанными слоями. Авторы показали, что

их подход на основе фильтров к реализации AlexNet имеет значительно меньшее количество MR, чем подход, не учитывающий никаких оптимизаций для реализации AlexNet (они заявили о сокращении количества MR с одного миллиарда до 100000 диапазонов). Другая архитектура, которая использует банки фотонных весов для реализации CNN, описана в [157]. Авторы описали архитектуру, которая реализует все слои CNN с использованием связанных блоков свертки, которые состоят из весовых банков, где настроенные MR принимают значения ядра, используя настройку фазы для управления энергией на своих резонансных длинах волн. Архитектура была протестирована с использованием набора данных MNIST [158] и продемонстрировала лучшее время выполнения, чем классификация на основе графических процессоров, с графическими процессорами AMD Vega FE, AMD M125, NVIDIA Tesla P100 и NVIDIA GTX 1080 Ti. Однако они не рассматривают какие-либо методы оптимизации модели для уменьшения количества MR, необходимого для ее представления, и сообщают об очень высоком потреблении мощности 100 Вт для массива модулятора 1024 MR в предлагаемой ими архитектуре. Ускоритель CNN, реализованный с использованием MR и мемристоров, описан в [132]. В этой работе веса загружаются в банк MR-весов через мемристоры, которые, в свою очередь, получают свои значения веса из внешней памяти через буферы SRAM. Архитектура включает отдельные уровни, необходимые для реализации CNN. Сверточный слой состоит из банка фотонных весов на основе мемристора. Уровень активации (уровень ReLU) построен с использованием SOA. В работе также используется полностью оптический аналоговый компаратор, предложенный в [130], для реализации уровня maxpooling. 21 Эти три слоя образуют единый слой выделения признаков. Два слоя выделения признаков соединены между собой с помощью интерфейсного слоя, который демодулирует выходные данные предыдущего слоя maxpool, генерирует соответствующие значения электронного напряжения и затем передает их в мемристоры следующего слоя выделения признаков. Эта работа была сосредоточена на распознавании рукописных цифр с использованием набора данных MNIST и показала лучшее время выполнения по сравнению с ускорителем Caffeine на базе FPGA [159] и ускорителем ISAAC на основе мемристора [20] в различных тестах. В дальнейшем архитектура была расширена авторами в [208].

Архитектура направлена на предотвращение искажения веса на основе тепловых перекрестных помех, которое может происходить в архитектурах на основе протоколов B&W. Вариант черно-белого протокола для реализаций MLP был исследован в [37], где авторы описали архитектуру «безбарьерного веса и агрегирования». Этот метод накопления значений веса из банков весов был разработан для преодоления возмож-

ного искажения значений веса из-за тепловых изменений и перекрестных тепловых помех. Предлагаемая архитектура «безбарьерного веса и агрегата» для банков MR-веса разделяет длины волн и взвешивает их параллельным образом вместо каскадного подхода, предложенного в [154].

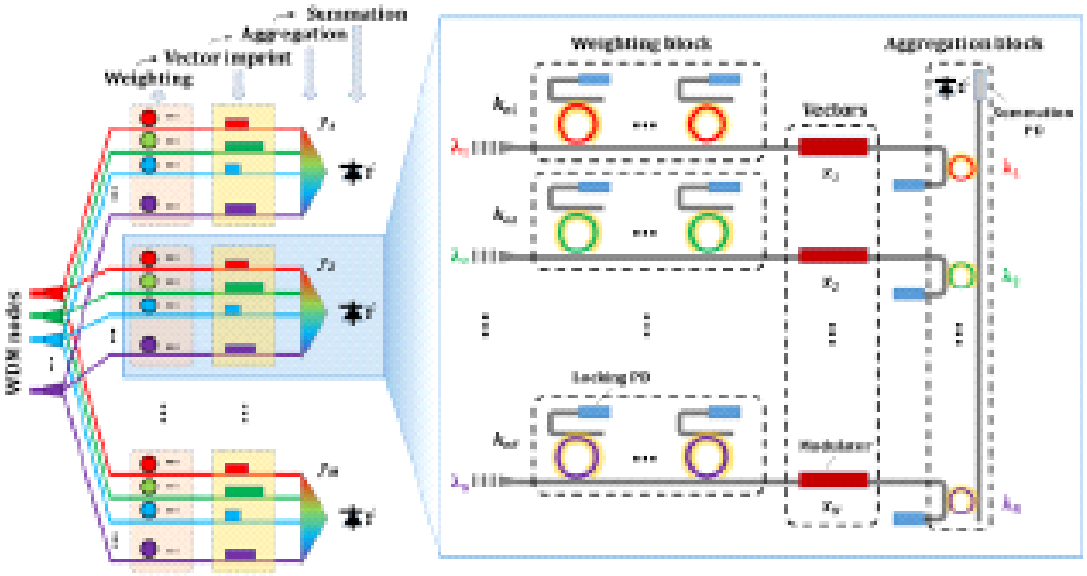


Рисунок 15: Весовая и агрегированная архитектура для оптических матрично-векторных умножителей (OMM) из [37]

Процесс обновления входной матрицы упрощен, чтобы противодействовать задержке, вызванной обновлением весов в банке весов. Это достигается путем кодирования ядра непосредственно во входную матрицу блока оптического матричного умножения (OMM), который показан на рисунке 15. Управление архитектурой банка Hitless MR передается FPGA, которая использует PD и ADC для получения суммированные сигналы от OMM. OMM были реализованы с использованием MR, но хранилище векторов было реализовано с использованием MZI. Учитывая, что основная проблема с черно-белой печатью – это большое количество MR, неясно, может ли модифицированная архитектура решить эту проблему, поскольку в работе не уточняется количество MR, даже с подходом к минимизации входной матрицы, обсуждаемым в статье. MR широко используются для реализации протокола B&W, но архитектура, которая рассматривает микродиски вместо MR из-за меньшей площади и энергопотребления, описана в [160]. В этой работе исследуется разработка и реализация фотонных матрично-векторных умножителей (MVM), сумматоров и сдвигов, которые являются фундаментальными вычислительными компонентами для вывода CNN, с использованием микродисков (рисунок 16). MVM (рисунок 16 (a)) использует коэффициент пропускания массива микродисков для представления элементов одной матрицы и входную мощность в массив микродисков из массива LD

для представления другой матрицы. Выходная мощность массива микродисков является произведением двух матриц. Электрооптический полный сумматор (рисунок 16 (b)) использует логические элементы КМОП для вычисления бита распространения (Pn) и бита генерации (Gn), необходимых для полного сумматора, а также микродиски для вычисления суммы и переноса. Значения Pn и Gn используются для модуляции микродисков. Модулируя интенсивность света, один из оптических сумматоров на выходе реализует логический элемент «исключающее ИЛИ», а другой генерирует логический элемент «ИЛИ», тем самым реализуя все необходимые операции для операций суммирования и переноса. Авторы также предлагают бинарный шифтер с использованием микродисков, как показано на рисунке 16 (c). Операция переключения выполняется путем настройки состояния включения / выключения переключателей пересечения микродиска. Авторы также описывают упрощенные модели CNN, называемые мощностью двух квантованных моделей CNN (P2Q-CNN), чтобы избежать зависимости от АЦП и повысить точность вывода CNN с незначительным падением точности (ниже 1%). В этой альтернативной архитектуре вместо MVM используется комбинация фотонного сумматора и сдвига. Для тестирования архитектуры авторы использовали тесты, основанные на наборах данных MNIST и ImageNet. Они сравнили эту архитектуру с ускорителем PIM ISAAC на основе ReRAM [20] и показали в 13 раз лучшую производительность на ватт. Недавняя работа [161] объединила MR и MZI для разработки базового блока оптического умножения и накопления (OMAC), который используется в ускорителе для CNN, называемом PIXEL. В работе описаны две версии ускорителя: гибридная версия, которая оптически умножает и накапливает электрически, и полностью оптическую версию, которая умножается и накапливает оптически. Гибридная версия использует MR для реализации функции И с MR, управляемым с помощью синапсов, с накоплением сдвига, выполняемым электрически. Побитовая операция И вместе со сдвигом-накоплением используется как альтернатива оптоэлектрической операции MAC. Конструкция полностью оптического сдвига и ADD использует MZI для выполнения операции сдвига-накопления с малой задержкой и малым энергопотреблением оптически и путем каскадного объединения MZI. Синхронизация сигналов с выхода И достигается за счет задержки распространения в плечах ИМЦ. Для создания задержки распространения оптических сигналов используются ИМЦ больших размеров и 6-миллиметровые плечики волновода между этими ИМЦ.

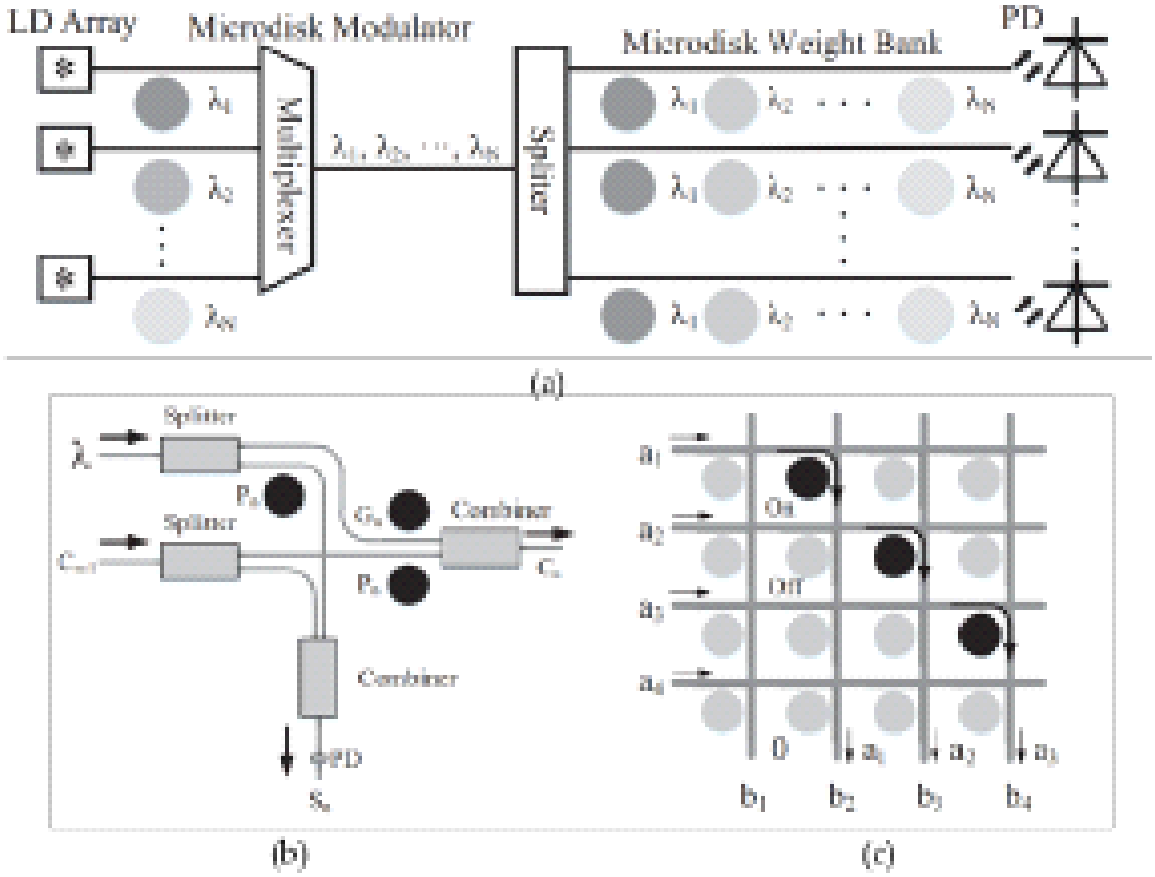


Рис. 16: (a) Микродисковый матрично-векторный умножитель (MVM) на основе микродиска, (b) полный электрооптический сумматор; (c) фотонный бит сдвиг, из [160].

Выходной сигнал этих взаимосвязанных MZI эффективно сдвигает входной бит. Предлагаемая архитектура имеет регистровые файлы для хранения веса фильтра в каждом OMAC. OMAC расположены в сетке, а выходы нейронов проходят через фотонные межсоединения как в x-, так и в y-измерениях. Синапсы предварительно загружаются в OMAC, и предлагаемый дизайн предполагает синхронизированное срабатывание нейронов для реализации функциональности MAC. Гибридный и полностью оптический подходы сравнивались с полностью электрической архитектурой с помощью моделирования для моделей ResNet, GoogleNet и ZFNet. Полностью оптический подход показывает лучшую энергоэффективность, чем полностью электрический, и в этом отношении сопоставим с гибридным подходом. Гибридный подход, основанный только на MR, занимает значительно меньшую площадь, чем полностью оптическая версия архитектуры, в которой используются как MR, так и MZI. Протокол B&W также использовался для сетей SNN, как описано в [90]. Предлагаемая архитектура использует лазерные нейроны в сочетании с двумя разными весовыми банками MR для взаимодействия с сигналами WDM. Банки весов используются для обозначения

возбуждающего и тормозящего веса. Веса накапливаются с помощью сбалансированной пары PD перед использованием для возбуждения лазерного нейрона. Один из PD в паре принимает мощность сигнала от банка возбуждающих весов, а другой принимает энергию от банка тормозных весов. Короткий провод включен для выполнения операции вычитания, тем самым считая значения из запрещающего банка отрицательными. Комбинация банков весов, PD и LD, действующих как пусковой механизм, моделирует базовый нейрон с импульсами и называется в этой работе узлом сети обработки (PNN). Сигналы WDM передаются между этими PNN с использованием широковещательного цикла (BL). Множественные широковещательные петли могут быть соединены друг с другом иерархическим образом через сопрягающиеся PNN. Взаимодействующие PNN – это PNN, которым поручено принимать выходные значения от одного BL и передавать их другому, по существу, действуя как оптический маршрутизатор для сигналов. Авторы [161] исследовали эту стратегию, чтобы учесть повторное использование спектра и улучшить параллельную обработку. В работе не предусмотрен экспериментальный раздел для демонстрации возможностей предложенной архитектуры. Скорее, в этой работе изучалась осуществимость нейронных сетей с пиковыми импульсами на основе протокола B&W. Ключевое наблюдение в работе заключается в том, как использование иерархической архитектуры широковещательного цикла обеспечит большую свободу пространственной компоновки, чем другие традиционные аппаратные нейроморфные системы.,Есть и другие случаи, когда MR используются для реализации SNN, кроме как через протокол B&W. Например, в [162] MR используются для реализации STDP на кристалле. Авторы [162] внедрили Ge2Sb5Te5 (GST), популярный и хорошо изученный материал ПКМ [163] – [166], поверх кольцевого волновода в кольцевом резонаторе. Это позволяло контролировать распространение света через порты путем простого изменения состояния GST. В этом случае PCM и его различные фазы действуют как память в синапсе. Авторы [162] также обсуждали потенциальную интеграцию нейрона интеграции и запуска с использованием MRs и GST в структуру SNN, состоящую из биполярных весов (веса с положительными и отрицательными значениями). Положительные и отрицательные взвешенные суммы вычисляются с использованием двух отдельных механизмов скалярного произведения и вводятся в два разных MR. Двунаправленное интегрирующее действие двух портов MR используется для расчета эффективного мембранного потенциала под действием биполярных взвешенных сумм. Выбросы на выходе генерируются, когда эффективный мембранный потенциал нейрона превышает пороговое значение. Получив стимул скалярного продукта, нейроны интегрируют свой мембранный потенциал на этом

временном шаге. В работе тестировалась эта архитектура с использованием набора данных MNIST и предполагалось, что механизм скалярного произведения работает безупречно.

Некоторые недавние попытки были направлены на устранение этих ограничений. Авторы в [37] предложили новую архитектуру, которая использует параллельное расположение MR вместо их каскадирования, хотя они используют неэффективные MZI для хранения векторов. Параллельное «безударное» расположение MR использовалось для уменьшения тепловых перекрестных помех между соседними MR и для достижения лучшего разрешения по весу. В других работах было предложено повторно использовать реализованные слои, например [156], а также методы для снижения энергопотребления и увеличения скорости работы за счет уменьшения использования электронных компонентов [160]. Также были предложения использовать микродиски вместо MR для дальнейшего увеличения плотности интеграции на чипах [161]. Но, как отмечено в [202], все варианты некогерентных архитектур могут страдать от низкой пропускной способности, поскольку электронные компоненты, такие как память, могут не работать с такой высокой частотой, как их фотонные компоненты. Для преодоления этих ограничений некогерентных архитектур необходимы дальнейшие исследования.

Работа подчеркивает жизнеспособность основанного на PCM STDP в нейроморфных архитектурах и демонстрирует это, демонстрируя более быстрые операции чтения / записи и низкое энергопотребление для фотонной архитектуры по сравнению с электронным аналогом. При их моделировании с использованием этой архитектуры была достигнута точность тестирования 98,3%. В связанной работе [167] обсуждается, как MR могут быть использованы для включения задержек спайков в фотонную SNN. В нескольких работах также были предложены архитектуры фотонных резервуарных вычислений (RC) на основе MR, такие как резервуар MR 5x5 для классификации цифровых образов с высокой скоростью передачи в [168]. В данной работе резервуар образован случайно связанными между собой MP. Смоделированная архитектура смогла достичь ошибки классификации всего 0,5%, предлагая скорость передачи данных до 160 Гбит / с для цифровых слов восьмибитовой длины для распознавания битовых образов в настраиваемом наборе данных. Авторы в [169] исследовали дизайн коллектора на основе вихревой топологии 4x4, который использует MR. В работе также были продемонстрированы основные логические операции. В таких архитектурах узлы состоят из нелинейных элементов (MR) и являются частью повторяющейся сети, что является отклонением от исходной топологии завихрения, представленной в [170]. Эта архитектура широко использовалась в исследованиях фотонного RC. Завихрение в трактах данных

обеспечивает достаточное смешение входных сигналов / матрицы весов. Традиционно архитектуры резервуаров устанавливают свои узлы в состояние, близкое к нестабильности для правильной работы резервуара, чтобы гарантировать, что они имеют достаточную память о прошлых входах и хорошо реагируют на новые входные данные. MR в [169] устанавливаются на эту рабочую точку после подробного анализа стабильности MR в работе и резонанса при различных значениях входной мощности, а также вызванной температурой оптической отстройки от резонанса. Таким образом, некогерентные микроархитектуры нейронов, использующие MR, являются одними из наиболее широко используемых компонентов для реализации архитектур фотонных нейронных сетей. Эти некогерентные архитектуры, использующие MR, охватывают SNN, ANN (MLP, CNN) и RC. Большинство архитектур SNN, MLP и CNN используют протокол B&W для передачи весов по сети. Некоторые усилия выявили несколько проблем с этим протоколом, таких как гетеродинные перекрестные помехи, искажающие значения весовых коэффициентов, и увеличивающееся количество MR, необходимых для реализации более крупных сетей, особенно когда используются более высокие степени WDM [37]. В других работах было предложено повторно использовать реализованные слои, как в [156], а также методы для снижения энергопотребления и увеличения скорости работы за счет уменьшения использования электронных компонентов [160]. Более того, были предложения использовать микродиски вместо MR для дальнейшего увеличения плотности интеграции в чипах, как в [160]. MR также использовались для распространения веса в SNN с использованием протокола B&W, как в [90], и для реализации STDP в фотонных SNN [162], [171]. Они также появляются в RC для создания узлов в резервуарах, которые состоят из случайно связанных узлов [169].

5.2 Сетевые архитектуры на основе оптической интерферометрии

Реализации нейронных сетей на основе оптической интерферометрии обычно основываются на манипуляции фазой и амплитудой электрического поля одной длины оптической волны. Это согласованные архитектуры, использующие согласованные микроархитектуры нейронов, обсуждаемые в разделе 4.2. Когерентные архитектуры во многом полагаются на MZI и широко используются в реализациях MLP. MZI реже использовались в качестве модуляторов интенсивности в некоторых реализациях SNN. Большое количество узлов, необходимых в резервуаре, и требования к большой площади MZI делают их не очень популярными в реализациях RC. Когерентные архитектуры на основе MZI используют универсальные линейные сетки MZI для реализации требуемых матрич-

ных умножений, необходимых для нейронных сетей. Веса контролируются путем управления фазой и амплитудой оптических сигналов, что осуществляется путем имплантации аттенюаторов и фазовращателей на плечах MZI. Это было продемонстрировано в [172], где 2 × 2 светоделители и фазовращатели в форме ИМЦ были запрограммированы для обеспечения независимого управления амплитудой и фазой света для набора оптических каналов. В работе [173] изготовлен и продемонстрирован оптический матричный умножитель 4x4 на основе MZI. Здесь архитектура построена на основе предпосылки, что идеальный многопортовый реконфигурируемый интерферометр на основе MZI NxN представляет собой специальную унитарную (SU) группу степени N, SU (N), которая состоит из n MZI с 24 N оптическими каналами, образующими матрица унитарного преобразования. В [173] структура состоит из секции SU (N = 4), за которой следует секция диагонального матричного умножения (DMM) (рисунок 17 (a)). Цифровой мультиметр может быть расширен в зависимости от приложения и может формировать полный SVD посредством каскадирования. Этот оптический матричный умножитель 4x4 использовался для реализации однослойной нейронной сети. Производительность нейронной сети оценивалась путем классификации 50 выборок данных синтетического линейно разделяемого многомерного гауссовского набора данных, для которых она смогла достичь точности 72%. В работе [174] описана другая реализация сетки на основе матричного SVD, которая может реализовывать произвольные неунитарные матрицы с использованием MZI. Методология на основе разложения по сингулярным значениям (SVD) используется для выполнения разложения матриц в унитарные матрицы, и эти упрощенные матрицы реализуются на кристалле. Более конкретно, SVD – это процесс, с помощью которого матрица может быть разложена на три матрицы, две унитарные матрицы V и U и диагональную матрицу, состоящую из ненулевых сингулярных значений Σ. Процесс SVD может быть реализован в сетке MZI с помощью диагональной матрицы, которая реализует амплитуду и фазу, в то время как универсальные унитарные матрицы следуют схемам, предложенным в [172] и [175]. Окончательная архитектура этого подхода показана на рисунке 17 (b). (a) (b). Авторы в [32] предложили архитектуру, которая использует подход на основе SVD для реализации необходимых матричных вычислений (рисунок 18), где векторы кодировались в интенсивности и фазе света, а затем подавались в каждый уровень сети. SVD используется для разложения матриц для умножения на унитарные матрицы, которые могут быть закодированы в сетку MZI. Сигналы работы SVD и кодирования для MZI были сгенерированы с помощью цифрового компьютера. После того, как матрицы закодированы в сетку MZI, можно выполнить умножение матриц между ними, позволяя оптическому сиг-

налу пассивно проходить через сетку. Ключевые преимущества SVD – это уменьшенная сложность работы и уменьшение размерности, что помогает снизить стоимость эксплуатации имеющейся модели DNN. Каждый уровень этой предложенной модели состоит из блока оптических помех (OIU) и блока оптической нелинейности (ONU). В этой работе функциональность ONU была реализована с помощью цифровой электроники, тогда как OIU был реализован в фотонной интегральной схеме, которая выполняла оптическое матричное умножение с использованием подхода SVD, как описано в [174].

В работе [32] обсуждалось, как использовать такую архитектуру для распознавания гласных. Они также использовали прямое распространение с методом конечных разностей вместо обратного распространения для обучения архитектуры. Архитектура смогла достичь 76,7% точности классификации гласных, что ниже 91,7%, достигнутой той же архитектурой, реализованной на обычном 64-битном цифровом компьютере. Авторы объясняют более низкую точность ограниченным вычислительным разрешением (24 бита в отличие от 64 бита обычного компьютера) оптической нейронной сети.

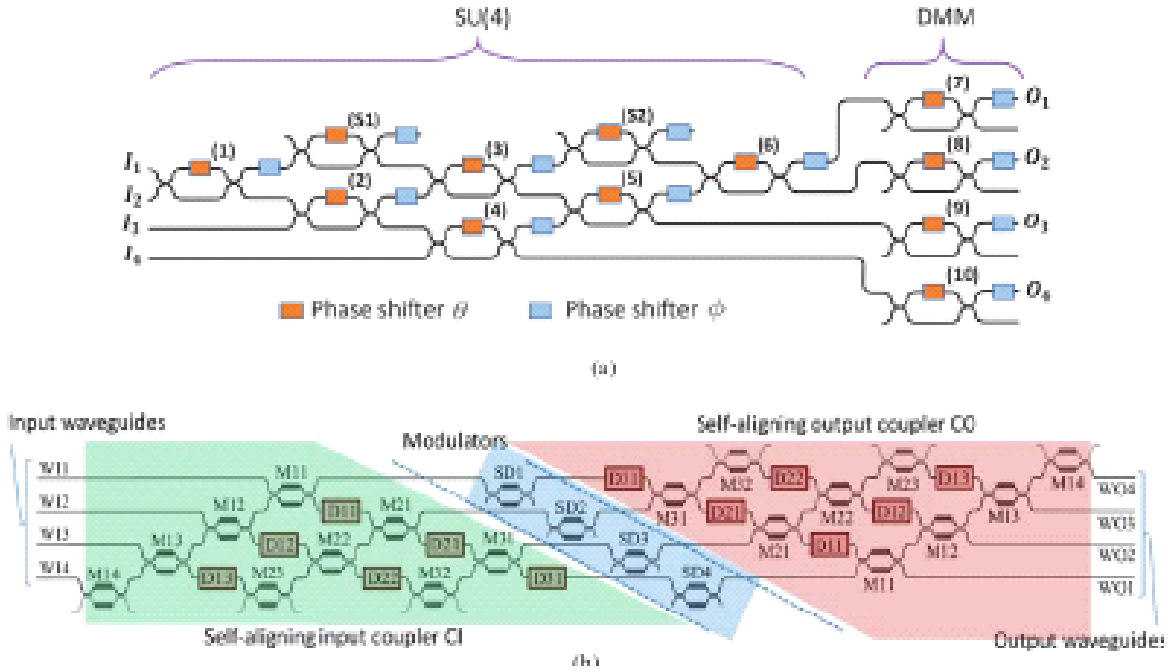


Рисунок 17: (a) Реконфигурируемый линейный оптический процессор на основе MZI 4 × 4 из [173]; SU – специальная унитарная группа, а DMM – единица умножения диагональных матриц; (b) Универсальная линейная сетка для разложения по сингулярным значениям (SVD) на основе MZI, как описано в [174]. Матрицы, участвующие в SVD, реализованы: V зеленым, диагональная матрица Σ синим и UT красным.

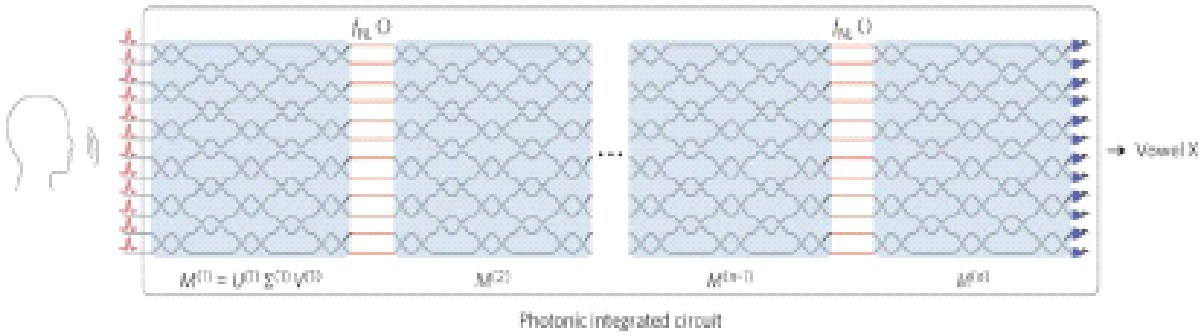


Рисунок 18: Архитектура из [32], в которой используется SVD для реализации умножения матриц для распознавания гласных с использованием фотонной интегральной схемы (PIC).

Было отмечено, что MZI имеют гораздо большую площадь основания (которая может составлять до нескольких миллиметров [161] или от десятков до сотен микрометров), чем их аналоги (например, MR), и, как отмечалось [132], эта большая площадь в сочетании с накоплением Количество фазовых ошибок в сетке на основе MZI может ограничивать масштабируемость нейронных сетей, построенных с помощью MZI

Было проведено исследование, например, в [176], в котором основное внимание уделялось сокращению общего потребления площади этими архитектурами, будь то использование методов сокращения весовых матриц, представленных сетками MZI, или использование других устройств в тандеме.

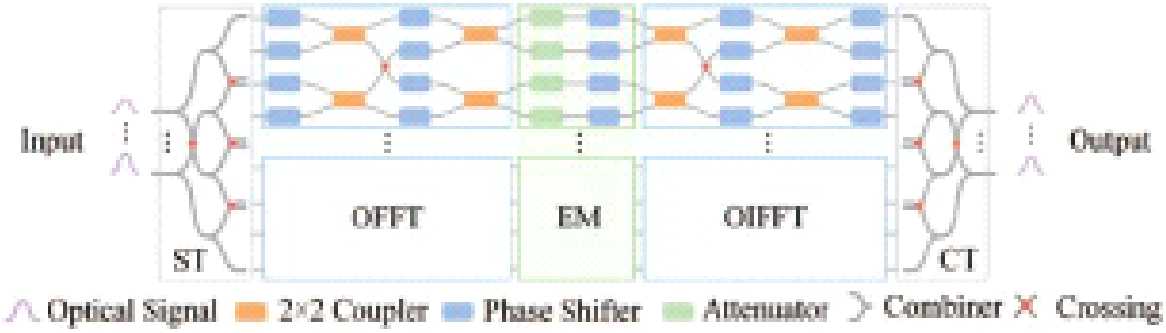


Рисунок 19: Архитектура фотонной структурированной нейронной сети на основе БПФ-ОБПФ, описанная в [144]. Основанный на БПФ анализ модели направлен на упрощение модели и, таким образом, снижение потребления энергии и площади сеток MZI.

В работе [176] продемонстрирована методология, основанная на быстром преобразовании Фурье (БПФ) для уменьшения площади и энергопотребления сеток MZI, используемых для реализации MLP. Это достигается за счет разрежения сети за счет уменьшения общего количества используемых весов, тем самым сжимая нейронную сеть. Предлагаемая архитектура основана на структурированных нейронных сетях

с циркулянтным матричным представлением. Структурированные нейронные сети – это класс нейронных сетей, которые специально разработаны для снижения вычислительной сложности, весовые матрицы которых регуляризованы с использованием композиции структурированных подматриц [177]. Структурированные нейронные сети используют циркулянтные весовые матрицы, которые можно эффективно вычислить с помощью БПФ и обратного БПФ (ОБПФ). Весовые матрицы дополнительно сокращаются с использованием регуляризации Group Lasso [178], и эти операции могут быть реализованы в сетках MZI с использованием каскадных аттенюаторов / усилителей и фазовращателей (рисунок 19). Авторы [176] адаптировали эту методологию из-за трудности сокращения архитектур на основе SVD. Архитектура была протестирована с использованием набора данных MNIST в сравнении с архитектурами на основе SVD, чтобы показать, насколько эффективен их метод в сокращении общей площади, занимаемой сетками MZI. Результаты показывают, что архитектура смогла достичь точности тестирования, близкой к 98,5%, при значительном сокращении общего потребления площади. Авторы [144] предложили реализацию бинарной нейронной сети на основе когерентной MZI с весами, ограниченными до +1 или -1. Функция активации – это функция символического решения, которая преобразует любое действительное число, сопоставленное с ним, в +1 или -1. Веса бинаризации кодируются в MZI путем сдвига напряжений на внутренних и внешних фазовращателях на плечах MZI. Действительная и мнимая части оптического сигнала с двусторонней поляризацией, модулированного синфазной и квадратурной составляющими (IQ), используются для обучения модели в симуляции. Входом в модель является реальная и мнимая часть сигнала, а выходом – предсказание позиции входа нейронной сетью. В этой работе была протестирована архитектура, состоящая из семи скрытых слоев, для классификации по ближайшим соседям совокупности, образованной действительной и мнимой частями сигнала DP-QPSK 100 ГГц. Точность классификации, близкая к 100%, была достигнута для входных сигналов с высоким отношением сигнал / шум (SNR), в то время как точность, близкая к 90%, была достигнута для сигналов с низким SNR. Как упоминалось ранее, архитектуры на основе MZI обычно когерентны и используют только одну длину волны. Но MZI также использовались для реализации архитектур на основе WDM или некогерентных архитектур. Например, в работе [133] был продемонстрирован ускоритель умножения фотонных матриц с использованием MZI и решетчатых волноводов с 26 решетками (AWG) – многомодовых помех (MMI). Один блок сбалансированного детектора с ответвителем AWG-MMI может успешно выполнять матричное умножение с использованием WDM и схемы когерентного гомодинного де-

тектирования. MZI используются как модуляторы интенсивности, которые передаются в умножитель (рисунок 20). В работе [143] продемонстрирована полностью оптическая WDM RNN, использующая SOA-MZI в качестве блока активации, включенного в контур задержки обратной связи. Чтобы имитировать полностью функциональный Gated-Recurrent-Unit (GRU), авторы интегрировали механизм стробирования (SOA-MZI), чтобы обеспечить гибкую реконфигурацию функций забывания в GRU. SOA встроены в плечи MZI и действуют как преобразователи длины волны модуляции перекрестного усиления. RNN была построена с использованием этого устройства и протестирована с использованием WDM с четырьмя входами. Полезность RNN была протестирована путем запуска тестового приложения для финансового прогнозирования с использованием набора данных FI-2010. Стробируемая оптическая RNN смогла получить более высокий балл F1 (41,85%), чем оптическая и обычная RNN.

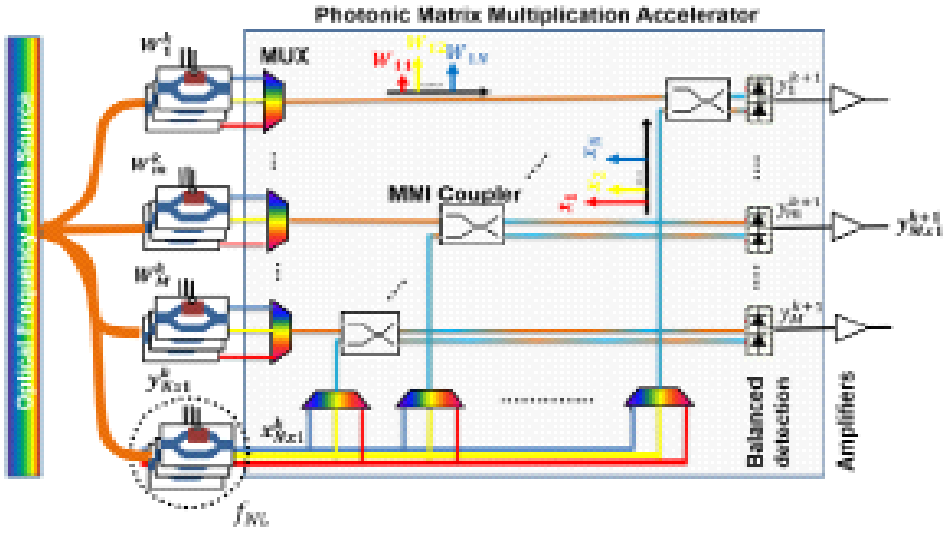


Рисунок 20: Каскадная аналоговая структура искусственной нейронной сети с прямой связью со схемой фотонного матрично-векторного умножителя и нелинейностью модулятора Маха-Цендера, как показано в [133].

Таким образом, когерентные микроархитектуры нейронов, которые используют устройства интерферометра, такие как MZI, широко используются в архитектурах фотонных нейронных сетей из-за их способности эффективно представлять матрицы для операций нейронной сети, но за счет большей площади накладных расходов, чем MR, и восприимчивости. к искажению фазового шума. Основной принцип MZI для работы нейронной сети основан на настройке фазы и амплитуды проходящей оптической длины волны, что может быть легко достигнуто путем интеграции фазовых и амплитудных тюнеров в плечах MZI. MZI обычно располагаются в виде сетки в тех работах, где они используются, при этом SVD также используется для эффективного представления матриц. Уменьшение площа-

ди MZI и искажения фазового шума предпринимается с помощью подходов регуляризации (например, [32], [144]) или с помощью моделей нишевых нейронных сетей для уменьшения общего количества MZI ([176]). Обычно архитектуры, использующие MZI, используют согласованные принципы для работы, но MZI также использовались в некогерентных подходах, которые используют WDM, например, в реализации RNN в [133], которая использует SOA и MZI в комбинации. Они не используются ни в каких архитектурах на основе RC, вероятно, из-за требований к большой площади, которые реализация на основе MZI потребует для реализации большого количества нелинейных узлов в реализации RC.

5.3 Сетевые архитектуры на основе дифракционной оптики

Дифракционная оптика на кристалле также используется для реализации архитектур фотонных нейронных сетей. Очевидным преимуществом использования этих методов является пассивная реализация необходимых функций за счет использования физики дифракционной оптики. Это отличается от MR и MZI, поскольку они используются в качестве активных устройств, которые требуют активной настройки (как в случае MR) или механизмов управления для управления фазой (как в случае MZI). Различные архитектуры обсуждают реализации MLP за счет интеграции дифракционной оптики на кристалле. Они обычно используют соединители AWG / Star вместе с контроллерами поляризации и SOA для достижения различных функций, необходимых для реализации нейронной сети. Архитектура, описанная в [179], описывает одну из таких реализаций, в которой AWG используются для уменьшения шума и увеличения разрешения накопленных значений веса, демонстрируя операции нейроморфного взвешенного сложения в кросс-коммутации 8x8 InP. Веса умножаются на оптические сигналы путем настройки усиления SOA. Затем выходные сигналы объединяются для накопления результатов, и операция взвешенного сложения выполняется с использованием PD для получения результирующего оптического тока. В этой системе высокоточная операция умножения и накопления с 4-битной точностью достигается с ошибкой менее 0,2. Авторы утверждают, что эту систему можно масштабировать для формирования жизнеспособных фотонных DNN. В работе [134] изучалась комбинация AWG (рисунок 21 (a)) и MZI для реализации CNN (рисунок 21 (b)). Распространение в свободном пространстве в AWG использовалось для имитации приблизительной операции дискретного преобразования Фурье (ДПФ). Каскадирование двух операций ДПФ с фазовой и амплитудной маской между ними использовалось для представления операции свертки. Слой объединения был реализован как фильтр нижних частот, который пропускает только низкочастотные компоненты ДПФ.

Фильтр был реализован с тремя AWG с фазовой и амплитудной маской между первыми двумя. Наконец, полностью связанный слой был реализован в виде сетки MZI с настраиваемыми аттенюаторами / усилителями в ее руках. Сетка MZI реализует SVD для представления унитарной матрицы, полученной в результате операций DFT. Авторы использовали алгоритм БПФ Кули-Тьюки [180], чтобы уменьшить количество используемых MZI и, таким образом, уменьшить объем реализации. Алгоритм Кули-Тьюки БПФ использует комбинацию ДПФ для создания аппроксимации непрерывного БПФ и чрезвычайно популярен в приложениях, основанных на БПФ. В этой работе также изучалось, как шум в масках, применяемых к выходным данным AWG, повлияет на точность архитектуры для задачи классификации набора данных MNIST

В работе исследовалось, как различные источники шума будут влиять на точность тестирования архитектуры, с учетом гауссовой амплитуды, фазы и сложного шума, добавленного к матрице AWG. Было показано, что архитектура устойчива к шуму после обучения с зашумленными входными сигналами. Было заявлено, что за счет повторного обучения выходного слоя с шумом точность архитектуры существенно улучшилась даже при сильно зашумленных входах.)

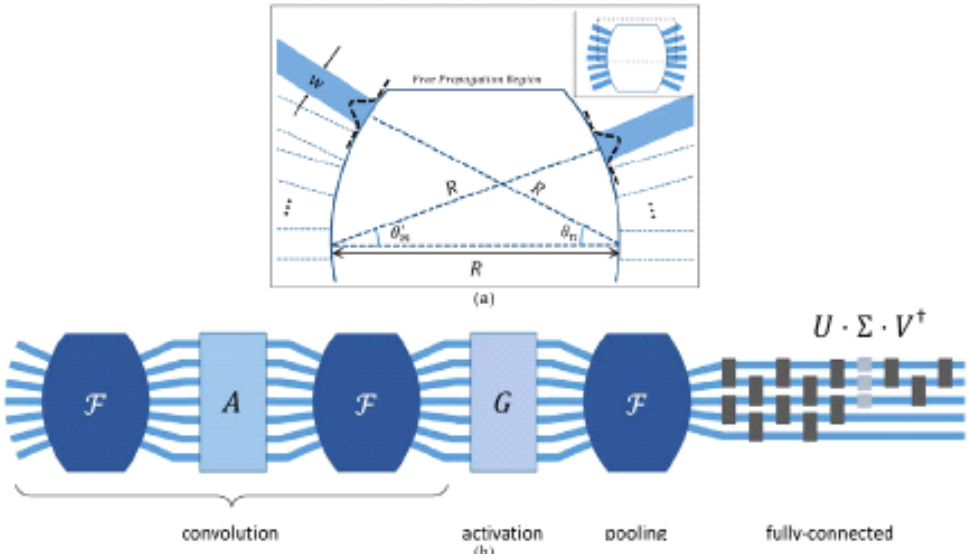


Рисунок 21: (a) Схема звездообразного ответвителя N × M или AWG. R – радиус конфокальных окружностей, составляющих область распространения в свободном пространстве. θn – угол n-го входного волновода, θm – угол m-го выходного волновода. w – параметр ширины моды волновода. (b) Реализация CNN на основе AWG, в которой используется тот факт, что оптические сигналы, проходящие через область свободного распространения звездообразного ответвителя, подвергаются дискретному преобразованию Фурье (ДПФ). A и G – маски фильтров [134].

В работе [34] описан резервуар, основанный на топологии завихрения 4x4, со слоем считывания, состоящим из нелинейных оптических модуляторов (рисунок 22). Используемый резервуар состоит из пассивных эле-

ментов, как в [169]. Известные подходы, адаптированные здесь, включают демонстрацию использования 4-импульсной амплитудной модуляции (4-PAM) в настройке расчета резервуара, где булевы операции, такие как XOR, являются эталоном. Кроме того, авторы представили архитектуру RC, в которой используются столбчатые кремнеземные рассеиватели с полостью в качестве пассивного элемента в коллекторе. Для экспериментов и моделирования авторы увеличили свой резервуар до 20x20 узлов. Они смоделировали эту архитектуру, показанную на рисунке 23, с помощью моделирования FDTD. Архитектура также продемонстрировала возможности классификации, будучи обученной отличать раковые клетки от нормальных. Эффективность этой безмаркированной классификации сравнивалась с предыдущей работой [129], в которой использовались столбчатые рассеиватели без полости, что заставило работу [129] использовать волны с более низкой длиной волны (УФ) для резервуара. Подход с использованием УФ для 28 этой задачи оказался непрактичным из-за высокой стоимости УФ-лазеров и возможного повреждения клеток, которое может вызвать УФ-излучение. Сообщается, что новая архитектура [34], основанная на столбчатых рассеивателях с резонатором, достигла точности, сопоставимой с подходом [129]. Встроенный дифракционный механизм использования VCSEL для формирования дифракционно связанной решетки VCSEL был использован для формирования резервуара в [181]. В данной работе предлагалось устанавливать веса с помощью пространственного модулятора. Перед архитектурой была поставлена задача распознавания заголовков, и она могла распознавать до 5-битных заголовков. В работе [182] описана крупномасштабная система, в которой используется дифракционная механика в считывающем слое, который является полностью оптическим и состоит из цифровых микрозеркал. Но нелинейность реализована в электрической области, что сильно ограничивает частоту обновления до 5 Гц. Эта работа продемонстрировала архитектуру с 2025 нелинейными узлами, реализованную как пиксель в пространственном модуляторе света (SLM). SLM будет отображать текущее состояние резервуара в виде спекл-шаблона, который можно считывать с помощью камеры, и следующее состояние, необходимое для резервуара, вычисляется и кодируется в SLM. Перед архитектурой была поставлена задача предсказать следующий шаг в нелинейном хаотическом временном ряду Макки-Гласса [183] с нормированной среднеквадратической ошибкой (NMSE) в качестве критерия для оценки производительности архитектуры. Было показано, что архитектура позволяет достичь значения NMSE 0,013 для задачи прогнозирования. Другая архитектура коллектора, в которой используется SLM, описана в [184]. Эта архитектура также использовала резервуар на основе SLM и работала на частоте 640 Гц, которую авторы относят к лучшему оборудованию SLM. Эта архитектура также была протестирована с ис-

пользованием предсказания хаотического временного ряда Макки-Гласса. Архитектура имеет до 16385 узлов, опять же в виде пикселей в SLM, и, как сообщается, имеет значение NMSE ниже 0,3.

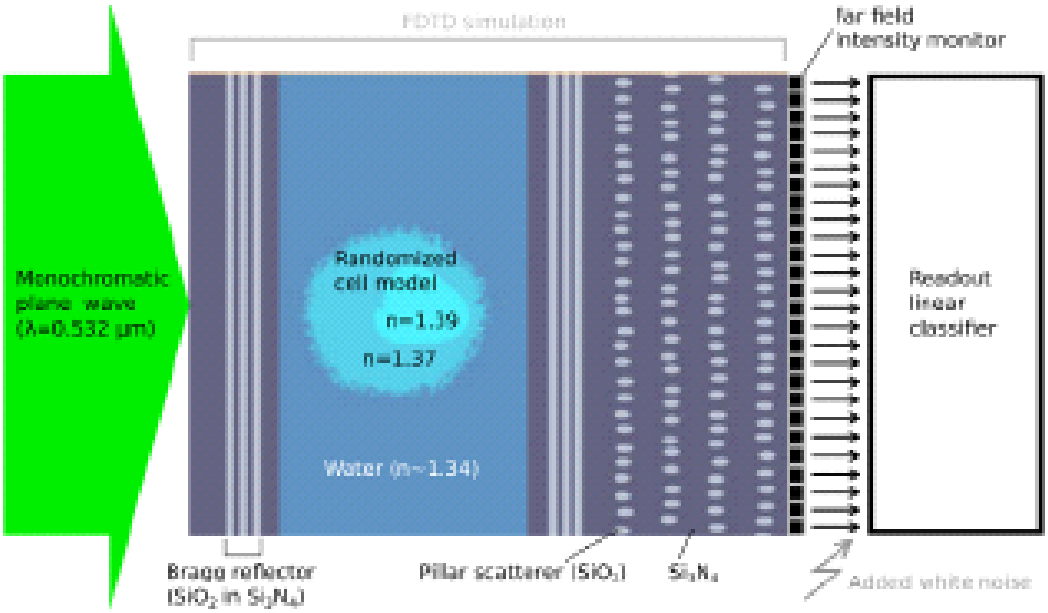


Рисунок 22: Подробная диаграмма из [34], изображающая моделирование FDTD их архитектуры на основе рассеивателя, показывающая различные смоделированные компоненты.

Таким образом, дифракционная оптика использовалась для реализации архитектур MLP, CNN, RNN и RC. Эти реализации используют разнообразный набор устройств, таких как встроенные в кристаллы AWG, пассивные элементы, такие как столбчатые рассеиватели, дифракционно связанные VCSEL и SLM. Эти архитектуры, как правило, используют пассивные свойства оптических устройств для достижения необходимых функциональных возможностей, таких как пассивное преобразование световых волн методом ДПФ, когда они проходят через AWG, или использование SLM для формирования огромных резервуаров. Часто архитектуры, использующие дифракционную оптику, также используют другие устройства, такие как MZI, SOA и VCSEL. Однако потребность в специально разработанных устройствах (SLM, микрозеркала, рассеиватели) препятствует программируемости и реализации компактных и масштабируемых реализаций. В результате реализации на основе дифракционной оптики не так популярны для ускорения нейронных сетей на уровне микросхем.

5.4 Архитектуры на основе оптического усиления и генерации

Здесь мы обсуждаем архитектуры фотонных нейронных сетей, которые используют SOA и VCSEL. Это известные реализации SNN для реализации STDP в сети. STDP считается фундаментальным механизмом пластичности синапсов человеческого мозга [185] – [187]. Вот как веса

присваиваются синапсам в мозге, в зависимости от временных отношений между пресинаптическими и постсинаптическими спайками. Вес, связанный с синапсом, увеличивается, если пресинаптический спайк появляется перед постсинаптическим спайком, и уменьшается в противном случае. Этот метод обычно реализуется в фотонных нейроморфных архитектурах с использованием плоских полупроводниковых оптических усилителей (SOA). В [188] DNN реализована и экспериментально проверена с помощью SOA. Функции смещения и активации реализуются с помощью цифровой электроники. Значение смещения добавляется к данным после обнаружения. В работе использовалась функция активации tanh. Эта архитектура реализует одну операцию нейрона путем смещения до шести SOA: один SOA в качестве предварительного усилителя, один SOA для выбора входного вектора и четыре SOA, действующих как модуляторы интенсивности для представления весов. Для представления работы уровня требуется в общей сложности четыре взвешенных сложения, которые выполняются смещением 21 SOA: один SOA предварительного усилителя, четыре SOA для выбора входных векторов и 16 SOA, действующих как веса. AWG также использовались в архитектуре для мультиплексирования / демультиплексирования длин волн. В результате была предложена нейронная сеть, состоящая из трех слоев нейронов. Классификация цветов ириса Фишера использовалась для проверки точности архитектуры, на которой при моделировании была достигнута точность прогноза 85,8% по сравнению с точностью 95% в цифровой электронике. Работа [189] демонстрирует контролируемое обучение, управляемое оптическим STDP, с использованием SOA и модулятора электропоглощения (EAM). Линейная комбинация эффектов уменьшения усиления в SOA и насыщения поглощения в EAM используется для реализации эффектов STDP. Для обучения устройства обработки импульсов использовалась выборка импульсов учителя, представляющая ожидаемый выходной сигнал последовательности импульсов, при этом фотонный STDP автоматически настраивал свое усиление, чтобы процессор импульсов соответствовал выборке импульсов учителя. Следуя этой модели, авторы [190] реализовали обучение с подкреплением, основанное на вознаграждении, которое стало возможным благодаря фотонному блоку STDP, построенному с двумя SOA. Это была эмуляция биологического поведения синапсов STDP и того, как мозг обучается с помощью принципов обучения с подкреплением. Здесь был введен новый модулирующий элемент путем изменения текущей инъекции в SOA и использован для имитации функции вознаграждения, необходимой для обучения с подкреплением. Работа экспериментально продемонстрировала, как функция вознаграждения настраивается фотонным STDP в зависимости от подкрепления. В работе [123] продемонстрирован фотонный модуль STDP для контролируемого обучения и неконтролируемого распо-

знавания образов на основе единой SOA. Предлагаемая установка впервые продемонстрировала обобщенный алгоритм Хеббиана [191] для синаптической модификации, который в нейробиологии называется зависимой от активности синаптической пластичностью (ADSP). SNN является фотонным, но вычисление корреляции между постсинаптическими и пресинаптическими сигналами было рассчитано с использованием ЦП, наряду с вычислением правила обновления и управлением банком весов на основе SOA. В [192] SOA с вертикальным резонатором (VCSOA) [193] – [195] наряду с VCSEL использовались для формирования фотонных SNN. VCSOA рассматриваются как VCSEL, работающие ниже порога генерации, что обеспечивает простоту интеграции с VCSEL из-за их структурного сходства и низкого энергопотребления. Авторы основывали эту реализацию STDP на основе VCSOA на своей предыдущей работе в [125], в которой была представлена теоретическая и математическая модель для достижения фотонного STDP с использованием VCSOA. SNN был протестирован с помощью задания на распознавание произвольных паттернов спайков. Результаты показывают, что время постсинаптического всплеска сходится с временем всплеска входной последовательности всплесков посредством контролируемого обучения. Авторы [196] обсуждали полностью связанный фотонный SNN, состоящий из возбудимых VCSEL со встроенным насыщающимся поглотителем, для реализации обучения импульсной последовательности посредством обучения с учителем. Авторы включили фотонный STDP в алгоритм классического удаленного контролируемого метода (ReSuMe), чтобы реализовать контролируемое обучение SNN. В работе [145] были представлены быстрые системы VCSEL-нейронов для нейроморфных фотонных приложений в двух различных архитектурах, а именно с одним VCSEL-нейроном, подверженным задержанной оптической обратной связи, и двумя взаимно связанными VCSEL-нейронами. Это имитировало работу биологических нейронных цепей сетчатки. Взаимосвязанные нейроны VCSEL использовали для имитации связи между биполярными клетками и ганглиозными клетками сетчатки глаза, при этом нейрон VCSEL представлял фоторецепторы. Используя эти VCSEL-нейроны, исследование успешно имитировало нейронные схемы ВКЛ и ВЫКЛ в глазу. В [197] для образования резервуара использовались связанные SOA. Но использование активных элементов, таких как SOA, сделало бы мощность архитектуры резервуара неэффективной, хотя использование активных элементов значительно сокращает площадь, занимаемую архитектурой. Они обошли эти проблемы, используя пассивные элементы в архитектуре RC из [169]. В работе [169] был продемонстрирован резервуар с использованием только пассивных элементов: волноводы, разветвители и сумматоры были единственными компонентами, используемыми в резервуаре. Коллектор был реализован в виде 16-узловой квад-

ратно-ячеистой сети с несколькими петлями обратной связи. В архитектуре, описанной в [169], требуемая нелинейность больше не в пределах коллектора и реализуется на уровне считывания с использованием PD. Выходной сигнал каждого узла представляет собой линейную суперпозицию комплексных амплитуд входных волноводов этого узла. На уровне считывания комплексные амплитуды узлов коллектора преобразуются в действительные уровни мощности, которые затем используются в качестве входных данных для линейного классификатора. Архитектура из [169] была создана для выполнения базовых логических задач и распознавания заголовков до 5-битных заголовков с использованием предложенной архитектуры. Авторы также продемонстрировали способность архитектуры распознавать произносимые цифры и сообщили о минимальной частоте ошибок слов (WER) 4,5% для их связного резервуара на основе SOA. В [198] электрически модулированные кремниевые нанолазеры (СНЛ) используются в качестве резервуарного слоя их RC-архитектуры. Контур задержки SNL используется для генерации виртуальных узлов, а временное мультиплексирование используется для формирования резервуара. Веса устанавливаются с использованием случайной матрицы весов, вводимой через входной слой, в то время как взвешивание 30 и линейное суммирование происходит на выходном слое. Оптимизация веса выполняется путем минимизации ошибки наименьших квадратов между текущим и целевым весами. Перед тестированием архитектуры стояла задача предсказать следующий шаг в хаотическом временном ряду Санта-Фе [199]. Производительность архитектуры оценивалась путем вычисления нормированной среднеквадратичной ошибки (NMSE) между прогнозируемыми и целевыми значениями. Скорость обратной связи SNL была точно настроена для тестирования архитектуры и ее производительности. Сообщалось о NMSE 0,0359 для скорости обратной связи 10 нс-1. Таким образом, обсуждения в этом разделе в основном относятся к реализации SNN с использованием SOA и лазеров. Различные работы, перечисленные в этом разделе, сосредоточены на реализациях STDP для SNN, по большей части с использованием SOA и VCSEL для достижения синаптической пластичности на кристалле. Есть также несколько архитектур RC, предназначенных для приложений машинного обучения, которые также используют SOA и лазеры. Мы обсудили резервуар, построенный из SOA, реализующих нелинейные узлы, и резервуар полностью пассивных фотонных элементов. Реализация на основе пассивных элементов была исследована, чтобы обойти ограничения мощности и скорости активных элементов, таких как SOA на пласте. Мы также обсудили недавнюю работу, в которой нано-лазеры на кристалле, называемые SNL, использовались для формирования RC-архитектуры, в которой петли задержки лазера использовались для фор-

мирования виртуальных узлов, работающих с временным мультиплексированием.

5.5. Резюме

5.5 Резюме Литература, касающаяся архитектур фотонных нейронных сетей, обширна, равно как и методы и устройства, используемые для реализации этих архитектур. В этом разделе мы рассмотрели различные архитектуры и разделили литературу на реализации на основе резонаторов, реализации на основе интерферометра, реализации на основе дифракционной оптики и реализации на основе оптического усиления / генерации. Мы предоставили обзор литературы по архитектурам, описанной в разделе 5 в таблице 3. В таблице есть ссылки на работы (первый столбец); устройства, занимающие видное место в архитектуре (второй столбец); краткое изложение приложений, рассматриваемых как часть экспериментов (третий столбец); были ли в работе сфабрикованы результаты или моделирование, или и то, и другое, или нет (четвертый столбец); и важные результаты, представленные в документе (пятая колонка). Знак «-» в таблице представляет информацию, которая не предоставлена.

Таблица 3: Краткое изложение предыдущей работы над архитектурами фотонных нейронных сетей.

Reference	Devices utilized	Application	Fabricated (F) or Simulated (S)	Results achieved
[154]	MRs	Lorenz Attractor simulation to benchmark against a traditional CPU based CTRNN.	F	Reports 294× acceleration in simulation over traditional CPU based CTRNN.
[156]	MRs and MZMs	*AlexNet* CNN model	S	Claims 5 orders of magnitude faster speeds than fully electrical implementations.
[157]	MRs	MNIST classification using CNNs.	S	Faster when compared to GPU based implementations (2.8 to 1.4 times faster) and 0.75 times the power consumption.
[132]	MRs and SOAs. Weight fed into MR banks using memristor arrays.	Various benchmarks including MNIST tested on photonic CNNs.	S	Reduction in operation cost when compared to GPU based implementations, with up to 25× better computational efficiency.
[37]	MRs and MZIs	MNIST classification using MLPs.	S	Higher than 95% accuracy achieved at 14 bit resolution and custom MLP with 2048 neurons in hidden layer, for both types of weights. Non-negative weights give lower accuracy.
[160]	Microdisks	Image classification with CNNs.	S	13× better performance per Watt than ISAAC.
[161]	MRs and MZIs	Image classification with CNNs.	S	All optical design consumes only 5.1% the energy needed by all electrical accelerator, while being 31.9% faster.

[90]	MRs	This was an exercise to prove the feasibility of B&W based SNNs. No application-based experiment was conducted in this work.	–	–
[162]	GST embedded MRs	MNIST classification with MLPs.	S	98.06% accuracy.
[168]	MRs	High-bit-rate digital pattern classification using RC.	S	Classification error of 0.5% at 160 Gbps for 8-bit-length digital words.
[169]	MRs	Demonstration of Boolean operations using RC. Detailed analysis of XOR operations.	S	Demonstrated XOR operations at an error rate of 0.1. Also explored the relationship between error rate and input power modulation and optical detuning.
[172]	MZIs	Mathematical discussion of phase and amplitude control for unitary operator representation, using MZIs.	–	–
[173]	MZIs	Single layer neural network using the 4x4 optical processor described in the work, set to classifying data samples	F + S	Demonstrated an accuracy of 72% in classification of data samples.
[174]	MZIs	Mathematical and theoretical discussion of MZI based unitary matrix representation, and consequently, how a universal linear device may be fashioned. No application-based testing done.	—	–
[175]	MZIs	Mathematical and theoretical discussion of MZI based unitary matrix representation, with added discussion into error and loss tolerance of such a device.	–	–
[144]	MZIs	Binary neural network set to nearest neighbor classification of a constellation formed from 100 GHz DP-QPSK signal.	S	Close to 100% accuracy in classification achieved for high SNR signal, while accuracy close to 90% was achieved for low SNR signal.
[32]	MZIs	Photonic DNN for vowel recognition	F + S	Achieved 76.7% accuracy in vowel recognition. Lower accuracy attributed to limited resolution (24 bits).
[176]	MZIs	MNIST dataset classification using Structured Neural Network.	S	98.5% accuracy.
[133]	MZMs and MMIs	Analog feed-forward ANN with photonic MVM and MZM non-linearity demonstrated using a 2-by-1 vector dot-product experiment. Energy efficient binary multiplication demonstrated in simulation.	S	–
[143]	SOA-MZIs	RNN benchmarked using a finance forecasting application utilizing FI-2010 dataset	S	Gated optical RNN achieved an F1 score of 41.85%
[179]	AWGs and SOAs	Demonstration of precise 4-bit multiplication and accumulation operation	F + S	Error less than 0.2
[134]	AWG and MZIs	MNIST classification with CNN architecture. CNN implemented using Cooley-Tukey FFT algorithm, with AWGs used to implement DFT photonically.	S	Various noise sources (amplitude, phase and linear noises) and their combinations introduced to the CNN; 99.6% accuracy for 14280 parameter CNN.
[34]	Pillar silica scatteres	XOR computation and label-less classification of cancer cell images from healthy cells	S	20×20 node reservoir achieves symbol error rate below 5%.

[129]	Laser diodes	Label-less classification of cancer cell images from healthy cells	S	–
[181]	Diffractively coupled VCSELs	Demonstrated header recognition up to 5-bit headers.	S	–
[182]	SLM	Mackey-Glass chaotic time series prediction	S	Achieves an error of 0.013 for the prediction task.
[184]	SLM	Mackey-glass chaotic time series prediction	S	Reports NMSE below 0.3 for time series prediction task.
[188]	SOAs and AWGs	DNN implementation. Tested on Fisher's Iris classification.	F + S	Prediction accuracy of 85.8% achieved.
[189]	SOA and EAM	Experimental demonstration of photonic STDP and its utilization for supervised learning.	S	–
[190]	SOAs	Theoretical discussion and experimental demonstration of photonic STDP implementation using feedback signals. Demonstrated STDP used for reward based reinforcement-learning demonstration.	S	–
[123]	SOAs	Supervised and unsupervised pattern recognition. Demonstrated Hebbian algorithm for synaptic modification	S	–
[192]	VCSOA and VCSELs	SNN for learning and recognizing arbitrary spike patterns	S	–
[196]	VCSEL-SAs	SNN for learning and recognizing arbitrary spike patterns	S	–
[145]	VCSELs	SNN to simulate biological retinal neuronal circuitry. Simulated the ON and OFF stages of the retinal neuron circuitry.	S	–
[197]	SOA	Spoken digit recognition using RC.	S	The work reports a minimum Word Error Rate (WER) of 4.5% for their coherent SOA based reservoir
[170]	Passive photonic elements	Successful recognition of up to 5-bit headers and spoken digit recognition using RC.	S	Reports error rate "very close to" 0%
[198]	SNLs	Santa-Fe chaotic time series prediction	S	NMSE of 0.0359 obtained while the SNL is tuned to a feedback rate of $10ns^{-1}$

Используемые эталонные устройства. Созданные приложения (F) или смоделированные (S). Достигнутые результаты [154] Моделирование аттрактора Лоренца M.P. для сравнения с традиционной CTRNN на базе ЦП. F Сообщает о 294-кратном ускорении моделирования по сравнению с традиционным CTRNN на базе ЦП. [156] MR и MZM AlexNet CNN модель S утверждает, что скорость на 5 порядков выше, чем у полностью электрических реализаций. [157] Классификация MRs MNIST с использованием CNN. S Быстрее по сравнению с реализациями на базе графического процессора (в 2,8–1,4 раза быстрее) и в 0,75 раза больше энергопотребления. [132] MR и SOA. Вес загружается в банки MR с помощью массивов мемристоров. Различные тесты, включая MNIST, протестированы на фотонных CNN. S Снижение эксплуатационных расходов по сравнению с реализациями на базе графического процессора, с увеличением вычислительной эффективности до 25 раз. [37] MR и MZI классификация MNIST с использованием MLP. S Точность выше 95% достигается при разрешении 14 бит и настраиваемом MLP с 2048 нейронами в скрытом слое для обоих типов весов. Неотрицательные веса дают более низкую точность. [160] Микродиски Классификация изображений с помощью CNN. S В 13 раз лучшая производительность

на ватт, чем у ISAAC. [161] MR и MZI Классификация изображений с помощью CNN. S Вся оптическая конструкция потребляет всего 5,1% энергии, необходимой для всех электрических ускорителей, при этом на 31,9% быстрее. 31 [90] MR. Это было упражнение, чтобы доказать возможность создания SNN на основе черно-белых изображений. Прикладных экспериментов в данной работе не проводилось. _ _ [162] GST встроил классификацию MRs MNIST с MLP. S Точность 98,06%. [168] MRs Классификация цифровых шаблонов с высокой скоростью передачи битов с использованием RC. S Ошибка классификации 0,5% при 160 Гбит / с для цифровых слов длиной 8 бит. [169] MRs Демонстрация булевых операций с использованием RC. Детальный анализ операций XOR. S Продемонстрированы операции XOR с частотой ошибок 0,1. Также исследовалась взаимосвязь между частотой ошибок и модуляцией входной мощности и оптической расстройкой. [172] MZI. Математическое обсуждение управления фазой и амплитудой для представления унитарного оператора с использованием MZI. _ _ [173] MZIs Однослойная нейронная сеть, использующая описанный в работе оптический процессор 4x4, настроенный на классификацию выборок данных F + S. Показала точность классификации выборок данных 72%. [174] MZI. Математическое и теоретическое обсуждение представления унитарной матрицы на основе MZI и, следовательно, того, как можно сконструировать универсальное линейное устройство. Тестирование на основе приложений не проводилось. _ _ [175] MZI. Математическое и теоретическое обсуждение основанного на MZI представления унитарной матрицы с дополнительным обсуждением устойчивости к ошибкам и потерям такого устройства. _ _ [144] MZI. Двоичная нейронная сеть, настроенная на классификацию ближайшего соседа совокупности, сформированной из сигнала DP-QPSK 100 ГГц. S Точность классификации, близкая к 100%, достигается для сигнала с высоким ОСШ, а точность, близкая к 90%, достигается для сигнала с низким ОСШ. [32] MZIs Photonic DNN для распознавания гласных F + S Достигнута 76,7% точности распознавания гласных. Более низкая точность объясняется ограниченным разрешением (24 бита). [176] Классификация наборов данных MZI MNIST с использованием структурированной нейронной сети. S Точность 98,5%. [133] MZM и MMI Аналоговая ИНС с прямой связью с фотонной нелинейностью MVM и MZM, продемонстрированная с помощью эксперимента с векторным скалярным произведением 2 на 1. Энергоэффективное двоичное умножение продемонстрировано в моделировании. S _ [143] SOA-MZIs RNN, протестированная с помощью приложения финансового прогнозирования с использованием набора данных FI-2010. S Gated оптическая RNN получила оценку F1 41,85% [179] AWG и SOA Демонстрация точной операции 4-битного умножения и накопления F +

S Ошибка менее 0,2 [134] Классификация AWG и MZI MNIST с архитектурой CNN. CNN реализован с использованием алгоритма Кули-Тьюки FFT, с AWG, используемым для реализации DFT фотонно. • Различные источники шума (амплитудные, фазовые и линейные шумы) и их комбинации, вводимые в CNN; Точность 99,6% для 14280 параметров CNN. [34] Пиллярный кремнезем рассеивает вычисление XOR и безмаркированную классификацию изображений раковых клеток из здоровых клеток. S 20 × 20 узловой резервуар достигает коэффициента ошибок символа ниже 5%. 32 [129] Лазерные диоды Безмаркировочная классификация изображений раковых клеток из здоровых клеток S _ [181] Дифракционно связанные VCSEL Продемонстрированное распознавание заголовков до 5-битных заголовков. S _ [182] SLM Прогнозирование хаотических временных рядов по Макки-Глассу S Достигает ошибки 0,013 для задачи прогнозирования. [184] SLM Прогнозирование хаотических временных рядов по методу Макки S Сообщает NMSE ниже 0,3 для задачи прогнозирования временных рядов. [188] Реализация DNN SOA и AWG. Протестировано по классификации Fisher's Iris. F + S Достигнута точность прогноза 85,8%. [189] SOA и EAM Экспериментальная демонстрация фотонного STDP и его использования для обучения с учителем. S _ [190] SOA Теоретическое обсуждение и экспериментальная демонстрация реализации фотонного STDP с использованием сигналов обратной связи. Продемонстрированный протокол STDP, используемый для демонстрации обучения с подкреплением на основе вознаграждения. S _ [123] SOA Распознавание образов с учителем и без учителя. Продемонстрированный алгоритм Hebbian для синаптической модификации S _ [192] VCSOA и VCSELs SNN для обучения и распознавания произвольных паттернов спайков S _ [196] VCSEL-SAs SNN для обучения и распознавания произвольных паттернов спайков S _ [145] VCSELs SNN для имитации биологических нейронов сетчатки схема. Моделировали стадии включения и выключения нейронной схемы сетчатки. S _ [197] SOA Распознавание цифр с использованием RC. S В работе сообщается, что минимальная частота ошибок в словах (WER) составляет 4,5% для их когерентного резервуара на основе SOA [170]. Пассивные фотонные элементы. Успешное распознавание до 5-битных заголовков и голосовое распознавание цифр с использованием RC. S Сообщает о частоте ошибок, «очень близкой к» 0% [198] SNL Прогнозирование хаотического временного ряда в Санта-Фе S NMSE 0,0359, полученное, когда SNL настроен на скорость обратной связи 10 нс-1

6. ВЫЗОВЫ И ВОЗМОЖНОСТИ

Состояние – Кремниевые фотонные устройства art показали большие перспективы для создания искусственных нейронов. Архитектуры

глубокого обучения, построенные с использованием фотонных нейронов, поддерживают высокий уровень параллелизма при передаче и обработке весов за счет использования WDM, быстрого времени выполнения и низких затрат энергии. Однако существует несколько нерешенных проблем, связанных с эффективной реализацией различных нейронных функций с помощью кремниевых фотонных устройств, а также с целью достижения высокой надежности, масштабируемости и каскадности в реализациях архитектуры. Здесь мы суммируем проблемы и возможности для будущих исследований, необходимых для преодоления этих проблем.

• **Проблемы когерентности**: встроенные интерферометры (например, MZI) широко используются в архитектурах фотонных нейронных сетей из-за их способности эффективно представлять матрицы для операций нейронных сетей. Основные проблемы с MZI – это требования к большой площади и искажение фазового шума в сетках MZI. Из-за тепловых изменений и вариаций производственного процесса в MZI значения фазы могут отклоняться от своих целевых значений, что может повлиять на точность вывода нейронной сети, использующей их. Недавние исследования [200], [201] исследуют, как избежать этих проблем, учитывая эти проблемы на этапе обучения и настраивая фотонную нейронную сеть с учетом вариаций. Требуются дальнейшие исследования для более эффективного преодоления ограничений по площади и шуму этих когерентных архитектур.

• **Проблемы, связанные с некогерентностью**. Протокол широковещательной передачи (B&W) широко используется для реализации архитектур фотонных нейронных сетей. Некоторые усилия выявили возможные проблемы с этим протоколом, такие как гетеродинные перекрестные помехи, искажающие значения веса, и очень большое количествоMR необходимы для реализации более крупных сетей, особенно когда используется DWDM. Некоторые недавние попытки были направлены на устранение этих ограничений. Авторы в [37] предложили новую архитектуру, которая использует параллельное расположение MR вместо их каскадирования, хотя они используют неэффективные MZI для хранения векторов. Параллельное «безударное» расположение MR использовалось для уменьшения тепловых перекрестных помех между соседними MR и для достижения лучшего разрешения по весу. В других работах было предложено повторно использовать реализованные слои, например [156], а также методы для снижения энергопотребления и увеличения скорости работы за счет уменьшения использования электронных компонентов [160]. Также были предложения использовать микродиски вместо MR для дальнейшего увеличения плотности интеграции на чипах [161]. Но, как отмечено в [202], все варианты некогерентных

архитектур могут страдать от низкой пропускной способности, поскольку электронные компоненты, такие как память, могут не работать с такой высокой частотой, как их фотонные компоненты. Для преодоления этих ограничений некогерентных архитектур необходимы дальнейшие исследования.

MR необходимы для реализации более крупных сетей, особенно когда используется DWDM. Некоторые недавние попытки были направлены на устранение этих ограничений. Авторы в [37] предложили новую архитектуру, которая использует параллельное расположение MR, а не каскадирование их, хотя они используют неэффективные MZI для хранения векторов. Параллельное «безударное» расположение MR использовалось для уменьшения тепловых перекрестных помех между соседними MR и для достижения лучшего разрешения по весу. В других работах было предложено повторно использовать реализованные слои, например [156], а также методы для снижения энергопотребления и увеличения скорости работы за счет уменьшения использования электронных компонентов [160]. Также были предложения использовать микродиски вместо MR для дальнейшего увеличения плотности интеграции на чипах [161]. Но, как отмечено в [202], все варианты некогерентных архитектур могут страдать от низкой пропускной способности, поскольку электронные компоненты, такие как память, могут не работать с такой высокой частотой, как их фотонные компоненты. Для преодоления этих ограничений некогерентных архитектур необходимы дальнейшие исследования.

• **Вариации и надежность**: многие кремниевые фотонные устройства (например, MR, MZI) подвержены изменениям во времени разработки и времени выполнения. Процесс изготовления [203] и тепловые перекрестные помехи [89], а также старение устройства [204] могут значительно повлиять на надежность и производительность фотонных нейронных сетей, внося нежелательные перекрестные помехи, оптические фазовые сдвиги, резонансные дрейфы (например, в MR), накладные расходы на настройку и рассогласование тока фотодетектирования. Например, экспериментальные исследования показали, что резонансная длина волны в MC может сдвигаться на 4,79 нм внутри пластины из-за неизбежных изменений производственного процесса [205] и отклоняться на 0,1 нм / K [206] из-за тепловых изменений во время работы. Более того, кремниевые фотонные устройства по своей сути страдают от оптических потерь, которые ухудшают энергоэффективность, надежность и масштабируемость фотонных нейронных сетей [207]. Кроме того, конечная точность кодирования настроек фазы (например, в когерентных сетях) добавляет дополнительную неопределенность к значениям веса, полученным во время обучения сети, когда они отображаются на фазо-

вращатели как фазовые углы. Недавнее исследование [153] о влиянии неопределенностей – из-за процесса изготовления и тепловых изменений – в фотонных нейронных сетях показывает значительное снижение точности вывода фотонных нейронных сетей на 70%. Следовательно, необходимы дальнейшие исследования для повышения надежности кремниевых фотонных устройств.

• **Мощность и энергия**: нейроны O / E / O потребляют много энергии, потому что электрооптические и оптоэлектрические преобразования потребляют значительную мощность. Более того, O / E / O требует преобразования длины волны для реализации крупномасштабной нейронной сети, которая также потребляет дополнительную мощность. Следовательно, O / E / O может быть не лучшим выбором для достижения высокой энергоэффективности. Полностью оптические нейроны могут достичь большей энергоэффективности, но за счет более низкой скорости работы (что может увеличить потребление энергии) и меньшей каскадируемости (что затрудняет реализацию сложных функций). Внешние лазеры потребляют значительную часть общей мощности в фотонных нейронных сетях. Хотя такие лазеры менее подвержены тепловым изменениям, чем лазеры на кристалле, они несут дополнительные потери оптической мощности из-за необходимости соединять внешний источник света с устройствами на кристалле через соединительную структуру (например, решетчатые элементы связи). Более того, чтобы справиться с вариациями (как обсуждалось в предыдущем пункте), требуются накладные расходы на мощность и энергию для достижения надежности за счет пространственной, временной или информационной избыточности. Поскольку мощность является таким значительным ограничением при проектировании современных вычислительных чипов, существует острая необходимость в новых исследованиях для достижения энергоэффективных реализаций фотонных нейронных сетей.

• **Электронные контроллеры**. Дизайн фотонной нейронной сети был бы нереалистичным без учета проблем, связанных с электронным контроллером. Для фотонных нейронных сетей требуется электронный контроллер для управления (например, настройки и управления) и оркестровки фотонных устройств в сети (например, настройки MR и контролируемого обучения). Более того, контроллер должен обнаруживать и уменьшать смещение времени выполнения (например, из-за тепловых перекрестных помех) и поддерживать правильную работу оптических нейронов. Однако электронные контроллеры требуют больших задержек, и между электронным контроллером и оптической сетью существует несоответствие частот. Следовательно, необходимы дополнительные

исследования в отношении реализации высокоскоростных электронных контроллеров для фотонных нейронных сетей.

• **Обучение обратному распространению**: почти все архитектуры фотонных нейронных сетей в предыдущей работе сосредоточены на ускорении вывода. Существует потребность в изучении фотонных архитектур, которые могут эффективно поддерживать обучение нейронной сети. Это особенно сложно, потому что обучение (например, через обратное распространение) требует обратного потока информации от выходных слоев к входным слоям, что потребует дополнительных волноводов, сигналов и компонентов обработки для расчета градиентов и обновления значений весов. Некоторые недавние попытки начали исследовать такие архитектуры, например, [208], который предложил гибридный ускоритель на основе мемристора и фотоники, который также поддерживает обратное распространение. Необходимы дополнительные исследования для разработки поддержки обратного распространения с низким уровнем служебных данных с помощью фотоники.

• **Разрешение.** Разрешение по весу играет решающую роль в архитектурах ускорителей глубокого обучения. Для ускорения вывода желательно иметь более высокое разрешение для большей точности. Большинство предшествующих работ по фотонным нейронным сетям достигают очень низкого разрешения, например, работа [157], в которой достигается разрешение 6-7 бит, и работа [37], которая достигает разрешения по весу 14 бит. Некоторые предлагаемые архитектуры решают проблему более низкого разрешения за счет разделения весового представления между несколькими устройствами, например [160], или за счет использования побитового распараллеливания операций весовой матрицы, как в [161], для достижения разрешения весов в 16 бит. В работе [32] удалось достичь разрешения по весу 24 бита с использованием сеток MZI, но масштабируемость такой архитектуры сомнительна из-за большой площади, занимаемой MZI. В работе [32] не удалось достичь разрешения выше 24 бит из-за шума фазового кодирования в фазовращателях MZI. Основные проблемы в достижении хорошего разрешения в фотонных архитектурах связаны с перекрестными помехами, чувствительностью фотодетектора и шумом фотодетектора (дробовым шумом). Хотя в работе [37] представлен подробный анализ того, как тепловые перекрестные помехи влияют на фотонную чувствительность к значениям веса, даже межканальные и внутриканальные перекрестные помехи могут повлиять на достижимое разрешение. Таким образом, необходимы исследования для достижения эффективного подавления фотонных перекрестных помех, коррекции фазового шума и шумоустойчивого фотодетектирования для достижения лучшего разрешения в ускорителях фотонного глубокого обучения.

• **Масштабируемость:** многие из работ, обсуждаемых в этом обзоре, были сосредоточены на реализации небольших нейронных сетей [32], [37], [90], [162], чтобы подчеркнуть эффективность кремниевого фотонного ускорения. Другие работы сосредоточены на ускорении умножения векторов матриц и их повторном использовании на нескольких уровнях модели глубокого обучения, например, в [157], [160] и [161]. Основная проблема, с которой сталкиваются при реализации крупномасштабных сетей с использованием кремниевых фотонных устройств, – это потребление площади, учитывая, что основные компоненты в архитектуре фотонной нейронной сети могут достигать микрометров. Кроме того, потери, связанные с распространением и перекрестными помехами, накапливаются в более крупных архитектурах, включающих очень большое количество устройств, и потребление энергии может достигать очень высоких значений [156]. Сетки MZI, такие как представленные в [32], [174] и [179], сталкиваются с серьезными проблемами, связанными с потреблением площади (MZI намного больше, чем MR или микродиски) и фазовым шумом, что ограничивает их масштабируемость. Чтобы уменьшить проблемы масштабируемости, в некоторых работах рассматривается упрощенная версия модели нейронной сети в оборудовании с использованием методов регрессии [176] и эффективного вычисления свертки матриц с использованием методов БПФ [134], [177]. Чтобы реализовать масштабируемые конструкции фотонных ускорителей, необходимы исследования 1) новых подходов к сжатию моделей для снижения сложности кремниевого фотонного оборудования и 2) шумоустойчивых, компактных кремниевых фотонных устройств с низкими потерями, которые могут поддерживать высокую каскадность для реализации больших нейронных сетей.

REFERENCES

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006

[2] V. K. Kukkala, J. Tunnell, S. Pasricha, and T. Bradley, “Advanced driver-assistance systems: A path toward autonomous vehicles,” IEEE Consumer Electronics Magazine, vol. 7, no. 5, pp. 18–25, 2018

[3] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” The International Journal of Robotics Research, vol. 37, no. 4-5, pp. 421–436, 2018

[4] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media using geometric deep learning,” arXiv:1902.06673 [cs], Feb. 2019

[5] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for covid-19 (sars-cov2) pandemic: A review,” Chaos, Solitons & Fractals, p. 110059, 2020

[6] K. Kukkala, S. V. Thiruloga, and S. Pasricha, “Indra: Intrusion detection using recurrent autoencoders in automotive embedded systems,” arXiv preprint arXiv:2007.08795, 2020

[7] J. Gu, G. Neubig, K. Cho, and V. O. Li, “Learning to translate in real-time with neural machine translation,” arXiv preprint arXiv:1610.00388, 2016

[8] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” The bulletin of mathematical biophysics,

vol. 5, no. 4, pp. 115–133, 1943

[9] F. Rosenblatt, “The perceptron, a perceiving and recognizing automaton”, Report 85-460-1, Cornell Aeronautical Laboratory, 1957

[10] A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamuraet al.,“A million spiking-neuron integrated circuit with a scalable communication network and interface,” Science, vol. 345, no. 6197, pp.668–673, 2014

[11] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jainet al., “Loihi: A neuromorphic manycore processor with on-chip learning,” IEEE Micro, vol. 38, no. 1,pp. 82–99, 2018

[12] P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borcherset al., “In-datacenter performance analysis of a tensor processing unit,” Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017

[13] V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello, “A 240 G-ops/s Mobile Coprocessor for Deep Neural Networks,” CVPR Workshop, 2014

[14] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, “ShiDianNao: Shifting Vision Processing Closer to the Sensor,” International Symposium on Computer Architecture (ISCA), 2015

[15] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, “Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks,” FPGA, 2015

[16] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze, “Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices,” IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), vol. 9, no. 2, pp. 292-308, June 2019

[17] A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, “SCNN: An accelerator for compressed-sparse convolutional neural networks,” International Symposium on Computer Architecture (ISCA), 2017

[18] S. Markidis, S. W. Der Chien, E. Laure, I. B. Peng, and J. S. Vetter, “Nvidia tensor core programmability, performance & precision,” pp. 522–531, 2018

[19] A. Sodani, R. Gramunt, J. Corbal, H.-S. Kim, K. Vinod, S. Chinthamani, S. Hutsell, R. Agarwal, and Y.-C. Liu, “Knights landing:

Second-generation intel xeon phi product," IEEE micro, vol. 36, no. 2, pp. 34–46, 2016

[20] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P.Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016

[21] L. Song, X. Qian, H. Li, and Y. Chen, "Pipelayer: A pipelined reram-based accelerator for deep learning," 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 541–552, 2017

[22] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, "Recent progress in analog memory-based accelerators for deep learning," Journal of Physics D: Applied Physics, vol. 51, no. 28, p. 283001, 2018

[23] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, "A 55nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots," pp. 124–126, 2018

[24] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement," pp. 141–142,2018

[25] A. Amaravati, S. B. Nasir, J. Ting, I. Yoon, and A. Raychowdhury, "A55-nm, 1.0–0.4 v, 1.25-pj/mac time-domain mixed-signal neuromorphic accelerator with stochastic synapses for reinforcement learning in autonomous mobile robots," IEEE Journal of SolidState Circuits, vol. 54,no. 1, pp. 75–87, 2018

[26] M. M. Waldrop, "The chips are down for Moore's law," Nature News, vol.530, no. 7589, p. 144, 2016 [27] S. Pasricha, and N. Dutt. "On-Chip Communication Architectures", Morgan Kauffman, ISBN 978-0-12-373892-9, Apr 2008

[28] S. Pasricha, M. Nikdast, "A Survey of Silicon Photonics for Energy Efficient Manycore Computing" to appear, IEEE Design and Test, 2020

[29] L. Chrostowski, H. Shoman, M. Hammood, H. Yun, J. Jhoja, E. Luan,S. Lin, A. Mistry, D. Witt, N. A. Jaegeret al., "Silicon photonic circuit design using rapid prototyping foundry process design kits," IEEE J. Sel. Top. Quantum Electron., vol. 25, no. 5, pp. 1–26, 2019

[30] D. A. Miller, "Silicon photonics: Meshing optics with applications," Nature Photonics, vol. 11, no. 7, pp. 403–404, 2017

[31] A. R. Totovi´c, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, "Femtojoule per mac neuromorphic photonics: An energy and technology roadmap," IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 5,pp. 1–15, 2020

[32] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englundet al., "Deep learning with coherent nanophotonic circuits," Nature Photonics, vol. 11, no. 7, p. 441, 2017

[33] G. Van der Sande, D. Brunner, and M. C. Soriano, "Advances in photonic reservoir computing," Nanophotonics, vol. 6, no. 3, pp. 561–576, 2017

[34] A. Katumba, M. Freiberger, F. Laporte, A. Lugnan, S. Sackesyn, C. Ma,J. Dambre, and P. Bienstman, "Neuromorphic computing based on silicon photonics

and reservoir computing," IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 6, pp. 1–10, 2018

[35] G. Tanaka, T. Yamane, J. B. H´eroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, "Recent advances in physical reservoir computing: A review," Neural Networks, vol. 115, pp. 100–123, 2019

[36] A. N. Tait, M. A. Nahmias, Y. Tian, B. J. Shastri, and P. R. Prucnal, "Photonic neuromorphic signal processing and computing," in Nanophotonic Information Physics. Berlin, Germany: Springer, 2014

[37] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," Proceedings of the IEEE, 2020

[38] T. F. de Lima, H. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, "Machine learning with neuromorphic photonics," J. Lightw. Technol., vol. 37, no. 5, pp. 1515_1534, Mar. 1, 2019

[39] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: A survey," IEEE Access, vol. 7, pp. 175 827–175 841,2019

[40] A. N. Tait, A. X. Wu, T. F. De Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Microring weight banks," IEEE J. Sel.Top. Quantum Electron., vol. 22, no. 6, pp. 312–325, 2016

[41] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," Journal of Comparative Neurology, vol. 513, no. 5, pp. 532–541, 2009

[42] T. Takeuchi, A. J. Duszkiewicz, and R. G. Morris, "The synaptic plasticity and memory hypothesis: encoding, storage and persistence," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 369, no. 1633, p. 20130288, 2014

[43] A.L Hodgkin, A.F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve", The Journal of Physiology, Aug 1952

[44] A. Borisyuk, "Morris–lecar model," Encyclopedia of Computational Neuroscience. Springer, 2015

[45] S. Binczak, S. Jacquir, J.-M. Bilbault, V. B. Kazantsev, and V. I. Nekorkin, "Experimental study of electrical fitzhugh–nagumo neurons with modified excitability," Neural Networks, vol. 19, no. 5, pp. 684– 693, 2006

[46] M. Hayati, M. Nouri, D. Abbott, and S. Haghiri, "Digital multiplierless realization of two-coupled biological hindmarsh–rose neuron model," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 63, no. 5, pp. 463–467, May 2016

[47] L.F. Abbott, "Lapique's introduction of the integrate-and-fire model neuron (1907)" (PDF). Brain Research Bulletin, May 1999

[48] W. Maass. "Networks of spiking neurons: The third generation of neural network models", Neural Networks 10, 1997

[49] E. M. Izhikevich, "Simple model of spiking neurons," IEEE Transactions on neural networks, vol. 14, no. 6, pp. 1569–1572, 2003

[50] C. A. Runyan, E. Piasini, S. Panzeri, and C. D. Harvey, “Distinct time-scales of population coding across cortex,” Nature, vol. 548, pp.92–96,Jul. 2017

[51] J. Hines, “Stepping up to summit,” Comput. Sci. Eng., vol. 20, no. 2, pp. 78–82, 2018

[52] R. J. Douglas and K. A. C. Martin, “Recurrent neuronal circuits in the neocortex,” Current Biol., vol. 17, no. 13, pp. 496–500, 2004

[53] J. Hasler and B. Marr, “Finding a roadmap to achieve large neuromorphic hardware systems,” Front. Neurosci., vol. 7, Sep. 2013

[54] C.Mead, “Neuromorphic electronic systems,” Proc. IEEE, vol. 78, no. 10, pp. 1629–1636, Oct. 1990

[55] D. Tank and J. J. Hopfield, “Simple ‘neural’ optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit,” IEEE Trans. Circuits Syst., vol. 33, no. 5, pp. 533–541, May 1986

[56] I. Sourikopoulos, S. Hedayat, C. Loyez, F. Danneville, V. Hoel, E. Mercier, and A. Cappy, “A 4-fj/spike artificial neuron in 65 nm CMOS technology,” Frontiers in neuroscience, vol. 11, p. 123, 2017

[57] J. Shi, S. D. Ha, Y. Zhou, F. Schoofs, and S. Ramanathan, “A correlated nickelate synaptic transistor,” Nature Commun., vol. 4, Oct. 2013

[58] W. Xu, S. Y. Min, H. Hwang, and T. W. Lee, “Organic core-sheath nanowire artificial synapses with femtojoule energy consumption,”Sci. Advances, vol. 2, no. 6, Jun. 2016

[59] J. Zhu, Y. Yang, R. Jia, Z. Liang, W. Zhu, Z. U. Rehman, L. Bao,X. Zhang, Y. Cai, L. Songet al., “Ion gated synaptic transistors based on 2d van der waals crystals with tunable diffusive dynamics,” Advanced Materials, vol. 30, no. 21, p. 1800195, 2018

[60] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” Nature, vol. 521, no. 7550,pp. 61–64, 2015

[61] S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B. H. Lee, H. Hwang, B. Lee, and B.-g. Lee, “Electronic system with memristive synapses for pattern recognition,” Scientific reports, vol. 5, p. 10123, 2015

[62] I. Boybat, M. Le Gallo, S. Nandakumar, T. Moraitis, T. Parnell, T. Tuma,B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, “Neuro-morphic computing with multi-memristive synapses,” Nature communications, vol. 9, no. 1, pp. 1–12, 2018

[63] S. Hu, G. Qiao, Y. Liu, L. Rong, Q. Yu, and Y. Liu, “An improved memristor model connecting plastic synapse and nonlinear spiking neuron,” Journal of Physics D: Applied Physics, vol. 52, no. 27, p. 275402, 2019

[64] X. Jin, S. B. Furber, and J. V. Woods, “Efficient modelling of spiking neural networks on a scalable chip multiprocessor,” pp. 2812–2819, 2008

[65] Introducing a Brain-inspired Computer, [Online]. Available: https://www.research.ibm.com/articles/brain-chip.shtml.

[66] Beyond Today’s AI,

[Online]. Available:

ttps://www.intel.com/content/www/us/en/research/neuromorphic-computing.html.

[67] S. Moore, “Intels neuromorphic system hits 8 million neurons, 100million coming by 2020,” IEEE Spectrum, vol. 15, 2019

[68] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” nature, vol. 323, no. 6088, pp.533–536, 1986

[69] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger, “Neural acceleration for general-purpose approximate programs,” pp. 449–460, 2012

[70] NVIDIA Corp., NVIDIA A100 Tensor Core GPU Architecture, Whitepaper, 2020

[71] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, “Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory,” ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 27–39, 2016

[72] D. Verstraeten, S. Xavier-de-Souza, B. Schrauwen, J. A. K. Suykens, D. Stroobandt, and J. Vandewalle, “Pattern classification with CNNs as reservoirs,” Proc. Int. Symp. Nonlin. Theory Appl., Budapest, Hungary, pp. 101–104, 2008

[73] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, “Optoelectronic reservoir computing,” Scientific reports, vol. 2, p. 287, 2012

[74] R. Martinenghi, S. Rybalko, M. Jacquot, Y. K. Chembo, and L. Larger, “Photonic nonlinear transient computing with multiple-delay wavelength dynamics,” Phys. Rev. Lett., vol. 108, Jun. 2012

[75] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Guti´errez, L. Pesquera, C. R. Mirasso, and I. Fischer, “Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing,” Optics express, vol. 20, no. 3, pp. 3241–3249, 2012

[76] F. Duport, B. Schneider,A. Smerieri, M. Haelterman, and S.Massar, “All optical reservoir computing,” Opt. Express, vol. 20, no. 20, pp.22783–22795, Sep. 2012

[77] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, “Parallel photonic information processing at gigabyte per second data rates using transient states,” Nature Commun., vol. 4, 2013

[78] K. Hicke, M. A. Escalona-Mor´an, D. Brunner, M. C. Soriano, I. Fischer, and C. R. Mirasso, “Information processing using transient dynamics of semiconductor lasers subject to delayed feedback,” IEEE J. Sel. Top. Quantum Electron., vol. 19, no. 4, pp. 1 501 610–1 501 610, 2013

[79] M. C. Soriano, S. Ort´ın, D. Brunner, L. Larger, C. R. Mirasso, I. Fischer, and L. Pesquera, “Optoelectronic reservoir computing: tackling noise-induced performance degradation,” Optics express, vol. 21, no. 1, pp.12–20, 2013

[80] S. Ortin, M. C. Soriano, L. Pesquera, D. Brunner, D. San-Martin, I. Fischer, C. Mirasso, and J. Gutierrez, “A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron,” Scientific reports, vol. 5, p. 14945, 2015

[81] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, “Fully analogue photonic reservoir computer,” Sci. Rep., vol. 6, 2016

[82] M. Nikdast, G. Nicolescu, J. Trajkovic, and O. Liboiron-Ladouceur, “Deeper: Enhancing performance and reliability in chip-scale optical interconnection networks,” Proceedings of the 2018 on Great Lakes Symposium on VLSI, pp. 63–68, 2018

[83] A. N. Tait et al., “Silicon photonic modulator neuron,” Physical Review Applied, vol. 11, no. 6, p. 064043, June 2019

[84] P. R. Prucnal, B. J. Shastri, T. F. de Lima, M. A. Nahmias, and A. N. Tait, “Recent progress in semiconductor excitable lasers for photonic spike processing”, Adv. Opt. Photonics, vol. 8, no. 2, pp. 228–299, Jun. 2016

[85] T. F. de Lima et al., “Noise analysis of photonic modulator neurons,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, pp. 1–9, JanFeb 2020

[86] D. Liang and J. E. Bowers, “Recent progress in lasers on silicon,” Nat. Photonics, vol. 4, no. 8, Art. no. 8, Aug. 2010

[87] N. H. Zhu et al., “Directly modulated semiconductor lasers,” IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 1,pp. 1–19, Jan-Feb 2017

[88] D. Vantrease et al., “Corona: System implications of emerging nanophotonic technology,” ACM SIGARCH Computer Architecture News, vol. 36, no. 3, pp. 153–164, 2008

[89] A. Mirza, S. M. Avari, E. Taheri, S. Pasricha, and M. Nikdast, “Opportunities for Cross-Layer Design in High-Performance Computing Systems with Integrated Silicon Photonic Networks,” 2020 Design, Automation Test in Europe Conference Exhibition (DATE), 2020

[90] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Broadcast and Weight: An Integrated Network For Scalable Photonic Spike Processing,” JLT, vol. 32, no. 21, pp. 4029–4041, 2014

[91] S. Xiang, Y. Zhang, X. Guo, A. Wen, and Y. Hao, “Photonic generation of neuron-like dynamics using vcsels subject to double polarized optical injection,” Journal of Lightwave Technology, vol. 36, no. 19, pp. 4227–4234, 2018

[92] Z. W. Song, S. Y. Xiang, Z. X. Ren, S. H. Wang, A. J. Wen, and Y. Hao, “Photonic spiking neural network based on excitable vcselssa for sound azimuth detection,” Optics Express, vol. 28, no. 2, pp. 1561–1573, 2020

[93] I. Aldaya, C. Gosset, C. Wang, G. Campuzano, F. Grillot, and G. Castanon, “Periodic and aperiodic pulse generation using optically injected dfb laser,” Electronics Letters, vol. 51, no. 3, pp. 280–282, 2015

[94] G. Sarantoglou, M. Skontranis, and C. Mesaritakis, “All optical integrate and fire neuromorphic node based on single section quantum dot laser,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 5,pp. 1–10, 2019

[95] F. Koyama, “Recent Advances of VCSEL Photonics,” JLT, vol. 24, no. 12, pp. 4502–4513, 2006

[96] J. Van Campenhout, P. Rojo-Romeo, P. Regreny, C. Seassal, D. Van Thourhout, S. Verstuyft, L. Di Cioccio, J.-M. Fedeli, C. Lagahe, and R. Baets, "Electrically pumped InP-based microdisk lasers integrated with a nanophotonic silicon-on-insulator waveguide circuit," Optics express, vol. 15, 2007

[97] J. Robertson, E. Wade, and A. Hurtado, “Electrically controlled neuron-like spiking regimes in vertical-cavity surface-emitting lasers at ultrafast rates,” IEEE J. Sel. Top. Quantum Electron., vol. 25,no. 6, pp. 1–7, 2019

[98] J. Robertson, M. Hejda, J. Bueno, and A. Hurtado, “Ultrafast optical integration and pattern classification for neuromorphic photonics based on spiking vcsel neurons,” Scientific reports, vol. 10, no. 1, pp. 1–8, 2020

[99] A. Hurtado, K. Schires, I. Henning, and M. Adams, “Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems,” Appl. Phys. Lett, vol. 100, no. 10, p. 103703, 2012

[100] A. Levi, “Microdisk lasers,” Solid-state electronics, vol. 37, no. 4-6, pp.1297–1302, 1994

[101] L. Mahler, A. Tredicucci, F. Beltram, C. Walther, J. Faist, B. Witzigmann, H. E. Beere, and D. A. Ritchie, “Vertically emitting microdisk lasers,” Nature Photonics, vol. 3, no. 1, pp. 46–49, 2009

[102] Yisu Yang, Gligor Djogo, Moez Haque, Peter R. Herman, and Joyce K. S. Poon, "Integration of an O-band VCSEL on silicon photonics with polarization maintenance and waveguide coupling," Opt. Express 25, 2017

[103] K. Alexander, T. Van Vaerenbergh, M. Fiers, P. Mechet, J. Dambre, and P. Bienstman, “Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response”, Opt. Express, vol. 21, no. 22, pp. 26182–26191, 2013

[104] M. Tran, D. Huang, T. Komljenovic, J. Peters, A. Malik, and J. Bowers, “Ultra-Low-Loss Silicon Waveguides for Heterogeneously Integrated Silicon/III-V Photonics”, Appl. Sci., vol. 8, no. 7, pp. 1139-, 2018

[105] S. Nambiar, S. Purnima, and S. K. Selvaraja, "Grating-assisted fiber to chip coupling for SOI photonic circuits," Applied Sciences, vol. 8, 2018

[106] D. F. Siriani and K. D. Choquette, “Coherent Coupling of Vertical-Cavity Surface-Emitting Laser Arrays”, Semiconductors and Semimetals, vol. 86, pp. 226-264, 2012

[107] S. Maktoobi et al., “Diffractive Coupling for Photonic Networks: How Big Can We Go?”, IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, pp. 1–8, Jan. 2020

[108] X. Wu et al., “UNION: A Unified Inter/Intrachip Optical Network for Chip Multiprocessors”, IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 22, no. 5, pp. 1082–1095, May 2014

[109] H. Shabani, A. Roohi, A. Reza, M. Reshadi, N. Bagherzadeh, and R. F. DeMara, “Loss-Aware Switch Design and Non-Blocking Detection Algorithm for Intra-Chip Scale Photonic Interconnection Networks”, IEEE Trans. Comput., vol. 65, no. 6, pp. 1789–1801, Jun. 2016

[110] A. N. Tait, T. F. de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Microring Weight Banks for Neuromorphic Silicon Photonics”, 2018 Conference on Lasers and Electro-Optics (CLEO), May 2018

[111] A. N. Tait, T. F. de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Multi-channel control for microring weight banks”, Opt Express, vol. 24, no. 8, pp. 8895–8906, 2016

[112] T. V. Vaerenbergh et al., “Cascadable excitability in microrings,” Opt. Express, vol. 20, no. 18, pp. 20292–20308, Aug. 2012

[113] J. Xiang, A. Torchy, X. Guo, and Y. Su, “All-Optical Spiking Neuron Based on Passive Microresonator”, JLT, pp. 1–1, 2020

[114] Z. Ying et al., “Comparison of microrings and microdisks for high-speed optical modulation in silicon photonics,” Appl. Phys. Lett., vol.112, no. 11, p. 111108, Mar. 2018

[115] Z. Yu, J. Zheng, P. Xu, W. Zhang, and Y. Wu, “Ultracompact Electro-Optical Modulator-Based Ge2Sb2Te5 on Silicon,” IEEE Photonics Technol. Lett., vol. 30, no. 3, pp. 250–253, Feb. 2018

[116] P. Xu, J. Zheng, J. Doylend, and A. Majumdar, “Non-Volatile Integrated-Silicon-Photonic Switches using Phase-Change Materials”, 2019 Asia Communications and Photonics Conference (ACP), Nov. 2019

[117] N. Dhingra, J. Song, G. J. Saxena, E. K. Sharma, and B. M. A. Rahman, “Design of a Compact Low-Loss Phase Shifter Based on Optical Phase Change Material”, IEEE Photonics Technol. Lett., vol. 31, no. 21, pp. 1757–1760, Nov. 2019

[118] Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, “On-chip photonic synapse,” Sci. Adv., vol. 3, no. 9, p. e1700160, Sep. 2017

[119] C. D. Wright, Y. Liu, K. I. Kohary, M. M. Aziz, and R. J. Hicken, “Arithmetic and Biologically-Inspired Computing Using Phase-Change Materials”, Adv. Mater., vol. 23, no. 30, pp. 3408–3413, 2011

[120] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, “All-optical spiking neurosynaptic networks with selflearning capabilities”, Nature, vol. 569, no. 7755, Art. no. 7755, May 2019

[121] S. Kim et al., “NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning”, 2015 IEEE International Electron Devices Meeting (IEDM), p. 17.1.1-17.1.4, Dec. 2015

[122] M. P. Fok, Y. Tian, D. Rosenbluth, and P. R. Prucnal, “Pulse lead/lag timing detection for adaptive feedback and control based on optical spike-timing-dependent plasticity,” Opt. Lett., vol. 38, no. 4, pp. 419–421, 2013

[123] R. Toole et al., “Photonic Implementation of Spike-Timing-Dependent Plasticity and Learning Algorithms of Biological Neural Systems”, JLT, vol. 34, no. 2, pp. 470–476, Jan. 2016

[124] F. Marino and S. Balle, “Experimental study of a broad area vertical-cavity semiconductor optical amplifier”, Opt. Commun., vol. 231, no. 1, pp. 325–330, Feb. 2004

[125] S. Xiang et al., “Numerical Implementation of Wavelength-Dependent Photonic Spike Timing Dependent Plasticity Based on VCSOA”, IEEE J. Quantum Electron., vol. 54, no. 6, pp. 1–7, Dec. 2018

[126] S. R. Restaino, “Introduction to Liquid Crystals for Optical Design and Engineering”, 2015, PDF ISBN: 9781628416619, Print ISBN:9781628418071

[127] R. Bruck et al., “All-optical spatial light modulator for reconfigurable silicon photonic circuits,” Optica, vol. 3, no. 4, pp. 396–402, Apr. 2016

[128] A. Lugnan et al., “Photonic neuromorphic information processing and reservoir computing,” APL Photonics, vol. 5, no. 2, p. 020901, Feb. 2020

[129] A. Lugnan, J. Dambre, and P. Bienstman, “Integrated pillar scatterers for speeding up classification of cell holograms,” Opt. Express, vol. 25, no. 24, pp. 30526–30538, Nov. 2017

[130] P. Li et al, “All-optical Analog Comparator”, Nature Comm., 2016

[131] Aikawa, Yohei. "Ultracompact optical comparator for 4-bit QPSK-modulated signal based on silicon photonic waveguide." IEEE Photonics Journal VOL. 11, no. 3, pp. 1-10, 2019

[132] D. Dang, J. Dass, R. Mahapatra, “ConvLight: A Convolutional Accelerator with Memristor integrated Photonic Computing”, IEEE 24th International Conference on High Performance Computing (HiPC), 2017

[133] M. B. On, H. Lu, H. Chen, R. Proietti and S. J. Ben Yoo, "Wavelength-Space Domain High-Throughput Artificial Neural Networks by Parallel Photoelectric Matrix Multiplier," Optical Fiber Communications Conference and Exhibition (OFC), 2020

[134] J. R. Ong, C. C. Ooi, T. Y. L. Ang, S. T. Lim and C. E. Png, "Photonic Convolutional Neural Networks Using Integrated Diffractive Optics," IEEE J. Sel. Top. Quantum Electron, vol. 26, no. 5, pp. 1-8, Sept.-Oct. 2020

[135] S. Y. Xiang et al., “Cascadable Neuron-Like Spiking Dynamics in Coupled VCSELs Subject to Orthogonally Polarized Optical Pulse Injection”, IEEE J. Sel. Top. Quantum Electron., vol. 23, no. 6, pp. 1–7, 2017

[136] Z. Zhang, Z. Wu, D. Lu, G. Xia, and T. Deng, “Controllable spiking dynamics in cascaded VCSEL-SA photonic neurons”, Nonlinear Dyn., vol. 99, no. 2, pp. 1103–1114, 2019

[137] X. Zhuge, J. Wang, and F. Zhuge, “Photonic Synapses for Ultrahigh-Speed Neuromorphic Computing”, Phys. Status Solidi RRL –Rapid Res. Lett., vol. 13, no. 9, p. 1900082, 2019

[138] U. H. Lodish et al., “Molecular Cell Biology”, Macmillan, 2008.

[139] J. Robertson, T. Deng, J. Javaloyes, and A. Hurtado, “Controlled inhibition of spiking dynamics in VCSELs for neuromorphic photonics: theory and experiments”, Opt. Lett., vol. 42, no. 8, pp. 1560–1563, 2017

[140] A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, “Demonstration of WDM weighted addition for principal component analysis”, Opt. Express, vol. 23, no. 10, pp. 12758–12765, 2015

[141] G. M. Alexandris et al., “Neuromorphic photonics with coherent linear neurons using dual-IQ modulation cells”, J. Lightw. Technol., vol. 38, no. 4, pp. 811–819, Feb. 2020

[142] F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, “Temporal summation in a neuromimetic micropillar laser”, Opt. Lett., vol. 40, no. 23, pp. 5690–5693, Dec. 2015

[143] G. M. Alexandris et al., "All-Optical WDM Recurrent Neural Networks With Gating," IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 5, pp. 1-7, Sept.-Oct. 2020

[144][151] B. J. Shastri, A. N. Tait, T. F. de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, “Principles of Neuromorphic Photonics”, ArXiv180100016 Phys., pp. 1–37, 2018

[152] Y. Zhao, D. Lombardo, J. Mathews, and I. Agha, “Low control-power wavelength conversion on a silicon chip”, Opt. Lett., vol. 41, no. 15, pp. 3651–3654, Aug. 2016

[153] S. Banerjee, M. Nikdast, and K. Chakrabarty, “Modeling Silicon-Photonic Neural Networks under Uncertainties,” IEEE/ACM Design, Automation and Test in Europe (DATE) Conference and Exhibition, 2021

[154] A. N. Tait et al., “Neuromorphic photonic networks using silicon photonic weight banks”, Sci. Rep., vol. 7, no. 1, Art. no. 1, Aug. 2017

[155] E. N. Lorenz, “Deterministic nonperiodic flow”, Journal of Atmospheric Sciences, 1963

[156] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, “PCNNA: A Photonic Convolutional Neural Network Accelerator”, 31st IEEE International System-on-Chip Conference (SOCC), 2018

[157] V. Bangari et al., “Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)”, IEEE J. Sel. Top. Quantum Electron., Volume: 26, Issue: 1, Jan.-Feb. 2020

[158] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, 1998

[159] C. Zhang, Zhenman Fang, Peipei Zhou, Peichen Pan and Jason Cong, "Caffeine: Towards uniformed representation and acceleration for deep convolutional neural networks," ICCAD, 2016

[160] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, L. Jiang, “HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers”, Design, Automation Test in Europe Conference Exhibition (DATE), 2019

[161] K. Shiflett, D. Wright, A. Karanth, and A. Louri, “PIXEL: Photonic Neural Network Accelerator”, IEEE HPCA 2020

[162] I. Chakraborty, G. Saha, A. Sengupta and K. Roy, “Toward Fast Neural Computing using All-Photonic Phase Change Spiking Neurons”, Nature, August 2018

[163] M. Wuttig, N. Yamada, “Phase-change materials for rewriteable data storage”, Nat. Mater. 6, 824–832, 2007

[164] S. R. Ovshinsky, Reversible electrical switching phenomena in disordered structures”, Phys. Rev. Lett. 21, 1450–1453, 1968

[165] W. H. P. Pernice, H. Bhaskaran, “Photonic non-volatile memories using phase change materials”, Appl. Phys. Lett. 101, 171101, 2012

[166] C. Rios, P. Hosseini, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, “On-chip photonic memory elements employing phase-change materials”, Adv. Mater. 26, 1372–1377, 2014

[167] T. Van Vaerenbergh, M. Fiers, P. Bienstman and J. Dambre, "Towards integrated optical spiking neural networks: Delaying spikes on chip," 2013 Sixth "Rio De La Plata" Workshop on Laser Dynamics and Nonlinear Photonics, Montevideo, 2013

[168] C. Mesaritakis, V. Papataxiarhis, and D. Syvridis, “Micro ring resonators as building blocks for an all-optical high-speed reservoircomputing bit-pattern-recognition system,” J. Opt. Soc. Amer. B, vol. 30, no. 11, pp. 3048–3055, Nov. 2013

[169] F. D. Coarer et al., “All-Optical Reservoir Computing on a Photonic Chip Using Silicon-Based Ring Resonators,” IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 6, Nov/Dec 2018

[170] K. Vandoorne et al., “Experimental demonstration of reservoir computing on a silicon photonics chip”, Nature March 2014

[171] C. Mesaritakis, M. Skontranis, G. Sarantoglou and A. Bogris, “Micro-Ring-Resonator Based Passive Photonic Spike-TimeDependentPlasticity Scheme for Unsupervised Learning in Optical Neural Networks”, OFC 2020

[172] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, “Experimental realization of any discrete unitary operator”, Phys. Rev. Lett., vol. 73, no. 1, pp. 58–61, Jul. 2002

[173] F. Shokraneh, S. Geoffroy-Gagnon, M. S. Nezami and O. Liboiron-Ladouceur, "A Single Layer Neural Network Implemented by a 4×4 MZI-Based Optical Processor", IEEE Photonics Journal, vol. 11, no. 6, pp. 1-12, Dec. 2019

[174] D. A. B. Miller, “Self-configuring universal linear optical component [Invited],” Photon. Res., vol. 1, no. 1, p. 1, Jun. 2013

[175] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, “Optimal design for universal multiport interferometers”, Optica, vol. 3, no. 12, pp. 1460–1465, Dec. 2016

[176] J. Gu et al., “Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture”, ASPDAC 2020

[177] Z. Li, S. Wang, C. Ding et al., “Efficient recurrent neural networks using structured matrices in fpgas,” ICLR Workshop, 2018

[178] J. Friedman, T. Hastie, and R. Tibshirani, “A note on the group lasso and a sparse group lasso,” arXiv preprint arXiv:1001.0736, 2010

[179] B. Shi, D. Bunandar, D. Englund and R. Stabile, “WDM Weighted Sum in an 8x8 SOA-Based InP Cross-Connect for Photonic Deep Neural Networks”, Photonics in Switching and Computing (PSC), 2018

[180] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” Math. Comput., vol. 19, no. 90, 1965

[181] D. Brunner and I. Fischer, ``Reconfigurable semiconductor laser networks based on diffractive coupling," Opt. Lett., vol. 40, no. 16, pp. 3854_3857, 2015

[182] J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, “Reinforcement learning in a large-scale photonic recurrent neural network,” Optica, vol. 5, pp. 756–760, 2018

[183] M. C. Mackey, L. Glass, “Oscillation and chaos in physiological control systems,” Science, vol. 197, pp. 287-289, 1977

[184] J. Dong, M. Rafayelyan, F. Krzakala, and S. Gigan, “Optical reservoir computing using multiple light scattering for chaotic systems prediction,” IEEE J. Sel. Top. Quantum Electron. Vol. 26, pp. 1–12, 2020

[185] G. Bi and M. Poo, “Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type,” J. Neurosci., vol. 18, no. 24, pp. 10464–10472, Dec. 1998

[186] L. F. Abbott and S. B. Nelson, “Synaptic plasticity: Taming the beast,” Nature Neurosci., vol. 3, pp. 1178–1183, Nov. 2000

[187] G. Q. Bi and M. M. Poo, “Synaptic modification by correlated activity: Hebb’s postulate revisited,” Annu. Rev. Neurosci., vol. 24, pp. 139–166, Mar. 2001

[188] B. Shi, N. Calabretta and R. Stabile, “Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect”, IEEE J. Sel. Top. Quantum Electron, VOL. 26, NO. 1, Jan/Feb 2020

[189] M. P. Fok, Y. Tian, D. Rosenbluth, and P. R. Prucnal, “Pulse lead/lag timing detection for adaptive feedback and control based on optical spike timing-dependent plasticity,” Opt. Lett., vol. 38, no. 4, pp. 419–421, Feb. 2013

[190] Q. Ren, Y. Zhang, R. Wang, and J. Zhao, “Optical spike-timing dependent plasticity with weight-dependent learning window and reward modulation,” Opt. Express, vol. 23, no. 19, pp. 25247–25258, Sep. 2015

[191] N. Caporale and Y. Dan, “Spike timing-dependent plasticity: a Hebbian learning rule,” Annual Review of Neuroscience, vol. 31, 2008

[192] S. Xiang et al., “STDP-Based Unsupervised Spike Pattern Learning in a Photonic Spiking Neural Network With VCSELs and VCSOAs”, IEEE J. Sel. Top. Quantum Electron, vol. 25, no. 6, Nov-Dec 2019

[193] A. Hurtado, I. D. Henning and M. J. Adams, “Effects of parallel and orthogonal polarization on nonlinear optical characteristics of a 1550 nm VCSOA,” Opt. Express, vol. 15, no. 14, pp. 9084–9089, Jul. 2007

[194] M. D. Sánchez, P. Wen, M. Gross, and S. C. Esener, “Rate equations for modeling dispersive nonlinearity in Fabry-Perot semiconductor optical amplifiers,” Opt. Express, vol. 11, no. 21, pp. 2689–2696, Oct 2003

[195] A. Hurtado and M. J. Adams, “Two-wavelength switching with 1550 nm semiconductor laser amplifiers,” J. Opt. Netw., vol. 6, no. 5,pp. 434–441, May 2007

[196] Z. Song et al., “Spike Sequence Learning in a Photonic Spiking Neural Network Consisting of VCSELs-SA With Supervised Training”,IEEE J. Sel. Top. Quantum Electron, VOL. 26, NO. 5, Sept/Oct 2020

[197] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen and P. Bienstman, “Parallel reservoir computing using optical amplifiers”, IEEE Trans. Neural Netw., vol. 22, no. 9, 2011

[198] X. Xing Guo et al., “High-Speed Neuromorphic Reservoir Computing Based on a Semiconductor Nanolaser With Optical Feedback Under Electrical Modulation”, IEEE J. Sel. Top. Quantum Electron, VOL. 26, NO. 5, Sept/Oct 2020

[199] A. S. Weigend and N. A. Gershenfeld, “Time series prediction: Forecasting the future and understanding the past,” 1993,

[Online]. Available:

http: //www-psych.stanford.edu/andreas/Time-Series/SantaFe.html

[200] Y. Zhu et al., “Countering Variations and Thermal Eects for Accurate Optical Neural Networks”, IEEE ICCAD 2020

[201] M.Y.-S. Fang et al., “Design of optical neural networks with component imprecisions”, Opt. Express vol. 27, pp. 14009-14029, 2019

[202] F. Zokae et al., “LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks”, Design, Automation Test in Europe Conference Exhibition (DATE), 2020

[203] M. Nikdast, G. Nicolescu, J. Trajkovic, and O. Liboiron-Ladouceur, “Modeling fabrication non-uniformity in chip-scale silicon photonic interconnects,” Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016.

[204] S. V. R. Chittamuru, I. G. Thakkar, and S. Pasricha. "Analyzing voltage bias and temperature induced aging effects in photonic interconnects for manycore computing." SLIP. 2017.

[205] J. S. Orcutt, A. Khilo, C. W. Holzwarth, M. A. Popovic, H. Li, J. Sun, T. Bonifield, R. Hollingsworth, F. X. Kartner, H. I. Smith, V. Stojanovic, and R. J. Ram “Nanophotonic integration in state-of-the-art CMOS foundries,” Optics Express, vol. 19, pp. 2335-2346, 2011

[206] A. Mahendra, C. Xiong, X. Zhang, B. J. Eggleton, and P. H. W. Leong, “Multiwavelength stabilization control of a thermo-optic system with adaptive reconfiguration,” Applied Optics, vol. 56, no. 4, pp. 1113-1118, 2017

[207] W. R. Clements, William R., P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, “Optimal design for universal multiport interferometers,” Optica, vol. 3, no. 12, pp. 1460–1465, 2016

[208] D. Dang, S. V. R. Chittamuru, S. Pasricha, R. Mahapatra, D. Sahoo, “BPLight-CNN: A Photonics-based Backpropag or Deep Learning”, to appear, ACM Journal on Emerging Technologies in Computing Systems (JETC), 2021.

ОТКАЗОУСТОЙЧИВОСТЬ И ПОМЕХОУСТОЙЧИВОСТЬ В ДИФРАКЦИОННЫХ ОПТИЧЕСКИХ НЕЙРОННЫХ СЕТЯХ В СВОБОДНОМ ПРОСТРАНСТВЕ

Панда С.С., Хехгде Р.С.
http://dx.doi.org/10.1364/ao.XX.XXXXXX

Дифракционные оптические сети в свободном пространстве – это класс обучаемых оптических носителей, которые в настоящее время исследуются как новая аппаратная платформа для нейронных двигателей. Фаза обучения таких систем обычно выполняется на компьютере, а затем полученные веса переносятся на оптическое оборудование («обучение вне поля»). Хотя этот процесс переноса веса имеет много практических преимуществ, он часто сопровождается сбоями, снижающими производительность изготовленного оборудования. Будучи аналоговыми системами, эти двигатели также подвержены снижению производительности из-за шумов на входах и во время оптоэлектронного преобразования. Рассматривая дифракционные оптические сети (DON), обученные для задач классификации изображений на стандартных наборах данных, мы численно изучаем ухудшение характеристик, возникающее из-за ошибок веса и внесенных шумов, а также методы для уменьшения этих эффектов. Режим обучения, основанный на преднамеренном сбое и внесении шума во время фазы обучения, оказывается лишь незначительно успешным с точки зрения обеспечения отказоустойчивости или помехоустойчивости. Мы предлагаем новый режим обучения с использованием условий регуляризации на основе градиента в целях обучения, которые, как обнаружено, придают некоторую степень отказоустойчивости и помехоустойчивости.

ВВЕДЕНИЕ

Продолжающиеся разработки в области искусственного интеллекта (ИИ) и технологий больших данных потребовали более глубокого изучения новых альтернативных аппаратных средств нейронной обработки. [1–3] в обычную твердотельную электронику. Оптика – это платформа [4–10], которая может потенциально улучшить скорость обработки [11] и энергоэффективность, особенно на этапе вывода. Существует долгая история исследований в области применения оптики для реализации нейронных сетей (NN) [12]. Возрождение интереса к этой области наблюдается [5, 13] с такими подходами, как когерентные нанофотонные схемы [14], нейроморфная фотоника [15] и дифракционные оптические нейронные сети. Явление дифракции обеспечивает естественный способ

реализации плотно взаимосвязанных элементов, которые напоминают трехмерную топологию биологических нейронных сетей. Рассматривая каноническую сеть, в которой нейроны расположены в последовательных плоскостях с межсоединениями, занимающими промежуточное пространство, Динк с соавторами [16] показывают, что площадь, необходимая для размещения N нейронов, масштабируется как $\sqrt{N}$. Архитектура, в которой размещены как нейроны, так и межсоединения, требования к площади масштабируются как N. Таким образом, дифракционные оптические связи могут позволить нейронные сети со значительно увеличенным числом нейронов (число в конечном итоге ограничено ограничениями, разработанными в [17]). Другим важным преимуществом является то, что поглощение света является единственным механизмом тепловыделения в оптических дифракционных межсоединениях (эти потери могут быть низкими при использовании свободного пространства). Недавний интерес был вызван отчетом о дифракционной глубокой нейронной сети (D2NN), созданным Лином и его сотрудниками [18], которые экспериментально продемонстрировали, что линейная пассивная дифракционная оптическая система, работающая в терагерцовом диапазоне частот, может достичь более 90% классификации. точность набора данных MNIST. С тех пор во многих статьях были предложены модификации к исходной статье о дифракционных сетях, написанной Лином и соавторами [18–21], включая гибридизацию с электронными слоями [22], множественные частотные каналы [23] и дифференцированное обнаружение для конкретных классов [24]. Некоторые авторы исследовали свойство преобразования Фурье установки 4f корелятора для реализации полностью оптического двумерного сверточного слоя перед сопряжением с электроникой [25–27]. Дифракционные нейронные сети в свободном пространстве были исследованы в 1980-х годах с использованием нелинейных фоторефрактивных материалов [28–30], а веса сети обучались аппаратно [31]. Нынешние воплощения дифракционных нейронных сетей в свободном пространстве значительно проще своих предшественников. Это полностью линейные системы, и веса нейронов (то есть обучаемые параметры) определяются на цифровом компьютере, а затем загружаются в оптическое оборудование во время изготовления («обучение вне помещения»). Программное обеспечение для обучения и тестирования глубоких нейронных сетей оптимизировано для простоты использования и привело к парадигме проектирования путем «обучения» (т.е. обучаемое фото 2

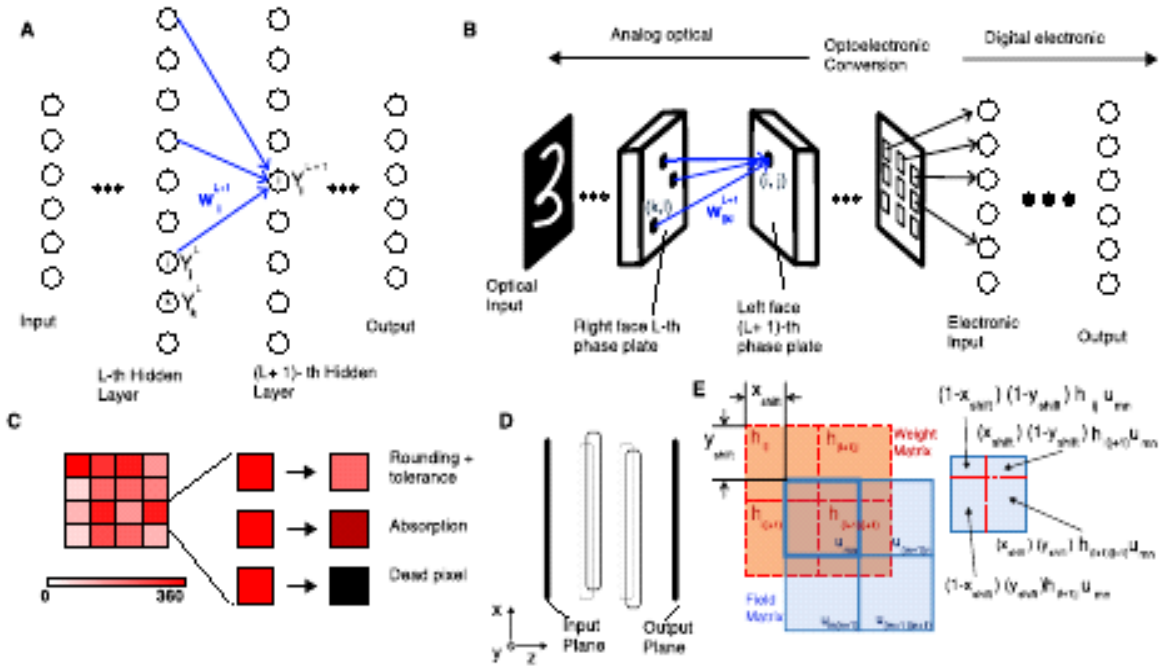


Рис. 1. A: схема многослойных глубоких нейронных сетей с прямой связью. B: гибридная аналого-оптическая / цифрово-электронная установка, состоящая из нескольких предварительно запрограммированных фазовых пластин, которая ведет себя как многослойная NN. C: часть разработанной чистой фазовой пластины и 3 вида некоррелированных сбоев пикселей, которые могут возникнуть: (i) округление и допуск, (ii) случайное поглощение и (iii) мертвый пиксель, который полностью блокирует свет. D: смещение фазовых пластин от их идеального положения, ведущее к коррелированным ошибкам. E: численная схема для моделирования распространения в системе с боковыми смещениями, в которой используется среднее значение поля по пикселю для оценки модифицированного произведения Адамара

tonic media [32, 33]) и, в сочетании с разработками в области метаповерхностей [7, 34–37], обещает новые способы проектирования устройств для светлый мани Пуля. Обучение на месте (т.е. обучение с использованием оптического оборудования в контуре) [38] в принципе возможно для дифракционных сетей [31, 39] и обучаемых фотонных сред. Но его реализация значительно усложняет оптическую установку и может сделать дизайн свободного пространства неконкурентоспособным с альтернативными решениями, такими как интегрированная фотоника [40]. С другой стороны, проблема с обучением вне помещения состоит в том, что веса, полученные во время обучения на плаву, нельзя передать на оптическое оборудование с идеальной точностью, что приведет к ошибочной копии, которая может показать снижение производительности. Дополнительная сложность заключается в том, что оптические реализации (независимо от онлайн-обучения или автономного обучения) представляют собой аналоговые системы [41] и демонстрируют повышенную чувствительность к шумам сигнала и меньшую точность по сравнению с цифровыми нейронными сетями. Таким образом, для

практической системы чрезвычайно важно, чтобы характеристики плавно ухудшались при наличии неисправностей и шумов. Отказоустойчивость и помехоустойчивость будут важными показателями производительности, которые в конечном итоге определят конкурентоспособность дифракционных оптических нейронных двигателей и устройств, основанных на обучаемых фотонных средах. В частности, для масштабирования рабочей длины волны DON до видимых длин волн отказоустойчивость и помехозащищенность будут критическими проблемами. Несмотря на свою важность, этому аспекту пока не уделяется достаточного внимания в литературе. Об исследованиях неточности компонентов и стратегий проектирования для обеспечения устойчивости при наличии неточностей компонентов сообщалось в интегрированных фотонных нейронных двигателях [42, 43]; однако результаты и идеи не могут быть напрямую перенесены на устройства freespace. Менгу и его коллеги считают, что обучение DON инвариантно к изменениям масштаба, сдвига и поворота входного изображения [44]. Однако эту проблему можно решить, пополнив обучающий набор данных. Но увеличение набора обучающих данных не может обеспечить отказоустойчивость и иммунизацию шума. Shi и соавторы [45] рассмотрели обучение введению весового шума для шумов фазовых возмущений в наборе данных MNIST. Их исследование не проверяло сети на ряд возможных ошибок, и поэтому неясно, сработала ли их стратегия. Семенова [41] недавно сообщила об аналитическом исследовании неисправностей и шумов в аналоговых нейронных двигателях, но их сети намного меньше типичных DON. Отказоустойчивость и помехоустойчивость широко рассматривались [46–48] в литературе применительно к аналоговым электронным нейронным двигателям. Сбои в аналоговых электронных двигателях в основном бывают двух типов [46, 49]: (1) застрявшие сбои, когда выход нейрона не реагирует на его вход и постоянно привязан к низкому или высокому уровню; (2) пертурбативные неисправности аддитивного или мультипликативного характера. Природа неисправностей и шумов, возникающих в DON, отличается от аналоговой электроники по ряду причин: (1) обучаемые веса фазовой маски в DON ограничены диапазоном от 0 до 2π; (2) неисправности в DON могут иметь корреляции (например, сдвиг фазовой пластины); и (3) количество нейронов в DON может быть большим. Чтобы правильно оценить ухудшение характеристик DON, зависящее от сбоев и шума, важно создавать реалистичные модели, а затем разрабатывать новые схемы обучения, чтобы обеспечить определенную степень надежности. Литература по нейронным сетям показала, что отказоустойчивость и помехозащищенность не обязательно должны быть неотъемлемыми чертами DON [46]. Однако известно, что специализированные процедуры обучения могут в определенной степени спо-

собствовать повышению отказоустойчивости и помехоустойчивости. 3 В этой статье мы представляем всестороннее численное исследование, в котором изучаются эффекты снижения производительности из-за неисправностей и шумов сигналов на примере DON, обученного для задач классификации изображений на стандартных наборах данных (MNIST, Fashion MNIST и CiFAR-10 в оттенках серого). Мы разрабатываем реалистичные модели неисправного DON, чтобы оценить снижение производительности и разработать режим тренировки для придания устойчивости. Во-первых, мы оцениваем эффективность процедур обучения ex-situ на основе индукции неисправностей и шумов [45, 49, 50]. Мы обнаружили, что стандартные предписания по наведению неисправностей и шумов не работают в случае DON. Мы предлагаем альтернативные стратегии обучения, включающие упорядоченные цели [51], и показываем, что они могут обеспечить некоторую степень отказоустойчивости и помехоустойчивости. Работа вместе с соответствующим программным обеспечением будет полезна для направления дальнейших экспериментальных работ в дифракционных оптических сетях в свободном пространстве и может вдохновить на дальнейшую работу по разработке отказоустойчивых обучаемых оптических сред. После этого введения оставшаяся часть статьи организована следующим образом: (1) в разделе 2 мы описываем процесс имитации DON с ошибками и входным шумом на цифровом компьютере и дополнительно обсуждаем соответствующие стратегии обучения; (2) в разделе 3 мы сначала сравниваем производительность идеального DON (отсутствие сбоев или шумов), затем представляем и обсуждаем относительное ухудшение производительности, возникающее из-за различных видов сбоев и шумов, и, наконец, обсуждаем эффективность робастного обучения. стратегии; (3) статья завершается в разделе 4.

ЭМУЛЯЦИЯ НЕПРАВИЛЬНОГО ДОНА, НАДЕЖНЫЙ РЕЖИМ ОБУЧЕНИЯ

Процедура обучения ex-situ включает создание точной модели дифракционного оптического нейронного двигателя. Создание такого эмулятора в идеальном случае и включение ошибок и сигнального шума описано в этом разделе, а за ним следует обсуждение различных режимов тренировки, с которыми проводились эксперименты. А. Модель потока оптических сигналов в рамках теории скалярной дифракции. На рис. 1B представлена схема широкого класса дифракционно-оптических нейронных процессоров, рассматриваемых в этой статье. Система имеет аналого-оптический интерфейс, слой оптоэлектронного преобразования и, наконец, цифрово-электронный интерфейс. Оптическая часть начинается с входного слоя (который может включать преобразование входного электронного сигнала в оптическое изображение), за ним следует каскад

из одной или нескольких пиксельных фазовых пластин, разделенных свободным пространством, и заканчивается на уровне оптоэлектронного преобразования. Пиксели можно интерпретировать как «нейроны», а дифракционная связь в свободном пространстве между пикселями на разных пластинах аналогична взвешенным межсоединениям стандартной многослойной сети с прямой связью, показанной на рисунке 1A. Для наиболее общего описания потребуется спецификация векторы поля во всех пространственных точках оптической части системы. Во многих случаях скалярная теория дифракции и дискретное представление полей в определенных плоскостях адекватны для полного описания потока [52], и этот подход рассматривается здесь. На рисунке 1A показаны два промежуточных слоя (L-й и L + 1-й уровни) в общей нейронной сети прямого распространения, содержащей по N нейронов каждый. Активация i-го нейрона в L + 1-м слое YiL + 1 может быть выражена в терминах активаций YjL с j = 1 · · · N нейронов предыдущего слоя как: YiL + 1 = f (N ∑ j = 1 wiLj + 1YjL + biL + 1), где wiLj + 1 и biL + 1 – обучаемые веса и смещения L + 1-го слоя, а f – функция активации (обычно нелинейная функция, такая как ReLU). Рассмотрим свободное пространство между любыми двумя соседними фазовыми пластинами, расположенными на расстоянии d друг от друга, как показано на рисунке 1 B. Предполагая монохроматическое когерентное освещение и пренебрегая обратными отражениями на фазовых пластинах, дискретизированное поле u для пикселя размером P × P в (i, j) в левой части L + 1-й фазовой пластины можно выразить через поля v во всех пикселях в правой части L-го слоя как [52]:

$$u_{i,j}^{L+1} = \sum_{k=1}^{N} \sum_{l=1}^{N} w_{ijkl}^{L+1} v_{k,l}^{L}, \quad (1)$$

$$w_{ijkl}^{L+1} = \left[\frac{1}{j\lambda} + \frac{1}{2\pi r_{ijkl}} \right] \frac{d \exp\left(j \frac{2\pi}{\lambda} r_{ijkl}\right)}{r_{ijkl}^2}, \quad (2)$$

где i, j = 1 · · · N – индекс пикселей на L + 1-й фазовой пластине и k, l = 1 · · · N – индекс пикселей на L-й фазовой пластине, wijkl связывает отдельные пиксели и зависит от расстояния $r_{ijkl} = \sqrt{P^2(i-k)^2 + P^2(j-l)^2 + d^2}$ между ними и длиной волны освещения λ. Это уравнение, описывающее преобразование поля, в общем случае является тензором. Когда апертура фазовой пластинки d P ∗ N и d λ, пропагатор Френеля может быть использован для аппроксимации этого преобразования поля (известно, что приближение остается справедливым даже для меньших расстояний во многих случаях [52]). В правой

части L + 1-й фазовой пластинки поле v возникает из-за локальных изменений как амплитуды, так и фазы и задается произведением Адамара:

viL, j + 1 = uiL, j + 1hiL, j + 1 и изученный вес hiL, j + 1 = AiLj + 1 exp (jΦiLj + 1),

где AiLj + 1 и ΦiLj + 1 – локальные изменения амплитуды и фазы, передаваемые пикселем в местоположении (i, j). Мы кратко отметим различия между традиционными нейронными двигателями и оптическими дифракционными нейронными двигателями. Веса, фигурирующие в уравнении 2, не обучаются. Обучаемые параметры появляются только на фазовых пластинах. Оптическая часть заканчивается оптоэлектронным преобразовательным слоем, который состоит из массива фотодетекторов (PD). Количество фотодетекторов обычно соответствует количеству классов в наборе данных m (10 для всех наборов данных, рассматриваемых в этой статье). Предполагается, что каждый из частичных разрядов реагирует на подобласть области, охватываемой фазовыми пластинами. Принимая во внимание бесшумный частичный размах и непрерывное освещение волны, фототок в i-м частном разряде определяется выражением:

$$c_i = \eta_Q \sum_{l=l_i}^{l_i+N_{pd}} \sum_{m=m_i}^{m_i+N_{pd}} |u_f(l,m)|^2 p^2, \quad (3)$$

где Npd – ширина PD в количестве пикселей, P – ширина одного пикселя, ηQ – его квантовая эффективность, а uf – двумерное дискретизированное поле в плоскости PD. Максимальный фототок в PD, cimax ограничен максимальной интенсивностью пикселей на входной плоскости Imax и задается как ηQNp2dP2Imax. Индекс фототока, который имеет самый высокий фототок, обозначает класс входного изображения. В качестве альтернативы, массив фототоков также может подаваться на электронную нейронную сеть [22, 53].

A.1. Сверточные слои и оптическая нелинейность.

Классический 4f-коррелятор можно использовать для создания оптического сверточного слоя путем размещения фазовой маски в плоскости Фурье. Колберн с соавторами [26] рассматривают многократно сложенное входное изображение 4 с физически разделенными корреляторами 4f для реализации множественных сверток одного входного изображения. Чанг и его коллеги [25] сообщили, что, когда изображение вводится с достаточным заполнением нулями (см. Рисунок 1 D), выходные данные можно интерпретировать как стек сверток с пространственным смещением входного изображения с разными ядрами.. Здесь сверточный слой реализуется путем добавления программируемой фазовой пласти-

ны в плоскости Фурье идеального коррелятора 4f. Таким образом, фазовая пластина изменяет амплитуду и фазы различных компонентов пространственной частоты. После выполнения взвешенного суммирования входных данных и добавления члена смещения нейрон в традиционной нейронной сети применяет к нему нелинейную функцию активации. В оптике эквивалентный подход состоит в рассмотрении фазовой пластины, где фазовая и / или амплитудная характеристика пикселя зависит от локальной интенсивности. Конкретная форма локальных изменений зависит от наличия / отсутствия и конкретного типа зависящей от интенсивности оптической нелинейности в фазовой пластине. Для трех рассматриваемых здесь различных случаев: (1) наличие насыщающейся нелинейности поглощения, (2) наличие нелинейности керровского типа и (3) незначительная нелинейность, мы рассматриваем следующие упрощенные выражения:

$$A_{ij}^{L+1} = \begin{cases} \begin{cases} \dfrac{\sqrt{|u_{i,j}^{L+1}|^2 - I_{sat}}}{|u_{i,j}^{L+1}|} & \text{if } |u_{i,j}^{L+1}|^2 \geq I_{sat} \quad \text{Sat. absorber} \\ 0 & \text{otherwise} \end{cases} \\ 1 \qquad\qquad \text{Otherwise} \end{cases}$$

$$\Phi_{ij}^{L+1} = \begin{cases} \Phi_{ij}^{0} + 2\pi \dfrac{|u_{i,j}^{L+1}|^2}{I_{full}}, & \text{Kerr-type nonlinearity} \\ \Phi_{ij}^{0} & \text{Otherwise.} \end{cases} \tag{4}$$

(4)

Для чистых фазовых масок, рассматриваемых в уравнении 4, набор фаз Фi0j является обучаемыми параметрами. Параметры Ifull и Isat являются зависящими от материала константами, которые характеризуют степень нелинейности (Ifull – это локальная интенсивность, которая дает сдвиг фазы 2π, а Isat – это пороговая интенсивность, при которой поглощение полностью насыщается. B. Модель неисправности и шума We различать неисправности (которые относятся к весам) и шумы (которые относятся к распространяющимся оптическим полям и процессу оптоэлектронного преобразования). Рассматривая архитектуру каскадных программируемых фазовых пластин, следующие неисправности (см. рисунок 1 C и D) учитываются в модели сбоя: (1) фазовые сбои, возникающие из-за округления, (2) сбои по фазе из-за производственных допусков и межпиксельной перекрестной связи; (3) неоднородное случайное поглощение в пикселях в только фазовой конструкции; (4) дефекты, возникающие из-за смещения пластин в поперечном направлении; и (5) мертвые пиксели, которые полностью блокируют свет. Поскольку нанесение тонких пленок является высокоразвитой технологией, дефекты, возникающие из-за расслоения пластин рваные раны в продольном

направлении и повороты пластины здесь не учитываются. Рассмотрены два источника шума: (1) шум на входном каскаде; и (2) шум в значениях фототока ci, возникающий из-за шума в частичных разрядах. Сбои по фазе, возникающие из-за округления, производственных допусков и межпиксельной связи в любых двух пикселях, можно рассматривать как некоррелированные друг с другом. Кроме того, эти сбои не коррелируют с ошибками по амплитуде и битыми пикселями. Шумы на входе и выходе также не коррелируют друг с другом. Однако неисправности, возникающие в результате смещения пластин, влияют на несколько пикселей одновременно. При наличии неисправностей в i, j-м пикселе на пластине L + 1 мы можем записать преобразование локального поля в виде:

$$v_{i,j}^{L+1} = u_{i,j}^{L+1} A_{ij}^{L+1} \exp\left(j\Phi_{ij}^{L+1}\right) \times A^{U} \exp\left(j(\Phi^{N} + \Phi^{R})\right), \quad (5)$$

где AU ~ U (к, 1) – равномерно распределенная случайная величина, обозначающая потери случайного поглощения, ΦN ~ N (0, Г) – нормально распределенная случайная величина, обозначающая аддитивный фазовый шум, а ΦR – ошибка аддитивного фазового шума, возникающая из-за дискретизации фазы. Мертвый пиксель моделируется установкой его выходного поля на ноль. Позиции отдельных фазовых пластин могут быть смещены от идеального положения (положения, предполагаемого при обучении), как показано на рисунке 1. Эти сдвиги изменяют входные данные для каждого пикселя в зависимости от его положения, и изменения по пикселям коррелируются. Этот источник коррелированного шума может быть реализован численно по схеме, показанной на рисунке 1. Для простоты плоскости входа и выхода считаются идеально выровненными друг относительно друга. Рассмотрим пластину, которая сместилась от своего идеального положения на перемещения xshift ≤ P и yshift ≤ P в направлениях x и y соответственно. Матрица поля на выходе L-й пластины выражается как взвешенная сумма четырех членов (см. Рисунок 1 E):

$$\begin{aligned} v_{i,j}^{L} = u_{i,j}^{L} \times \Big(& (1 - |x_{shift}|)(1 - |y_{shift}|) h_{i,j}^{L} \\ & + |x_{shift}|(1 - |y_{shift}|) h_{i,j\pm1}^{L} \\ & + (1 - |x_{shift}|)|y_{shift}| h_{i\pm1,j}^{L} \\ & + |x_{shift}||y_{shift}| h_{i\pm1,j\pm1}^{L} \Big). \end{aligned} \quad (6)$$

Это выражение равносильно приближению среднего поля и становится все более достоверным, когда размер пикселя P становится меньше по сравнению с рабочей длиной волны. Знак индекса i в u в приведенном выше уравнении принимается как положительный или отрица-

тельный в зависимости от того, является ли сдвиг x положительным или отрицательным (и аналогично для j). На входе изображения учитывается случайный аддитивный шум. Процесс оптоэлектронного преобразования на этапе фотодетектора может вызвать ошибки, которые зависят от ширины полосы сигнала. Более быстрая работа нейронного двигателя увеличит полосу пропускания и, следовательно, уровень шума. Вместо детальной модели рассматривается аддитивный гауссов шум для стадии фотоопределения. При наличии шума в PD фототок в i-м PD определяется выражением:

$$c_i = \eta_Q \sum_{l=l_i}^{l_i+N_{pd}} \sum_{m=m_i}^{m_i+N_{pd}} |w_f(l,m)|^2 + c_{noise}, \quad (7)$$

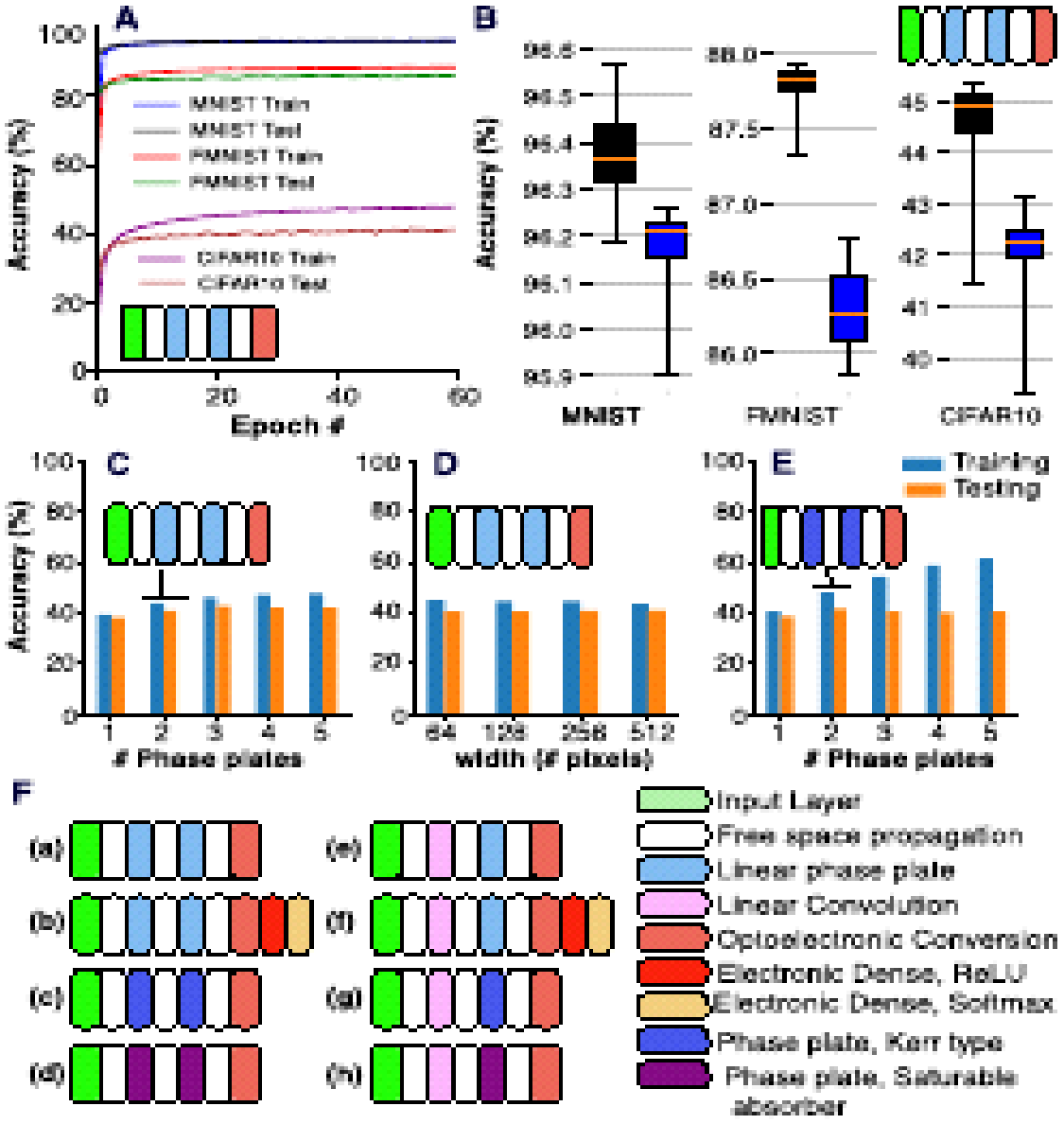
где cnoise – нормально распределенная случайная величина с нулевым средним.

C. Режим обучения для повышения устойчивости

В этой статье мы исследовали две стратегии повышения отказоустойчивости: (1) добавление весового шума во время обучения и (2) использование заданной пользователем целевой функции обучения. Аналогичным образом были исследованы две стратегии обеспечения помехоустойчивости: (1) ввод входного шума во время обучения и (2) использование заданной пользователем целевой функции обучения. Оба этих подхода (введение шума [49, 50] и добавление штрафных терминов регуляризации [51]) были рассмотрены в литературе по нейронным сетям. В документе оценивается эффективность этих стратегий в придании отказоустойчивости и помехоустойчивости результирующим сетям при загрузке на аналоговую аппаратную платформу. Стратегия введения весового шума добавляет преднамеренный шумовой член 5

Статистика точности обучения и тестирования для трех наборов данных для той же сетевой архитектуры, что и в A (после 60 эпох обучения) с использованием 20 различных инициализаций веса. C, E: Зависимость классификации изображений CiFAR-10 производительность (после 60 эпох) по глубине ДОН (количество фазовых пластин) и ширине ДОН (количество пикселей в каждой фазовой пластине). В сетях C, D используются линейные фазовые пластины, а в D – фазовые пластины с оптической нелинейностью типа Керра (см. Уравнение 4, Ifull = 2). F: восемь различных типов DON и типичная точность обучения (Tr) и тестирования (Te) после 60 эпох обучения на трех наборах данных. Набор тестовых данных – 10k для всех, обучающий набор данных – 60k для MNIST, F-MNIST и 50k для CiFAR-10. Фазовые пластины с нелинейностью типа Керра и с насыщающимся поглощением характеризуются Ifull = 2Imax и Isat = 0,1Imax соответственно, а пиковая интенсивность вход-

ного изображения (соответствующая белому пикселю) установлена на 1,0. к сетевым обучаемым весам в каждую тренировочную эпоху.



		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
MNIST	Tr	96.06	96.35	98.66	94.37	96.05	96.78	91.08	90.75
	Te	96.03	96.78	97.64	97.44	96.16	97.02	91.75	91.35
F-MNIST	Tr	87.54	87.91	88.42	84.90	88.48	89.03	84.05	83.10
	Te	86.08	86.25	86.49	85.87	85.92	86.29	82.78	81.24
CIFAR-10	Tr	42.56	45.30	44.52	37.78	47.84	49.36	36.59	41.30
	Te	38.62	42.58	40.71	42.10	40.50	43.97	35.71	37.31

Рис.2 A: Повышение точности классификации при обучении и тестировании за периоды обучения на наборах данных MNIST, FMNIST и Cifar-10 для экземпляра сети, показанной на вставке. B:

Вводимый шум взят из нормального распределения с нулевым средним и заданным стандартным отклонением. Стратегия впрыска шума

аналогична, но добавляет шум на вход. Для второго подхода модифицированная цель обучения определяется, как показано ниже, с использованием набора взвешенных членов регуляризации с использованием весовых констант λf, λi и λt: – M ∑ c = 1 y0, c · ln pc Категориальные потери кроссэнтропии + λfVar 迸 翎鼸 喰 ∂pc ∂hi 翎 鼸 喰 迸 регуляризация отказоустойчивости + λiVar 鼸 刀 翎 ⺊ ∂pc ∂xi 刀 翎 ⺊ регуляризация помехоустойчивости + λt max | hm, n – hm − 1, n | регуляризация инвариантности к сдвигу + λt max | hm, n – hm + 1, n | + λt max | hm, n – hm, n − 1 | + λt max | hm, n – hm, n + 1 |, 1 ≤ c ≤ M, 1 ≤ i ≤ W 1 ≤ i ≤ N2 1 ≤ m, n ≤ N. (8) В приведенном выше уравнении члены имеют были идентифицированы. Выход слоя фотодетектора p представляет собой массив из 10 элементов (т.е. M = 10), а y – массив из 10 элементов в формате однократного кодирования, которые являются обучающими метками. Функция var обозначает дисперсию массива различных градиентов. Регуляризация отказоустойчивости использует градиенты вывода относительно обучаемых весов сети. С другой стороны, термин регуляризации иммунизации шума использует градиенты выхода по отношению к входам сети. Интуиция, лежащая в основе предлагаемых условий регуляризации, заключается в том, что они уменьшают зависимость сети от индивидуальных весов или входных данных, что добавляет степень устойчивости [51]. Другими словами, эта регуляризация пытается гарантировать, что один вес или вход не более значим, чем другой. Требование устойчивости должно быть сбалансировано с требованием точности сети. Наконец, минимизация среднего значения разности фаз между соседними пикселями на фазовой пластине придаст определенную степень устойчивости к рассогласованиям фазовой пластины.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Численные эксперименты были выполнены на платформе глубокого обучения Keras [54] с бэкэндом Tensor Flow на рабочей станции с процессором Intel ™ i9–7920X с графической картой NVIDIA ™ GeForce GTX 1080 с 12 ГБ памяти. Исходный код реализации, наборы данных и сохраненные модели доступны в сети [55]. Все сети были обучены в течение 60 эпох (время обучения 15 минут) с редкими категориальными потерями кроссэнтропии с использованием оптимизатора NADAM [54]. Использовались три хорошо известных набора данных: MNIST, Fashion MNIST и CiFAR-10 (после преобразования в оттенки серого). Если не указано иное, поля и фазовые маски используют дискретизацию 256 на 256 пикселей размером 0,5λ0 (λ0 – длина волны в свободном пространстве). Слой оптоэлектронного преобразования (реализованный как средний пул) в некоторых случаях является последним слоем. В других

случаях используется один дополнительный электронный полносвязный слой с активацией ReLU. Выход последнего слоя – это одномерный вектор (размер 10) с индексом самого большого элемента, рассматриваемого как предсказанный класс.

В Обучение введению шума со стандартным отклонением 5%. Неисправности и шумы, возникающие во время тестирования, обозначены синим цветом.

Идеальная производительность для различных архитектур и масштабирование размера

В этом подразделе мы рассмотрим следующий вопрос: (1) как способность прогнозирования идеального DON (отсутствие сбоев и шумов) зависит от его конструкции (т. Е. Типа используемых слоев) и его сложности (количество обучаемых параметров)? Можно заметить, что для фиксированной структуры (определенное количество слоев и нейронов на слой) в процессе обучения могут быть получены различные наборы весов в зависимости от инициализации. Было обнаружено, что можно найти много различных наборов весов, которые имеют сходные характеристики классификации. Некоторые наборы весов могут быть более надежными, чем другие. Типичный сценарий обучения представлен на рисунке 2A для сети с двумя программируемыми фазовыми пластинами. Видно, что точность классификации сети быстро возрастает, а затем почти падает. Обучение DON очевидно, потому что во всех случаях оно обеспечивает лучшую классификацию, чем чистая случайность (10%). Однако видно, что производительность набора данных CiFAR-10 составляет менее 50%. Эти результаты согласуются с предыдущими отчетами, и, насколько нам известно, точность более 50% вообще не сообщалась.

На рисунке 2B видно, что некоторое улучшение производительности возможно за счет выполнения нескольких итераций и выбора наилучшего набора весов. На рисунке 2C – E мы рассматриваем эффект расширения фазовой пластины и добавления нескольких слоев в бесшумном DON (что составляет масштабирование по ширине и глубине). Для нелинейной фазовой пластины с нелинейностью типа Керра (рисунок 2 E) константа $I_{full} = 2 * I_{max}$ (см. Уравнение 4) устанавливается равной 2-кратной максимальной интенсивности входного изображения. Вопреки утверждениям [22, 56] о том, что параметр глубины может значительно улучшить характеристики, видно, что добавление дополнительных слоев и расширение фазовых пластин в поперечном направлении улучшили точность лишь незначительно и оказались стратегией с уменьшением отдачи для случай линейных фазовых пластин [57].

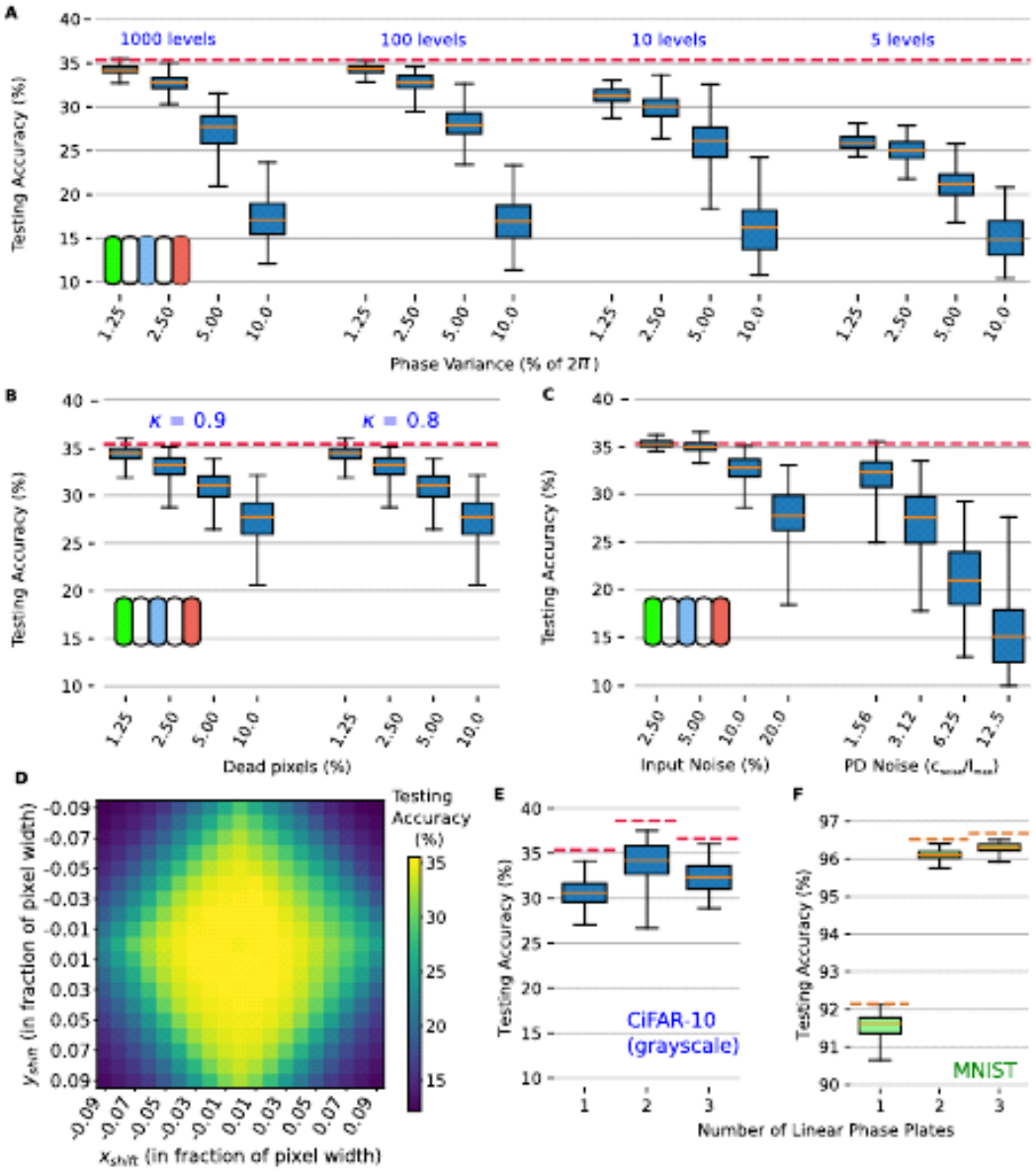


Рис. 3. Ухудшение точности классификации в зависимости от неисправности и шума для DON с линейными фазовыми пластинами. Только фазовые замыкания. Учитываются комбинации четырех уровней дискретизации фазы (1000, 100, 10 и 5) и четырех уровней дисперсии фазы (1,25%, 2,5%, 5% и 10%). B Только ошибки амплитуды. Комбинации двух констант затухания (0,8, 0,9) с четырьмя различными отношениями битых пикселей (1,25%, 2,5%, 5% и 10%). C Шум входной интенсивности и шум частичных разрядов. Четыре случая входного шума (2,5%, 5%, 10% и 20%) и шума фотодиода (PD) (1,56, 3,12, 6,25 и 12,5) с квантовой эффективностью, установленной на 90%. D Несоосность фазовой пластины в поперечной (x – y) плоскости. Учитываются комбинации D, E неисправностей и шумов: 100 фазовых уровней, фазовая ошибка 2,5%, затухание 0,9, битые пиксели 2,5%, входной шум 10% и шум частичных разрядов 1,56. B A – D рассматривается DON с однофазным слоем, в то время как в E, F рассматриваются три различных DON с одно-, двух- и трехфазным слоями соответственно. Красные пунктирные линии обозначают точность DON при отсутствии неисправностей и шумов.

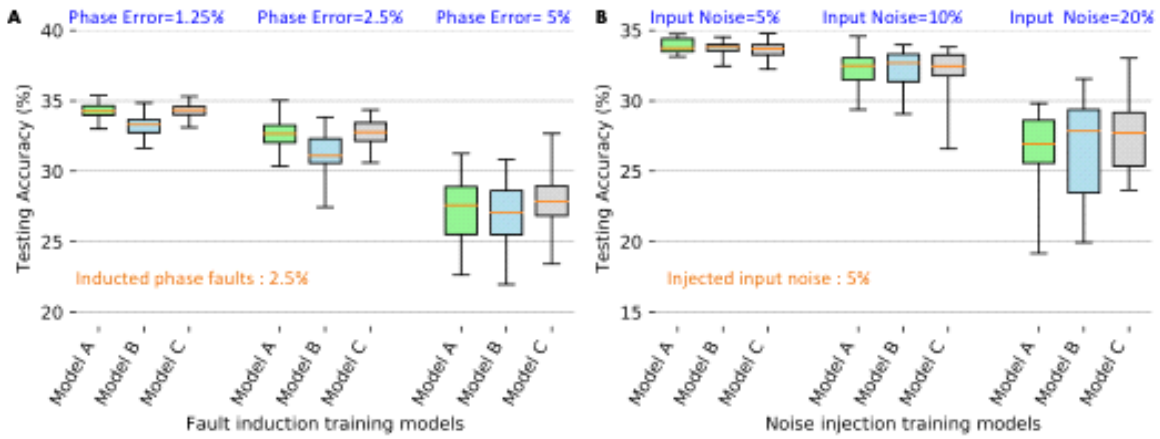


Рис. 4. Сравнительные характеристики моделей, обученных с введенными разломами и шумами (модель B и модель C), с моделью, обученной регулярно (модель A). Модель B изменяет вводимые неисправности / шумы в каждую функциональную эпоху, в то время как модель C меняет их каждые 5 эпох.A Обучение индукции неисправности с нормально распределенной случайной величиной с нулевым средним и фазовым углом 2,5% (часть стандартного отклонения 2π.

Перетренированность наблюдается в нелинейном случае (рисунок 2E) по сравнению с линейным случаем (рисунок 2A), что указывает на то, что способность к обучению сети увеличивается и что методы увеличения набора данных потенциально могут быть полезными. В следующем эксперименте рассматривались типичные характеристики классификации изображений восьми различных типов DON с двумя фазовыми пластинами каждая (см. Рисунок 2 F). В частности, это полезно для оценки относительной важности оптической нелинейности и оптической свертки. Точность обучения и тестирования, полученная в конце 60 эпох, отмечена в таблице на рисунке 2. Для нелинейной фазовой пластины с нелинейностью типа Керра (сети E и F) константа Ifull = 2 ∗ Imax (см. уравнение 4 устанавливается равным 2-кратной максимальной интенсивности входного изображения. Для нелинейной фазовой пластины с нелинейностью типа насыщающегося поглотителя (сети G и H) постоянная Isat = 0,1 ∗ Imax (см. уравнение 4 устанавливается равной 0,1 умноженной на максимальная интенсивность входного изображения. Отмечено, что производительность всех вариантов лучше всего с более простой задачей MNIST и хуже всего с задачей CiFAR-10. Однако удивительным результатом является то, что точность обучения и тестирования лишь незначительно улучшается при использовании сверточных слоев и оптической нелинейности. Добавление оптической нелинейности (столбец (e)) и / или электронной активации ReLU (столбцы (c)) и (d), как видно, приводит к некоторой степени перетренированности, указывающей на то, что обучение пропускная способность сети улучшается. Некоторые опасения вызывают n был поднят в связи с отсутствием оптической нелинейности и ее влиянием на работу

дифракционной оптической нейронной сети (DON) [58]. Результаты показывают, что простое наличие нелинейности не является гарантией улучшения качества классификации. Современные сверточные нейронные сети, которые используют каскад сверточных слоев и слоев максимального объединения (которые представляют собой тип нелинейности), достигают точности намного выше 90% на наборах данных CiFAR-10. Использование последовательных и объединяющих слоев приводит к резкому уменьшению размера обучаемых параметров, что, в свою очередь, приводит к лучшей производительности.

Ухудшение производительности из-за неисправностей и шумов

Различные виды неисправностей и шумов, возможные в DON, обсуждались ранее. В этом подразделе относительная важность (см. Уравнение 8), по сравнению с регулярно обученным DON (оранжевый). (i) yshift = 0 и (ii) yshift = 0,1 × P. В Фазовые маски ДОНов, обученных с различными значениями λt. Рассмотрены 9 различных видов неисправностей и шумов. Сначала обучается DON с одной программируемой фазовой пластиной, применяются отдельные неисправности и шумы в различной степени, и регистрируется точность классификации на испытательном наборе. Когда в приложении сбоя или шума используется случайное число, выполняется несколько прогонов для записи статистики снижения производительности. Было обнаружено, что во всех случаях 100 прогонов было достаточно для количественной оценки статистических вариаций. Во втором эксперименте комбинации неисправностей и шумов были рассмотрены для трех различных архитектур DON (с 1, 2 и 3 программируемыми фазовыми пластинами соответственно). Во-первых, комбинация фазовых ошибок, связанных с округлением и допуском, рассматривается как 3A для DON, который состоит из одной линейной фазовой пластины. Рассмотрены четыре уровня дискретизации фазы, которые комбинируются с четырьмя различными значениями стандартного отклонения по фазе (дискретная и непрерывная фазовые ошибки). Можно заметить, что с уменьшением уровней дискретизации точность классификации постепенно снижается. Однако уровни дискретизации 100 и 1000 показывают такой же уровень точности, как и DON, обученный с точностью до плавающей точки. Таким образом, 100 уровней фазовой дискретизации кажутся подходящим уровнем. Со всеми уровнями дискретизации добавляются случайные фазовые ошибки со средним значением 0 и стандартным отклонением в диапазоне 1,25%, 2,5%, 5% и 10%, а точность статистической классификации наносится на график. Можно заметить, что при фазовой ошибке 5% (18 °) и 10% (36 °) можно ожидать серьезного ухудшения точности, вплоть до 27% и 17% со-

ответственно. Наряду с фазовыми ошибками, битые пиксели могут быть источником неисправности в сети.

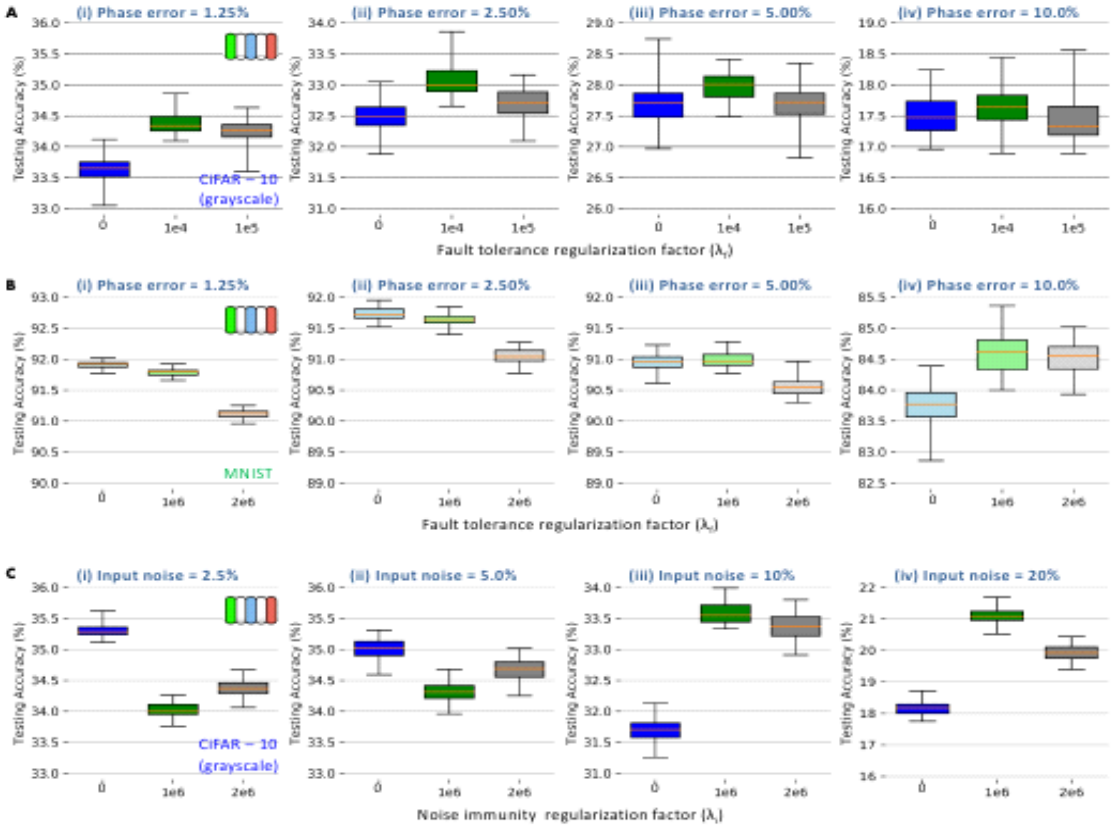


Рис. 5. Производительность моделей, обученных с использованием предложенного штрафа за регуляризацию (см. Уравнение 8). A Включен только член регулирования отказоустойчивости с тремя значениями параметра λf. Статистика классификации испытаний для четырех различных степеней фазовой ошибки показана в (i), (ii), (iii) и (iv) соответственно. B Эксперимент, аналогичный эксперименту A, проведенному с набором данных MNIST. C Включен только член регуляризации помехоустойчивости с тремя значениями параметра λi. Статистика классификации испытаний для четырех различных степеней шума входной интенсивности показана в (i), (ii), (iii) и (iv) соответственно.

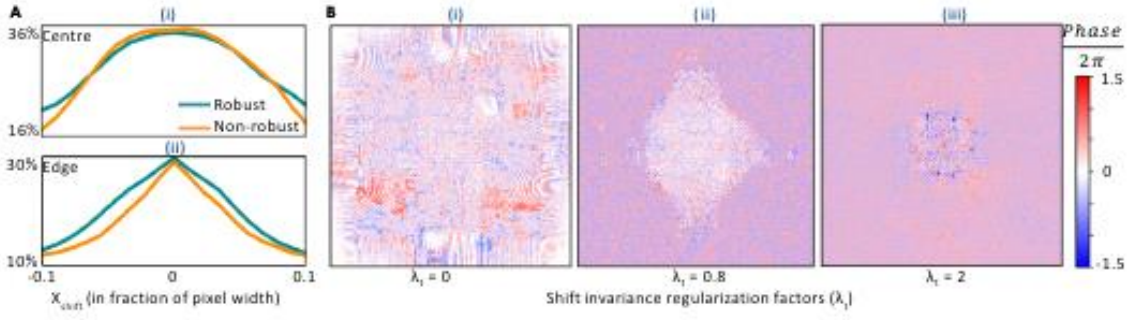


Рис. 6. Производительность модели, обученной с включенным членом регуляризации инвариантности сдвига (λt = 0,8)

Чтобы получить чувствительность точности по отношению к битым пикселям, выбираются случайные пиксели в линейных фазовых пластинах, и фазы для этих пикселей отключаются. Доли 1,25%, 2,5%, 5% и 10% от общего числа пикселей считаются битыми пикселями и равномерно распределяются случайным образом среди 256 × 256 пикселей фазовой пластины. Средняя точность для всех этих случаев составляет 34%, 33%, 31,5% и 28% соответственно. Исследование дисперсии точности классификации из-за битых пикселей выполняется в сочетании с константами затухания 0,8 и 0,9, как показано на рисунке 3B. Можно заметить, что точность классификации не очень чувствительна к случайному затуханию на фазовых пластинах; однако битые пиксели сильно снижают точность: 10% битых пикселей показывают среднюю точность 27%. Однако случайное затухание будет иметь значение при наличии шума частичных разрядов, поскольку они вносят вклад в ухудшение системного SNR. На рисунке 3D показана деградация точности классификации в DON с одной программируемой фазовой пластиной, когда фазовая пластина может быть не совмещена с входной плоскостью. Пластина перемещается в поперечной плоскости с шагом 0,01 ∗ P для всех возможных комбинаций сдвига x и y, и отображается график точности тепловой карты для каждой комбинации. Видно снижение точности (≈ 12%) при несовпадении в диапазоне ± 0,1 × ширины пикселя. Видно, что существует резкое ухудшение точности сдвигов субпикселей. Это происходит из-за того, что фазы двух соседних пикселей могут сильно различаться. Чувствительность точности классификации DON по отношению к шуму входной интенсивности и фототоковому шуму суммированы на рисунке 3C. Входной шум моделируется путем добавления случайной интенсивности поля к каждому из элементов поля в данных тестирования. Характеристики DON были изучены для 2,50%, 5%, 10% и 20% входного шума, и статистические данные деградации нанесены на график. Подобно входным шумам, шум фотодиода можно аппроксимировать добавлением случайного произвольного фототока на конце слоя фотодиода. Влияние шума частичного разряда нанесено на график для четырех значений шума частичного разряда (cnoise / Imax): 1,56, 3,12, 6,25 и 12,5. Точность классификации очень чувствительна как к входным шумам, так и к шумам фотодиодов. Суммарное влияние всех неисправностей и шумов на точность классификации рассматривается далее для ДОН с одним, двумя и тремя обучаемыми фазовыми слоями; протестировано на данных шкалы серого CiFAR-10 (рисунок 3E) и MNIST (рисунок 3E) соответственно. Комбинация неисправностей и шумов, рассматриваемая в этом эксперименте, следующая: 100 уровней дискретизации фазы, фазовая ошибка 2,5%, затухание 0,9, битые пиксели 2,5%, входной шум 10% и шум частичного разряда 1,56 с рассогласованием в

пределах ± 5% от ширина пикселя. Видно, что снижение производительности не зависит от сложности сети. Другими словами, наличие большего количества обучаемых параметров, кажется, придает некоторую устойчивость. Хотя учитываются коррелированные ошибки, не похоже, что ухудшение ухудшается при использовании нескольких фазовых пластин. Более того, при той же степени неисправностей и шумов сеть с более высокой производительностью в идеальных условиях будет продолжать работать лучше, чем ее менее производительный аналог. Мы также обнаружили, что использование разреженных категориальных потерь кроссэнтропии в обучении DON приводит к большему разбросу точности по сравнению с использованием категориальных потерь кроссэнтропии; заметной разницы в средней точности между двумя случаями не наблюдалось.

Производительность надежно обученного DON

В этом разделе для целей сравнения показана статистика регулярно обученного DON (не робастного) (под регулярно обученным мы подразумеваем DON, обученный только категориальным потерям кроссэнтропии без каких-либо условий регуляризации или ошибок. / шумовая инжекция). Обученные веса моделей, обученных по-разному, сначала подвергаются случайному взвешиванию одинаковой величины для 100 испытаний, а статистика точности классификации на тестовых данных набора данных CiFAR-10 сравнивается между моделями.

C.1. Индуцирование неисправности и внесение шума

Сначала представлены характеристики DON, обученных с помощью индукции неисправности (рисунок 4 A) и инжекции шума (рисунок 4 B). В первом эксперименте были оценены две стратегии возникновения неисправности. Одна из стратегий состоит в добавлении гауссовского случайного шума с нулевым средним на каждой итерации (модель A на рисунке 4 A). Во время обучения индукции неисправностей 2,5% случайной фазовой ошибки добавляются к каждой обучаемой фазе для разработки DON. Вторая стратегия была принята для модели C, где разломы вводились через каждые 5 эпох, позволяя модели изучать внутренние разломы в обучаемых слоях. Хотя во время обучения моделей B и C была достигнута хорошая точность обучения и тестирования, результаты тестирования модели B хуже, чем у модели A, а модель C немного превосходит модель A по большим фазовым ошибкам. Аналогичные исследования проводятся путем введения входных шумов во время обучения для достижения невосприимчивости к снижению производительности из-за шумов во время логического вывода. Каждая из моделей B и C обучается с использованием случайного гауссовского шума с фа-

зовой ошибкой 5% (нулевое среднее). Модели B и C различаются по частоте смены шумов (каждую эпоху для B, каждые 5 эпох для C). Хорошая точность обучения и тестирования была достигнута во время обучения моделей B и C. В отличие от модели, вызванной неисправностью, описанной выше, модель B и модель C незначительно превосходят A по обоим низким уровням шума. Модель C превосходит A по большим уровням шума при испытаниях. Обнаружено, что обучение наведению отказов и введению шума обеспечивает лишь незначительные улучшения для больших степеней отказов и уровней шума. Неэффективность обучения введению неисправности / шума 10 может быть объяснена тем фактом, что фаза может изменяться только в фиксированном интервале (от 0 до 2π), который ограничивает динамический диапазон. Сниженная эффективность обучения введению неисправностей также наблюдалась в аналоговых электронных нейронных двигателях с ограниченным динамическим диапазоном обучаемых весов [49]. Еще один недостаток обучения с использованием инъекции сбоя / шума заключается в том, что он удлиняет тренировочный процесс.

C.2. Добавление членов регуляризации

Во втором подходе цель обучения модифицируется путем изменения функции потерь на категориальную кросс-энтропийную потерю вместе с членами регуляризации (см. Уравнение 8). Рисунок 5 суммирует производительность моделей, обученных с предложенными штрафами за регуляцию. Во всех случаях сравнения производительности представлены для двух разных DON (модели B и C), обученных с двумя разными значениями соответствующих факторов регуляризации, а также для регулярно обучаемого DON (модель A). Рисунок 5 Коэффициенты регуляризации сбоя λf, равные 1e4 и 1e5, учитываются для обучающих моделей B и C. Характеристики с точки зрения чувствительности к сбоям фазы 1,25%, 2,5%, 5% и 10% показывают, что лучшая медианная точность наряду с уменьшением Для модели B наблюдается разброс на всех уровнях разломов. Модель C превосходит модель A, но уступает B. Эксперимент повторяется на данных MNIST с коэффициентами регуляризации, сохраненными на уровне 1e6 и 2e6 для моделей B и C (см. Рисунок 5 B. Можно заметить, что только при высокой фазовой ошибке (5 % и 10% в эксперименте) модели B и C превосходят модель A. Для следующего эксперимента параметр регуляризации помехоустойчивости (λi) был включен и установлен на 1e6 и 2e6 для моделей B и C соответственно. Точность тестирования обученных моделей для входного шума 2,5%, 5%, 10% и 20% суммирована на рисунке 5C. В этом эксперименте также видно, что модели B и C значительно превосходят модель A для более высоких уровней входного шума, в то время как неэффективно

для более низких уровней шума. Наконец, эффективность члена регуляризации инвариантности сдвига оценивается на рисунке 6 A. Модель, обученная члену регуляризации инвариантности сдвига, как видно, обеспечивает незначительное улучшение по сравнению с обычной моделью. фазовые маски показывают характерную однородность, когда применяется регуляризация. Общее наблюдение состоит в том, что большие значения параметра регуляризации λf и λi обеспечивают лучшую устойчивость при более высоких степенях сбоя / шума, но также приводят к снижению производительности. Четкое представление о диапазоне неисправностей / шумов в оборудовании поможет в оптимальном выборе этих факторов регуляризации. Кроме того, подход регуляризации не усложняет и не удлиняет процесс обучения, как подход к обучению с добавлением ошибок / шума.

ЗАКЛЮЧЕНИЕ

В итоге было представлено влияние отказов и шумов на ухудшение качества работы DON и режим тренировки для смягчения этих эффектов. Результаты численного исследования дают несколько идей для руководства экспериментальной деятельностью. Во-первых, выяснилось, что DON хорошо работает только с простыми наборами данных. Чтобы DON могли конкурировать с современными CNN, необходимо реализовать нелинейные операции, такие как объединение пикселей. Для более простых наборов данных нехватка высоконелинейных оптических материалов не может быть серьезным препятствием. Для более сложных задач наличие материалов с высокой оптической нелинейностью может оказаться недостаточным для улучшения рабочих характеристик. Мы обнаружили, что DON очень чувствительны к сбоям и шумам, а деградация, связанная с сбоями / шумом, более серьезна для задач, в которых DON не могут достичь адекватной производительности в идеальном случае. Мы показываем, что соответствующий режим тренировок может в определенной степени уменьшить проблему чувствительности. Рекомендуется выполнять обучение с несколькими инициализациями и коэффициентами регуляризации и четкой оценкой ожидаемых уровней сбоев и шума в оборудовании, чтобы получить вес, установленный перед загрузкой на оптическое оборудование. Наша работа была ограничена скалярной теорией дифракции, но этот подход мог быть полезен в обучаемых средах, использующих каскадные метаповерхности [37, 59]. Дальнейшие расширения этой работы могут рассмотреть улучшение комбинаций неисправностей и шумов и их взаимодействие с сетевой архитектурой и оптической нелинейностью. Можно изучить сочетание терминов «внесение неисправностей / шумов» и объективных штрафов.

ФИНАНСИРОВАНИЕ

Проект наномиссий, Департамент науки и технологий, правительство Индии (SN / NM / NS-65/2016).

Раскрытие информации.

Авторы заявляют об отсутствии конфликта интересов.

REFERENCES

1. K. Berggren, Q. Xia, K. K. Likharev, D. B. Strukov, H. Jiang, T. Mikolajick, D. Querlioz, M. Salinga, J. R. Erickson, S. Pi, F. Xiong, P. Lin, C. Li, Y. Chen, S. Xiong, B. D. Hoskins, M. W. Daniels, A. Madhavan, J. A. Liddle, J. J. McClelland, Y. Yang, J. Rupp, S. S. Nonnenmann, K.-T. Cheng, N. Gong, M. A. Lastras-Montaño, A. A. Talin, A. Salleo, B. J. Shastri, T. F. de Lima, P. Prucnal, A. N. Tait, Y. Shen, H. Meng, C. Roques-Carmes, Z. Cheng, H. Bhaskaran, D. Jariwala, H. Wang, J. M. Shainline, K. Segall, J. J. Yang, K. Roy, S. Datta, and A. Raychowdhury, "Roadmap on emerging hardware and technology for machine learning," Nanotechnology 32, 012002 (2021).
2. S. Foo, L. Anderson, and Y. Takefuji, "Analog components for the VLSI of neural networks," IEEE Circuits Devices Mag. 6, 18–26 (1990).
3. Q. Yang, X. Luo, P. Li, T. Miyazaki, and X. Wang, "Computation offloading for fast CNN inference in edge computing," in Proceedings of the Conference on Research in Adaptive and Convergent Systems, (ACM, Chongqing China, 2019), pp. 101–106.
4. H. Caulfield, J. Kinser, and S. Rogers, "Optical neural networks," Proc. IEEE 77, 1573–1583 (1989).
5. L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic Neural Networks: A Survey," IEEE Access 7, 175827–175841 (2019).
6. S. Jutamulia and F. T. S. Yu, "Overview of hybrid optical neural networks," Opt. & Laser Technol. 28, 59–72 (1996).
7. Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," Light. Sci. & Appl. 8, 42 (2019).
8. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Solving computer vision tasks with diffractive neural networks," in Optoelectronic Imaging and Multimedia Technology VI, vol. 11187 (International Society for Optics and Photonics, 2019), p. 111870T.
9. Y. Zheng and M. S. Asif, "Joint Image and Depth Estimation with Mask-Based Lensless Cameras," arXiv:1910.02526 [cs, eess] (2019).
10. H.-Y. S. Li, Y. Qiao, and D. Psaltis, "Optical network for real-time face recognition," Appl. Opt. 32, 5026 (1993).
11. S. Xu, X. Zou, B. Ma, J. Chen, L. Yu, and W. Zou, "Analog-to-digital conversion revolutionized by deep learning," arXiv:1810.08906 [physics] (2018).
12. P. Ambs, "Optical Computing: A 60-Year Adventure," Adv. Opt. Technol. 2010, 1–15 (2010).

13. X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu, “A Review of Optical Neural Networks,” IEEE Access 8, 70773–70783 (2020).
14. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, “Deep Learning with Coherent Nanophotonic Circuits,” Nat. Photonics 11, 441–446 (2017).
15. T. F. de Lima, H.-T. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, “Machine Learning With Neuromorphic Photonics,” J. Light. Technol. 37, 1515–1534 (2019).
16. N. U. Dinc, D. Psaltis, and D. Brunner, “Optical Neural Networks: The 3D connection,” arXiv:2008.12605 [cs] (2020). 11
17. S. Maktoobi, L. Froehly, L. Andreoli, X. Porte, M. Jacquot, L. Larger, and D. Brunner, “Diffractive Coupling For Photonic Networks: How Big Can We Go?” IEEE J. Sel. Top. Quantum Electron. 26, 1–8 (2020).
18. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, “All-optical machine learning using diffractive deep neural networks,” Science. 361, 1004–1008 (2018).
19. Y. Gao, S. Jiao, J. Fang, T. Lei, Z. Xie, and X. Yuan, “Multiple-image encryption and hiding with an optical diffractive neural network,” Opt. Commun. 463, 125476 (2020).
20. H. Dou, Y. Deng, T. Yan, H. Wu, X. Lin, and Q. Dai, “Residual D 2 NN: Training diffractive deep neural networks via learnable light shortcuts,” Opt. Lett. 45, 2688 (2020).
21. C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, and H. Chen, “Performing optical logic operations by a diffractive neural network,” Light. Sci. & Appl. 9, 1–7 (2020).
22. D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, “Analysis of Diffractive Optical Neural Networks and Their Integration With Electronic Neural Networks,” IEEE J. Sel. Top. Quantum Electron. 26, 1–14 (2020).
23. Y. Chen and J. Zhu, “An optical diffractive deep neural network with multiple frequency-channels,” arXiv:1912.10730 [physics, stat] (2019).
24. J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, “Class-specific differential detection in diffractive optical neural networks improves inference accuracy,” Adv. Photonics 1, 046001 (2019).
25. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, “Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification,” Sci. Reports 8, 1–10 (2018).
26. S. Colburn, Y. Chu, E. Shilzerman, and A. Majumdar, “Optical frontend for a convolutional neural network,” Appl. Opt. 58, 3179–3186 (2019).
27. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, “Fourier-space Diffractive Deep Neural Network,” Phys. Rev. Lett. 123, 023901 (2019).
28. K. Wagner and D. Psaltis, “Multilayer optical learning networks,” Appl. Opt. 26, 5061–5076 (1987).
29. A. D. Fisher, W. L. Lippincott, and J. N. Lee, “Optical implementations

of associative networks with versatile adaptive learning capabilities," Appl. Opt. 26, 5039–5054 (1987).
30. D. Psaltis and N. Farhat, "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," Opt. Lett. 10, 98 (1985).
31. J. Steck, S. Skinner, A. Cruz-Cabrera, M. Yang, and E. Behrman, "Backpropagation training of an optical neural network," in Proceedings of the Seventh International Conference on Microelectronics for Neural, Fuzzy and Bio-Inspired Systems, (1999), pp. 346–351.
32. E. Khoram, A. Chen, D. Liu, L. Ying, Q. Wang, M. Yuan, and Z. Yu, "Nanophotonic media for artificial neural inference," Photonics Res. 7, 823–827 (2019).
33. P. Camayd-Muñoz, C. Ballew, G. Roberts, and A. Faraon, "Multifunctional volumetric meta-optics for color and polarization image sensors," Optica. 7, 280–283 (2020).
34. X. Luo, "Engineering Optics 2.0: A Revolution in Optical Materials, Devices, and Systems," ACS Photonics 5, 4724–4738 (2018).
35. L. Wan, D. Pan, S. Yang, W. Zhang, A. A. Potapov, X. Wu, W. Liu, T. Feng, and Z. Li, "Optical analog computing of spatial differentiation and edge detection with dielectric metasurfaces," Opt. Lett. 45, 2070 (2020).
36. Z. Wu, M. Zhou, E. Khoram, B. Liu, and Z. Yu, "Neuromorphic metasurface," Photonics Res. 8, 46–50 (2020).
37. N. Mohammadi Estakhri, B. Edwards, and N. Engheta, "Inverse-designed metastructures that solve equations," Science. 363, 1333–1338 (2019).
38. M. Hermans, M. Burm, T. V. Vaerenbergh, J. Dambre, and P. Bienstman, "Trainable hardware for dynamical computing using error backpropagation through physical media," Nat. Commun. 6, 1–8 (2015).
39. T. Zhou, L. Fang, T. Yan, J. Wu, Y. Li, J. Fan, H. Wu, X. Lin, and Q. Dai, "In situ optical backpropagation training of diffractive optical neural networks," Photonics Res. 8, 940–953 (2020).
40. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation," Optica. 5, 864 (2018).
41. N. Semenova, X. Porte, L. Andreoli, M. Jacquot, L. Larger, and D. Brunner, "Fundamental aspects of noise in analog-hardware neural networks," Chaos: An Interdiscip. J. Nonlinear Sci. MACL2020, 103128 (2019).
42. M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," Opt. Express 27, 14009 (2019).
43. R. Burgwal, W. R. Clements, D. H. Smith, J. C. Gates, W. S. Kolthammer, J. J. Renema, and I. A. Walmsley, "Implementing random unitaries in an imperfect photonic network," arXiv:1704.01945 [quant-ph] (2017).
44. D. Mengu, Y. Rivenson, and A. Ozcan, "Scale-, shift- and rotation-

invariant diffractive optical networks," arXiv:2010.12747 [physics] (2020).
45. J. Shi, "A Diffractive Neural Network with Weight-Noise-Injection Training," arXiv:2006.04462 [cs, eess] (2020).
46. J. Bernier, J. Ortega, E. Vidal, I. Rojas, and A. Prieto, "A Quantitative Study of Fault Tolerance, Noise Immunity, and Generalization Ability of MLPs," Neural Comput. 12, 2941–2964 (2001).
47. P. Chandra and Y. Singh, "Fault tolerance of feedforward artificial neural networks- a framework of study," in Proceedings of the International Joint Conference on Neural Networks, 2003., vol. 1 (2003), pp. 489–494 vol.1.
48. X. Zeng and D. Yeung, "Sensitivity analysis of multilayer perceptron to input and weight perturbations," IEEE Transactions on Neural Networks 12, 1358–1366 (2001).
49. P. Edwards and A. Murray, "Fault tolerance via weight noise in analog VLSI implementations of MLPs-a case study with EPSILON," IEEE Transactions on Circuits Syst. II: Analog. Digit. Signal Process. 45, 1255–1262 (1998).
50. A. Murray and P. Edwards, "Enhanced MLP performance and fault tolerance resulting from synaptic weight noise during training," IEEE Transactions on Neural Networks 5, 792–802 (Sept./1994).
51. P. J. Edwards and A. F. Murray, "Penalty terms for fault tolerance," in Proceedings of International Conference on Neural Networks (ICNN'97), vol. 2 (1997), pp. 943–947.
52. J. W. Goodman, Introduction to Fourier Optics (Roberts and Company Publishers, 2005).
53. R. S. Hegde, "Digital electronic neural networks with analog nanophotonic frontends: A numerical study," in Nanophotonics and Micro/Nano Optics VI, vol. 11556 (International Society for Optics and Photonics, 2020), p. 115560I.
54. F. Chollet, "Keras-team/keras," Keras (2020).
55. R. Hegde, "Rshegde / opticaldiffnets — Bitbucket," https://bitbucket.org/rshegde/opticaldiffnets (2020).

ПОЛНОСТЬЮ ОПТИЧЕСКАЯ НЕЛИНЕЙНАЯ ФУНКЦИЯ АКТИВАЦИИ ДЛЯ НЕЙРОМОРФНЫХ ФОТОННЫХ ВЫЧИСЛЕНИЙ С ИСПОЛЬЗОВАНИЕМ ПОЛУПРОВОДНИКОВЫХ ЛАЗЕРОВ ФАНО

Расмуссен Т.С., Йи Ю., Морк Д.

2020 Optical Society of America https://doi.org/10.1364/OL.395235

Мы прогнозируем, что полупроводниковые лазеры Фано могут быть использованы для реализации полностью оптической нелинейной функции активации для нейроморфных фотонных вычислений. Используя оптическое управление зеркалом Фано, лазер может генерировать оптические импульсы с низкой пороговой энергией, частотой повторения гигагерц и подавлением на порядки величины между включенным и выключенным состояниями. Аналитические оценки пороговой энергии переключения, коэффициента экстинкции и периода рефрактерности хорошо согласуются с численными результатами.

По мере того, как цифровая электроника приближается к фундаментальному препятствию эффективности, все чаще используются альтернативные специализированные компьютерные стратегии [1]. К ним относятся нейронные сети для глубокого обучения с приложениями, например, для обработки данных в реальном времени, компьютерного зрения и обработки естественного языка. Реализация нейронных сетей с использованием архитектур на основе фотоники обещает увеличить скорость вычислений на порядки, одновременно получая отличную энергоэффективность [1]. Искусственные нейроны, составляющие сети, состоят из двух основных строительных блоков: функции линейного взвешивания и интегрирования (умножение и накопление) для сбора входных данных от других нейронов и функции нелинейной активации (NLAF) для определения порога [Рис. 1 (a)]. Последний определяет, срабатывает ли нейрон или нет, посредством сигмовидной реакции с подходящим рефрактерным периодом и генерацией выходного импульса [2]. Эти устройства должны быть одновременно компактными, энергоэффективными, быстрыми и подходящими для масштабируемой интеграции, что создает сложную задачу [3]. В последнее время появился ряд предложений по реализации NLAF, включая использование модуляторов электропоглощения [4] и модуляторов микрокольца [5], а также ряда полностью оптических реализаций [6–8]. Альтернативным вариантом является использование возбуждаемых полупроводниковых лазеров [1], для которых [1]. [2] содержит обширный обзор. В этой работе мы демонстрируем, как особый тип микроскопического лазера, так называемый лазер Фано [9], демонстрирует возбудимость и может использо-

ваться для реализации NLAF для нейроморфных фотонных вычислений, выполняя требования пороговых значений (сигмоидных отклик), обладающий рефрактерным периодом наносекунды и генерирующий оптические импульсы в ответ на входные импульсы с низким энергопотреблением и чрезвычайно малой площадью основания, менее 100 мкм2. Ключевой характеристикой лазера Фано является то, что одно или оба лазерных зеркала реализованы за счет интерференции Фано за счет взаимодействия континуума волноводных мод и близлежащего дискретного резонанса [10] [рис. 1 (б)].

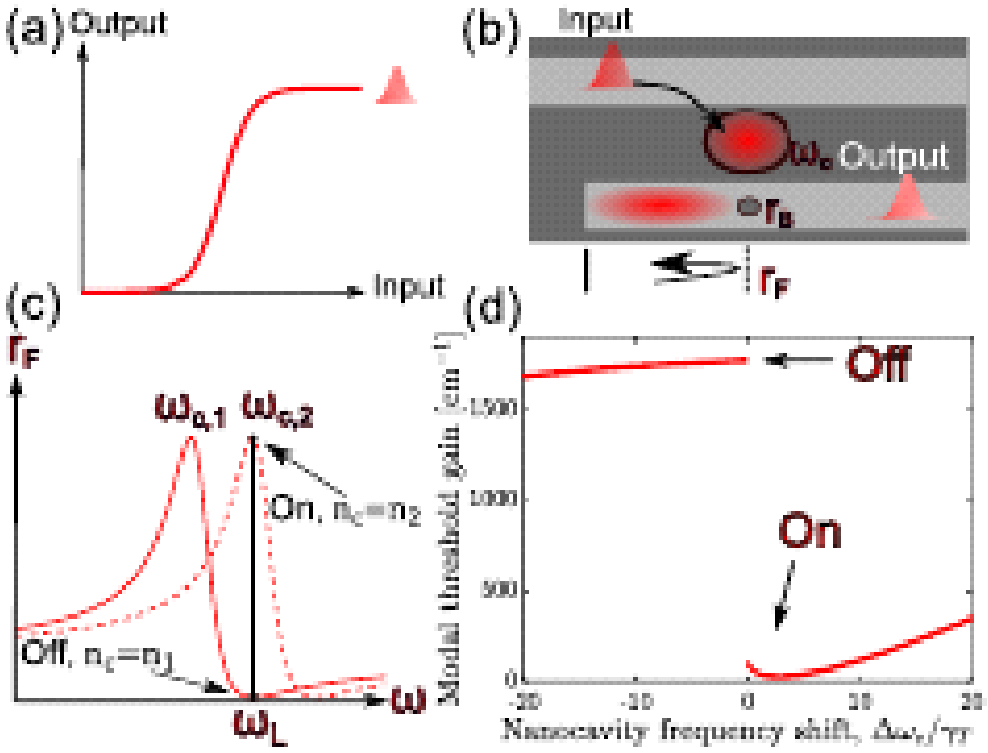


Рис. 1. (a) Схематическое изображение сигмовидной нелинейной функции активации. (b) Схема лазера Fano, работающего как NLAF с портами ввода и вывода. Правое лазерное зеркало rF образовано резонансом Фано из-за нанополости. (c) Коэффициент отражения Фано как функция частоты, демонстрирующий настройку между включенным и выключенным состояниями путем изменения показателя преломления нанополости с n1 на n2. (d) Модальное пороговое усиление лазера Fano как функция резонансной частоты нанополости, показывая выключенное и включенное состояния, соответствующие (c).

Это приводит к появлению узкополосного пика отражательной способности вблизи резонансной частоты дефектной нанополости [рис. 1 (в)] с характерной асимметричной формой Фано, образующей лазерное зеркало. Помещая активный материал в волновод между левым зеркалом и полостью дефекта, используя технологию скрытой гетероструктуры [11], формируется лазер, как показано ограниченным полем (красным) на рис. 1 (b). Это устройство ранее было реализовано на платформе фотонного кри-

сталла, демонстрируя одномодовую генерацию и самопульсирование [12,13], а также предсказания ширины полосы частотной модуляции выше 1 ТГц [14], а также исключительную устойчивость к внешней оптической обратной связи. [15]. Ref. [16] дает всесторонний обзор Фано лазерные исследования. Отражательная способность зеркала Фано равна [16]

$$r_F(\omega_c, \omega_L) = r_B + (it_B - r_B)\frac{\gamma_c}{i(\omega_c - \omega_L) + \gamma_T}. \quad (1)$$

Здесь rB и tB – недисперсионные коэффициенты отражения и пропускания частично передающего элемента (PTE) в волноводе [Рис. 1 (b)], γc – скорость связи волновода с нанорезонатором, γT – полная скорость затухания поля нанорезонатора, ωC – (настраиваемая) резонансная частота нанорезонатора, а ωL – частота входящего поля.

При rB > 0 отражательная способность характеризуется небольшим частотным разделением между ее экстремумами, что идеально подходит для применений в коммутации [17] и генерации импульсов. В частности, узкополосная отражательная способность и высокодисперсная фаза приводят к сильной частотной зависимости модального порогового усиления [9,14] [рис. 1 (г)]. Состояния с большим пороговым разделением усиления (включенное и выключенное состояния) близки по частоте и могут переключаться между ними путем настройки частоты нанополости. Это позволяет генерировать короткие оптические импульсы по схеме активной модуляции добротности [18]. Если резонанс изначально находится в выключенном состоянии, пороговое усиление слишком велико, чтобы устройство могло генерировать генерацию, а это означает, что большая инверсия населенностей может быть создана путем непрерывной накачки. Затем лазер переключается во включенное состояние путем динамической настройки резонансной частоты нанополости, при этом инверсия населенности сильно превышает пороговую инверсию. Чрезмерное усиление приводит к всплеску скорости стимулированного излучения, и излучается сильный импульс, истощающий плотность носителей. После излучения импульса резонанс нанополости возвращается в выключенное состояние, и плотность носителей в активной области восстанавливается. Динамическая настройка показателя преломления в нанополости и, следовательно, ее резонансной частоты может быть выполнена с использованием, например, электрооптического эффекта, эффекта Керра или плазменного эффекта, обусловленного возбуждением свободных носителей. Первые два требуют более сложной геометрии, такой как электроды [19] или бабочки для ограничения экстремального поля [20], поэтому в этой работе мы изучаем лазер Фано, настраиваемый на дисперсию носителей и заполнение зон в нанополости. Носители

возбуждаются за счет двухфотонного поглощения внешнего оптического импульса, как, например, в [5,11]. [17,21]. Динамика лазера исследуется численно с использованием нашей ранее разработанной модели лазера Фано во временной области [16], но в итерационной и дискретизированной версии, как представлено в [16]. [22] [Ур. (2), (16) и (17) в нем]. Нелинейности в нанополости реализованы с помощью формул. (3) и (10) статьи. [23] и параметры в нем, но с единственной постоянной времени τc для носителей в нанорезонаторе. Внешний запускающий импульс входит непосредственно как возбуждение в уравнение для плотности носителей в нанополости через скорость генерации G (t), представляя либо возбуждение посредством импульса, вводимого вертикально над мембраной, либо возбуждение с боковой связью через связь в плоскости волновод на другой стороне нанополости, т.е. красный входной импульс на рис. 1 (a), с той лишь разницей, что эффективность возбуждения. Используются реалистичные значения параметров, основанные на наших ранее изготовленных устройствах [12], но со скрытой активной областью квантовой ямы гетероструктуры в волноводе [11] вместо квантовых точек по всему образцу. Параметры в обозначениях [5]. [16]: α = 3, 嵒 = 0,12, vg = c / ng, ng = 3,17, g N = 1,3 · 10−18 м2, N0 = 1 · 1024 м − 3, rB = 0,4, τin = 270 фс, γc = 6 × 1011 с − 1, γT = 6.06 × 1011 с − 1, Rp = 3Rp, th, τs = τc = 0,28 нс. На рис. 2 (a) показан пример входного возбуждения нанополости и соответствующего выхода лазера. Понятно, что существует порог, за которым генерируются хорошо определенные оптические импульсы. Это пример собственной возбудимости [2] лазера Фано и обеспечивает пороговый механизм NLAF. На рисунке 2 (b) представлено более систематическое исследование, показывающее энергию выходного импульса (цветовую шкалу) как функцию ширины входного импульса и энергии импульса при работе при Rp = 3Rp, th, где Rp, th – порог лазерного излучения. скорость накачки в открытом состоянии при минимальном пороговом усилении на рис. 1 (d), с накачкой либо оптической, либо электрической. Электрический пороговый ток будет в диапазоне микроампер из-за небольшого размера устройства, что приведет к низкому энергопотреблению. Красная кривая представляет собой аналитическую оценку энергии переключения, заданной уравнением. (3), демонстрируя отличное согласие. Это ясно демонстрирует желаемый сигмовидный отклик с только шумом спонтанного излучения ниже порога и выходными импульсами в диапазоне ≈20 фДж за порогом, что приводит к коэффициенту экстинкции на несколько порядков величины. Этот сигмовидный отклик на настройку резонанса нанополости присущ механизму модуляции добротности лазера Фано из-за большой разницы в пороговом усилении и эффективной отражательной способности между включенным и выключенным состояниями.

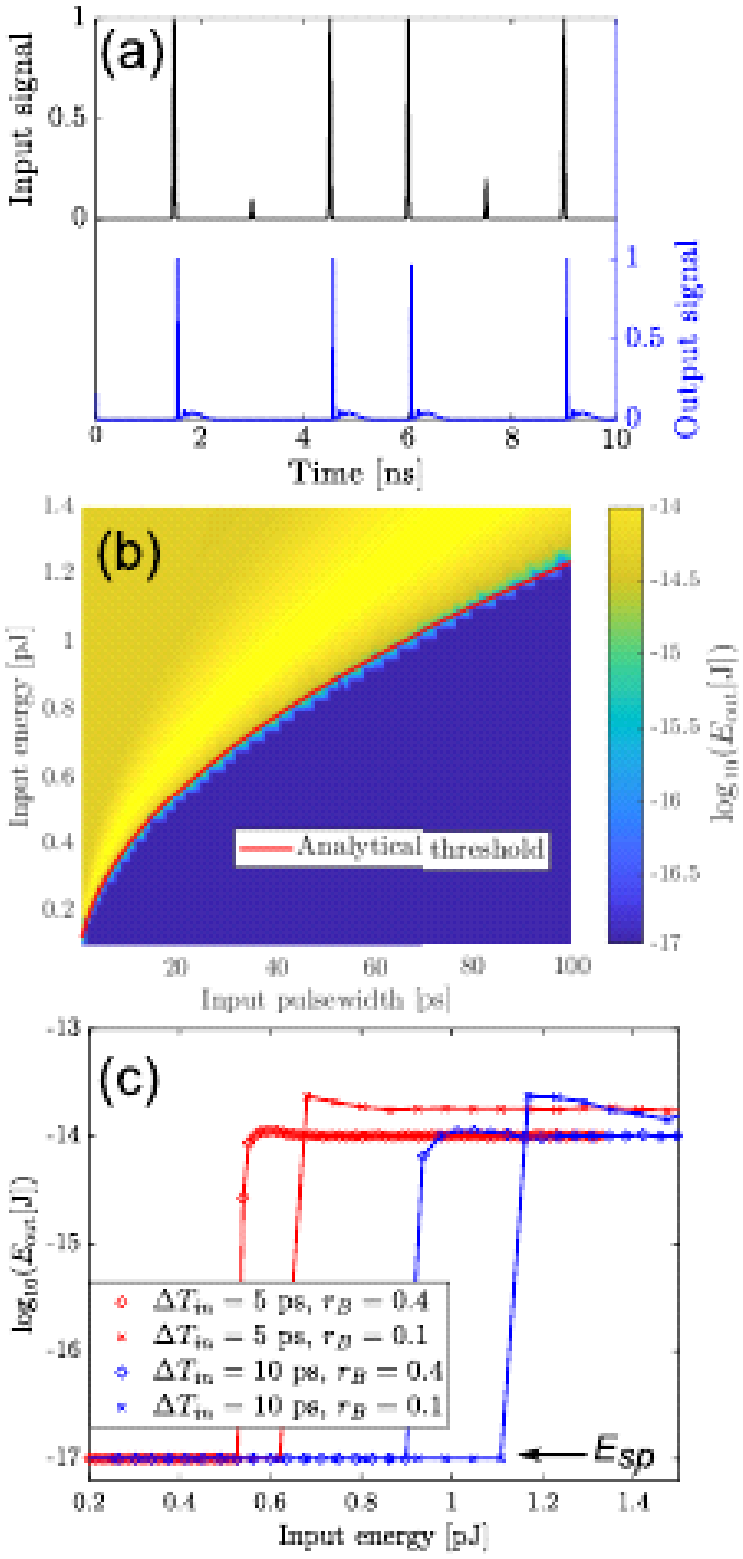


Рис. 2. (a) Расчетная динамика выходного сигнала для данного входного сигнала, показывающая генерацию импульсов и пороговое значение. (b) Энергия выходного импульса (цветовая шкала) как функция ширины входного импульса и энергии входного импульса. Красная кривая – это предсказание по формуле. (3). (c) Энергия выходного импульса как функция энергии входящего импульса для различной ширины входного импульса и отражательной способности PTE. Сплошные линии – ориентиры для глаз.

Если настройки резонанса недостаточно для переключения лазера из выключенного состояния во включенное состояние, свет не излучается, так как лазер не может достичь необходимого порогового усиления. Вместо этого достаточно сильная настройка резонанса означает, что лазер достигает включенного состояния [пересекает зазор на рис. 1 (d)] и излучает импульс. Свойства этого импульса определяются в первую очередь инверсией в закрытом состоянии, что означает, что после превышения порогового значения свойства импульса очень слабо изменяются в зависимости от входа, что приводит к быстрому плато в ответе. Эта зависимость от стационарного порогового усиления также дает возможность настраивать отклик, в частности, путем регулировки γT, rB и α, которые в первую очередь определяют резкость скачка порогового усиления.

На рисунке 2 (c) показана энергия выходного импульса как функция энергии входного импульса для различных постоянных значений ширины входного импульса. Пороговая входная энергия уменьшается с уменьшением ширины входного импульса, что связано с масштабированием эффективности двухфотонного поглощения с квадратом входной пиковой мощности. Видно, что все значения ширины импульса приводят к аналогичным кривым ввода-вывода с небольшими колебаниями амплитуды сразу за порогом. Причиной этого колебания является изменение эффективной отражательной способности во время генерации импульса, поскольку она сильно зависит от резонансной частоты относительно минимума порогового усиления. По мере увеличения входной энергии отклик выравнивается, потому что резонанс постоянно выходит за минимум, а затем высвобождает импульс во время распада носителей в нанополости. Настройка резонанса и энергия входящего импульса, необходимая для высвобождения импульса, в первом порядке определяются выражением

$$\Delta\omega_{r,th} \simeq \left(\gamma_T - \frac{r_B\gamma_T^2}{\gamma_s}\right)\left(\frac{\gamma_T\tau_{in}+1}{\gamma_T\tau_{in}}\right), \qquad (2)$$

$$E_{in,th} \simeq \sqrt{\Delta T\frac{\Delta\omega_{r,th}\gamma_S^2}{K_C G_{TPA}}}, \qquad (3)$$

где τin – время обхода лазера, KC – коэффициент дисперсии свободных носителей и заполнения полосы, GTPA – коэффициент двухфотонной генерации, γS – время, обратное времени хранения импульса возбуждения в нанополости, a,T – ширина входного импульса. (FWHM). Уравне-

ние (2) получается путем аппроксимации частоты и эффективной отражательной способности лазера как функции отстройки и сравнения с отражательной способностью в закрытом состоянии rB для оценки расстояния отстройки между включенным и выключенным состояниями. Он показывает, как можно снизить энергию триггера, уменьшив контраст между включенным и выключенным состояниями зеркала Фано, т.е. увеличив отражение ФЭП rB. Это также показывает, что полезно увеличить длину резонатора, так как это уменьшает продольный интервал между модами лазера, что, в свою очередь, снижает пороговую мощность переключения, с дополнительным преимуществом уменьшения порогового усиления лазера. Уравнение (3) получается путем преобразования порога настройки в энергию входного импульса, предполагая, что входной импульс гауссовский, и что высвобождение импульса намного быстрее, чем время жизни несущей. Максимально достижимый коэффициент экстинкции определяется выражением

$$\Delta E_{min} = \frac{N(g_{th,off})\hbar\omega_L V_{LC}}{E_{sp}}, \qquad (4)$$

где N – плотность несущих в закрытом состоянии, определяемая максимальным пороговым усилением gth, off : αi / ♀ + 1 / (2L ←) ln (1 / r 2B), которое, в свою очередь, определяется внутренним потери, αi, коэффициент ограничения поля, ᵖʟₑ, длина резонатора, L и rB. VLC – это объем активной области, а Esp – интегральная мощность спонтанного излучения за интервал обнаружения, при этом шумовой пол ниже порога переключения на рис. 2 (б) и 2 (с)

Уравнения (2) и (4) демонстрируют компромисс, присущий этому устройству, проиллюстрированный на рис. 2 (с), где и выходная энергия, и пороговая энергия увеличиваются с уменьшением rB. В общем, изменения, которые уменьшают пороговую энергию переключения, также имеют тенденцию к уменьшению степени экстинкции, например, увеличивая отражательную способность PTE rB или уменьшая межмодовый интервал по длине резонатора или ширине линии зеркала. Объяснение этому состоит в том, что достигаемое усиление в закрытом состоянии сильно зависит от расстояния между модами, поскольку достижение низкого порогового усиления для лазера Фано зависит от выравнивания резонанса нанополости с продольной модой, которая удовлетворяет фазовому условию двустороннего обхода. Собственное время восстановления (рефрактерный период) присутствует в схеме благодаря двум механизмам. Во-первых, механизм настройки основан на возбуждении свободных носителей в нанополости, которые распадаются со временем

жизни τc. Во-вторых, плотность носителей в резонаторе лазера сильно истощается во время высвобождения импульса и должна пополняться впоследствии, так что более медленный из двух процессов определяет фактическое время восстановления. Рисунки 3 (a) и 3 (b) демонстрируют это путем сравнения выходной динамики для входных возбуждений с различными задержками между импульсами запуска. Вторичный, близко расположенный импульс P2 дает только небольшой вторичный выходной импульс, потому что система еще не сбросила. Однако для большего интервала, показанного на P3, первый импульс почт

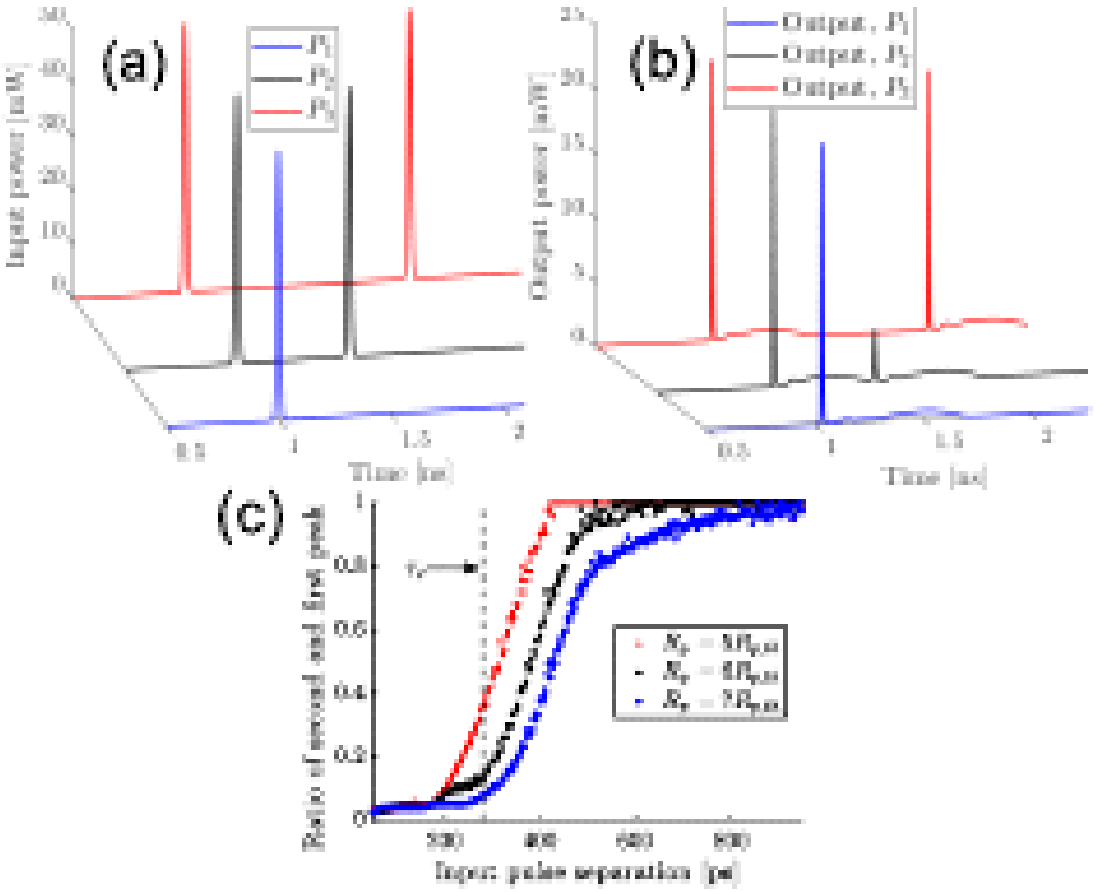


Рис. 3. (a) Три примера сигналов возбуждения P1, P2, P3. (b) Выходы лазера, соответствующие входным сигналам P1, P2, P3. (c) Отношение пиковых мощностей первого и второго импульсов в зависимости от разделения входных импульсов, демонстрирующее четкий рефрактерный период, ограниченный снизу τc (вертикальная пунктирная линия) и увеличивающийся с увеличением скорости накачки.

воспроизводится идентично, поскольку входной интервал приблизительно равен времени восстановления. Это поведение количественно определено более подробно на рис. 3 (c), который показывает соотношение вторичного пика и пика первого выходного импульса во временном сигнале как функцию временного разделения двух входных импульсов. Здесь отношение приближается к нулю для малых разносов, что соответствует, по существу, отсутствию вторичного выходного импульса,

пока в конечном итоге не приблизится к единице по мере увеличения расстояния, и два импульса не станут независимыми. Видно, что рефрактерный период увеличивается с увеличением скорости накачки, что может показаться нелогичным, но на самом деле также имеет место при обычном переключении добротности. Время восстановления 90% исходной плотности носителей, Ni, масштабируется приблизительно как t90 梮 τs log [10 (1 – Nf / Ni)], где τs – время жизни носителей L-резонатора. Начальная (в закрытом состоянии) плотность носителей, Ni, прямо пропорциональна скорости накачки, в то время как конечная (после высвобождения импульса) плотность носителей, Nf, уменьшается с увеличением скорости накачки, приблизительно подчиняясь обычным управляющим уравнениям активной модуляции добротности. [24]. По мере того как Nf уменьшается, а Ni увеличивается с увеличением скорости накачки, время восстановления, следовательно, также увеличивается, как показано на рис. 3 (c). Это означает, что также существует компромисс между максимальным коэффициентом гашения и минимизацией времени восстановления, и необходимо выбрать подходящую рабочую точку в зависимости от скорости и требований к энергии. Что касается энергопотребления и каскадности, текущая схема не допускает работу без усилителя, так как энергия выходного импульса примерно на один-два порядка ниже входной энергии. Однако неэффективное преобразование энергии связано с использованием двухфотонного поглощения для настройки резонатора и может быть значительно улучшено путем использования вместо этого линейного поглощения. Таким образом, усовершенствования в технологии нанопроизводства, позволяющие выращивать локализованную скрытую гетероструктуру с большей шириной запрещенной зоны внутри нанополости и, таким образом, возбуждение за счет линейного поглощения, значительно улучшат эффективность и снизят энергопотребление до пороговой энергии в десятки фемтоджоулей. В этом случае возможность каскадирования без усилителя в полностью оптической сети на кристалле на основе лазерных структур Fano была бы возможной, при условии, что подходящая система умножения-накопления может быть спроектирована внутри платформы, что приведет к крайней миниатюризации и сверхвысокой производительности. низкое энергопотребление. Также представляет интерес изучить приложения этой схемы просто как сверхкомпактный, низкоэнергетический и высокоскоростной NLAF для включения в существующие реализации на кристалле. Это по-прежнему требует интеграции лазеров на кремнии и эффективного соединения элементов многократного накопления с лазером Fano NLAF, что, вероятно, приведет к компромиссу между потреблением энергии и площадью устройства. В этом контексте отметим многообещающий прогресс интеграции фотонно-

кристаллических лазеров на кремнии [25,26]. Наконец, необходимы дальнейшие исследования гибридных электрооптических схем с электрическим управлением резонансом нанополости, поскольку это позволило бы включить функцию умножения-накопления просто путем введения интегрированных фотодетекторов. Учитывая заметный электрооптический эффект в фосфиде индия, это, вероятно, также обеспечит каскадность без усилителя, что значительно повысит потенциальную применимость.

В заключение, было показано, что полностью оптический NLAF для нейроморфной фотоники может быть реализован с использованием фотонно-кристаллических лазеров Фано. Лазерная система Fano по своей природе возбудима и имеет рефрактерный период, позволяющий работать в ГГц с подавлением на несколько порядков между включенным и выключенным состояниями. В то же время платформа фотонного кристалла позволяет сильно миниатюризировать компоненты, обеспечивая сверхмалые следы в диапазоне 100 мкм2 и малую энергию переключения в диапазоне 100 фДж. Таким образом, лазеры Фано являются многообещающими кандидатами на сверхбыстрые и компактные встроенные схемы нейроморфных фотонных вычислений.

Финансирование. Danmarks Grundforskningsfond (DNRF147); Виллум Фонден (8692); Европейский исследовательский совет H2020 (834410).

Раскрытие информации. Авторы заявляют об отсутствии конфликта интересов.

REFERENCES

1. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, Encyclopedia of Complexity and Systems Science (2018), pp. 1–37.
2. P. R. Prucnal, B. J. Shastri, T. F. de Lima, M. A. Nahmias, and A. N. Tait, Adv. Opt. Photon. 8, 228 (2016).
3. M. Miscuglio, G. C. Adam, D. Kuzum, and V. J. Sorger, APL Mater. 7, 100903 (2019).
4. R. Amin, J. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, T. El-Ghazawi, and V. J. Sorger, APL Mater. 7, 081112 (2019).
5. A. N. Tait, T. F. De Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Phys. Rev. Appl. 11, 064043 (2019).
6. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, Opt. Mater. Express 8, 3851 (2018).
7. G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, Opt. Express 27, 9620 (2019).
8. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu,

and S. Du, Optica 6, 1132 (2019).
9. J. Mork, Y. Chen, and M. Heuck, Phys. Rev. Lett. 113, 163901 (2014).
10. A. E. Miroshnichenko, S. Flach, and Y. S. Kivshar, Rev. Mod. Phys. 82, 2257 (2010).
11. A. Sakanas, Y. Yu, E. Semenova, L. Ottaviano, H. K. Sahoo, J. Mørk, and K. Yvind, in 2017 European Conference on Lasers and Electro-Optics and European Quantum Electronics Conference, (Optical Society of America, 2017), paper CB_5_3.
12. Y. Yu, W. Xue, E. Semenova, K. Yvind, and J. Mork, Nat. Photonics 11, 81 (2017).
13. T. S. Rasmussen, Y. Yu, and J. Mork, Laser Photon. Rev. 11, 1700089 (2017).
14. T. S. Rasmussen, Y. Yu, and J. Mork, Opt. Express 26, 16365 (2018).
15. T. S. Rasmussen, Y. Yu, and J. Mork, Phys. Rev. Lett. 123, 233904 (2019).
16. J. Mork, Y. Yu, T. S. Rasmussen, E. Semenova, and K. Yvind, IEEE J. Sel. Top. Quantum Electron. 25, 2900314 (2019).
17. D. Bekele, Y. Yu, K. Yvind, and J. Mork, Laser Photon. Rev. 13, 1900054 (2019).
18. W. G. Wagner and B. A. Lengyel, J. Appl. Phys. 34, 2040 (1963).
19. L.-D. Haret, X. Checoury, F. Bayle, N. Cazier, P. Boucaud, S. Combrié, and A. de Rossi, Opt. Express 21, 10324 (2013).
20. S. Hu, M. Khater, R. Salas-Montiel, E. Kratschmer, S. Engelmann, W. M. Green, and S. M. Weiss, Sci. Adv. 4, eaat2355 (2018).
21. T. Tanabe, M. Notomi, H. Taniyama, and E. Kuramochi, Phys. Rev. Lett. 102, 043907 (2009).
22. T. S. Rasmussen, Y. Yu, and J. Mork, Proc. SPIE 10939, 109390A (2019).
23. Y. Yu, E. Palushani, M. Heuck, N. Kuznetsova, P. T. Kristensen, S. Ek, D. Vukovic, C. Peucheret, L. K. Oxenløwe, S. Combrié, A. de Rossi, K. Yvind, and J. Mørk, Opt. Express 21, 31047 (2013).
24. A. E. Siegman, Lasers (University Science Books, 1986).
25. G. Crosnier, D. Sanchez, S. Bouchoule, P. Monnier, G. Beaudoin, I. Sagnes, R. Raj, and F. Raineri, Nat. Photonics 11, 297 (2017).
26. M. Takiguchi, A. Yokoo, K. Nozaki, M. D. Birowosuto, K. Tateno, G. Zhang, E. Kuramochi, A. Shinya, and M. Notomi, APL Photon. 2, 046106 (2017).

ПОЛНОСТЬЮ ОПТИЧЕСКИЕ РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ WDM СО СТРОБИРОВАНИЕМ

Мургиас-Александрис Дж., Дабос Дж., Пассалис Н., Тотович А., Тефас А., Плерос Н.

АННОТАЦИЯ

Нейроморфная фотоника выдвинула на первый план многообещающие нейронные сети (НС) с более высокими вычислительными скоростями по сравнению с электронными аналогами. В этом направлении исследовательские усилия были в основном сконцентрированы на разработке пиковых, сверточных и прямых (FF) -NN архитектур, направленных на решение сложных когнитивных проблем. Однако для решения сложных задач классификации временных рядов и прогнозирования современные модели глубокого обучения в большинстве случаев требуют использования рекуррентных NN (RNN) вместе с их стробированными вариантами, такими как Long-Short -Term-Memories (LSTM) и Gated-Recurrent-Units (GRU). Здесь мы экспериментально демонстрируем первую, насколько нам известно, полностью оптическую RNN со стробирующим механизмом, закладывая основу для полностью оптических LSTM и GRU. В предлагаемых схемах используется сигмоидальная активация на основе полупроводникового оптического усилителя (SOA) в волоконной петле, и они были проверены с использованием асинхронных сигналов с мультиплексированием по длине волны (WDM) с оптическими импульсами 100 пс. В версии Gated-RNN использовался вентиль SOA-Mach-ZehnderInterferometer (SOA-MZI), при этом выход RNN определял долю входного сигнала, которая требуется для входа в RNN. Наконец, сложная архитектура NN была обучена с использованием набора финансовых данных FI-2010 с использованием предложенных не-стробируемых и стробируемых-RNN, продемонстрировав выдающиеся результаты F1 в 41,68% и 41,85% соответственно, превосходя Multi-Layer Perceptron (MLP). базовых моделей в среднем на 6,49%. Ключевые слова: нейроморфная фотоника, программируемая фотоника, нейроморфные вычисления, оптические нейронные сети, рекуррентные нейронные сети.

ВВЕДЕНИЕ

Реальный интеллект (ИИ) начал проникать в нашу повседневную жизнь на всех уровнях, прежде всего через мобильные устройства и встроенные системы, которые могут распознавать объекты, определять, кто говорит, и даже реагировать на раздражители окружающей среды,

что, однако, требует огромного количества вычислительных ресурсов. [1], [2]. Этот неустанный поиск вычислительных ресурсов и обновленного оборудования искусственного интеллекта привел к появлению нейроморфных вычислительных архитектур, которые, в свою очередь, привели к созданию специализированного оборудования, вдохновленного мозгом [2] – [5], с беспрецедентной скоростью вычислений, которое обошло традиционные архитектуры фон Неймана. Поскольку нейроморфная электроника ограничивается электронной Тактовые частоты межсоединений, вычислительная мощность может расти в основном за счет соответствующего увеличения количества электронных нейронов и синапсов, в то время как задержка по своей природе ограничена задержкой межсоединения электронного сопротивления / емкости. Ожидается, что это изменится благодаря внедрению нейроморфной фотоники, которая обещает оснастить нейроморфные вычисления хорошо известными преимуществами полосы пропускания и задержки фотонов, в конечном итоге стремясь к невероятно высокой вычислительной скорости NN и, следовательно, революционизировать приложения со сверхмалой задержкой. требования к обучению [6] – [12]. Это уже стимулировало огромное количество исследований, направленных на разработку ключевых строительных блоков непосредственно в оптической области, включая банки весов WDM [13] – [15], когерентные линейные нейроны [16], блоки суммирования, а также активацию сигмоида и ReLU. функции [17] – [19]. Фотонные нейронные сети до сих пор в основном делали упор на схемах пиков [20], [21], FF [6], [22], сверточных [23] и резервных вычислений (RC) [24] – [28], с ограниченными возможностями. количество усилий [21], [29], [30], направленных на схемы, необходимые для архитектур RNN. Однако с формированием RNN и их стробированных вариантов, которые составляют основу продвинутых макетов LSTM и GRU, которые доминируют в приложениях классификации и прогнозирования временных рядов, переход к высокоскоростным оптическим RNN и механизмам стробирования позволил бы этому большому классу приложений AI получить скорость и Преимущества фотоники с задержкой в критических по времени задачах последовательной обработки данных, таких как распознавание текста, обработка естественного языка, финансы и т. д. Классификация временных рядов была успешно выполнена только с помощью ограниченного числа экспериментальных демонстраций фотонного RC [25], [31], [32], с RC, включающими, однако, довольно специальный автономный класс NN, который вряд ли может конкурировать с полномочиями реконфигурации и масштабирования обычных RNN [8] в направлении формирования стробированных вариантов LSTM и GRU. Таким образом, становится очевидным, что реализация фотонной RNN, способной работать полностью в оптической

области, является «золотым пятном», чтобы сделать скачок к более сложным, но поддерживаемым фотонами Gated RNN, таким как GRU. В этой работе мы предлагаем и впервые экспериментально демонстрируем, насколько нам известно, полностью оптическую WDM RNN и стробируемый рекурсивный механизм, основанный на оптическом соединении G. Mourgias-Alexandris, G. Dabos, N. Passalis, A. Тотович, A. Тефас и H. Плерос работают с факультета информатики Университета Аристотеля в Салониках, 54621 Салоники, Греция (электронная почта: mourgias@csd.auth.gr, ntamposg@csd.auth.gr, passalis @ csd.auth.gr, angelina@auth.gr, tefas@csd.auth.gr, npleros@csd.auth.gr). Полностью оптические рекуррентные нейронные сети WDM со стробированием Джордж Мургиас-Александрис, Джордж Дабос, Николаос Пассалис, Анджелина Тотович, Анастасиос Тефас и Никос Плерос

Активация двух сигмоидов [17] в рекурсивной волоконно-оптической петле модуль. Экспериментально продемонстрировано, что WDM RNN поддерживает рекуррентность и, следовательно, ее функции «памяти» как при синхронных, так и при асинхронных потоках входных данных с оптическими импульсами длиной 100 пс, передаваемыми по каналам с множеством длин волн. Более того, мы впервые вводим оптический стробирующий механизм в полностью оптическую схему RNN для частичного выполнения функции «забывания», которая традиционно реализуется в электронных GRU, представляя экспериментальные результаты с оптическими импульсами длительностью 100 пс, когда выход RNN диктует уровень входного сигнала, который может попасть в нейрон. Проверяя производительность этих полностью оптических RNN и стробирующих схем, наша работа предлагает многообещающий план развертывания сверхбыстрых фотонных альтернатив стробированных вариантов RNN, включая LSTM и GRU. Учетные данные приложений предлагаемых макетов без стробируемой и стробируемой RNN были оценены путем обучения их с помощью современных инструментов моделирования глубокого обучения на наборе финансовых данных FI2010, в результате чего была получена оценка F1 41,68% и 41,85% соответственно., в прогнозном выводе в области финансов, который превосходит модели, основанные на MLP, в среднем на 6,49%.

II. Повторяющиеся нейронные сети и архитектуры шлюза

На рисунке 1 представлены концептуальные схемы архитектур RNN с учетом развертывания оптики. Обычная компоновка архитектуры RNN изображена на рис. 1 (а), а возможные варианты, приближающиеся к ГРУ, изображены на рис. 1 (б) и (с). В случае традиционных схем каж-

дый входной сигнал xi (t) должным образом взвешивается с помощью определенного значения веса wi перед входом в этап суммирования. Затем суммированные сигналы, включая смещение b, направляются в функцию активации f (·) для получения соответствующей нелинейности. Результирующий сигнал состоит из вывода y (t) RNN, в то время как копия вывода задерживается однократным экземпляром t для реализации рекуррентного сигнала y (t-1). Наконец, сигнал y (t-1) направляется обратно на этап суммирования, где он взвешивается с коэффициентом wr. Повторяющийся макет регулируется следующим уравнением: $y\ (t) = f\ (\sum wixi\ (t) + wry\ (t-1)\ M\ i = 0 + b)$ (1) с i = 0,1… M – число входов. Двигаясь вперед к более сложным вариантам RNN, которые демонстрируют лучшую точность по сравнению с обычными, современные реализации в основном включают LSTM и GRU. Преимущество этих сложных вариантов над обычными RNN проистекает из их функций временной памяти и, в частности, из их способности выборочно «забывать» избыточные данные. Другими словами, RNN может решить, какую информацию следует выбросить и какую новую информацию следует принять во внимание, что требует использования рекурсивной схемы стробирования. На рисунке 1 (b) показана RNN, включающая рекурсивный механизм стробирования, который отвечает либо за разрешение, либо за блокировку суммированных входных сигналов. Рекуррентный сигнал y (t-1) после взвешивания с коэффициентом wf используется для управления переходом суммированных входных сигналов через дополнительный вентиль, таким образом частично реализуя функцию «забывания» сети. Комбинируя схемы на рис. 1 (a) и (b), можно получить полностью оптическую RNN с полным механизмом стробирования, как показано на рис. 1 (c). Эта комбинация вводится как средство для объединения операций «сброса» и «обновления» с целью реализации гораздо более простой функции «забыть», которая может использоваться для реализации аллоптических ГРУ по аналогии с программно реализованным минимальным стробируемым блоком (MGU) [33]..

III. Полностью оптическая рекуррентная нейронная сеть.

Рисунок 2 иллюстрирует предлагаемую RNN, основанную на традиционной версии, изображенной на рисунке 1 (a). Фотонная RNN упрощает включение блока оптической активации на основе SOA в контур задержки обратной связи с использованием волокон [17]. Используемая сигмовидная функция активации показывает максимальный контраст 7,5 дБ. На рис. 2 (b) представлена теоретическая модель фотонной активации сигмоида вместе с ее экспериментальной передаточной

функцией, которая почти полностью совпадает с экспериментально измеренной кривой [17]. Подгонка логистической сигмовидной функции к экспериментальной кривой подтвердила сигмовидную реакцию блока активации. Работа предлагаемой RNN регулируется (1), где M = 0. Повторяющиеся возможности предлагаемой RNN были подтверждены с использованием экспериментальной установки, показанной на рис. 2 (a). Генератор сигналов произвольной формы (AWG) использовался для генерации периодического двоичного электрического сигнала формы «1000», содержащего K = 4 символа, где каждый символ имеет период T = 400 пс с длительностью оптического импульса 100 пс. Генерируемый электрический сигнал был преобразован в соответствующий оптический сигнал x0 (t) с помощью оптического модулятора из ниобата лития (LiNbO3), который подавался непрерывным лазерным лучом на λ0 = 1549,2 нм, в то время как выигрыш в весе был обеспечен с помощью переменной -Оптический аттенюатор (VOA). Затем взвешенный входной сигнал w0 · x0 (t) был направлен в блок активации сигмовидной кишки и разделен на 2 идентичных сигнала, которые вводились в SOA-MZI в качестве управляющих сигналов вместе с непрерывными лазерными лучами λ1 = 1550.0 нм и λ2. = 1548,1 нм, которые служили входным и вспомогательным сигналами соответственно, реализуя схему с дифференциальным смещением [17]. Сигнал с преобразованием длины волны, который появился на выходе, показанном на

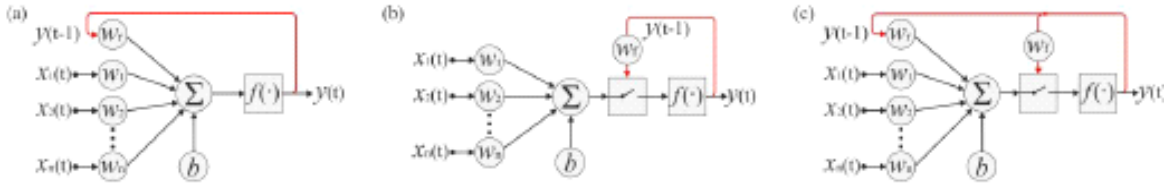


рис. 1. Архитектура RNN с точки зрения фотонных развертываний. (a) Обычный RNN с (b) промежуточным стробирующим механизмом. (c) RNN с полным запорным механизмом.

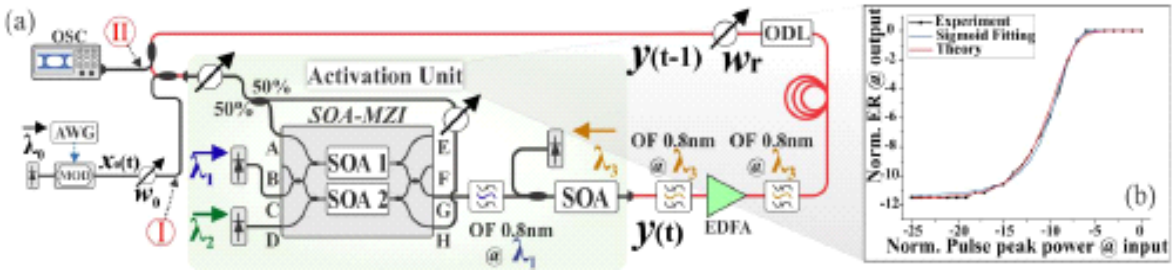


Рис. 2. (a) Экспериментальная установка, используемая для оценки повторяющихся возможностей предлагаемой RNN, (b) Экспериментальная и теоретическая передаточная функция блока активации сигмовидной кишки [16]

SOA-MZI был отфильтрован полосовым оптическим фильтром (OF) с полосой пропускания 3 дБ 0,8 нм на λ1 перед вводом следующего SOA

в качестве управляющего сигнала вместе с непрерывным лазерным лучом на λ3 = 1551,1 нм. Таким образом, SOA работает как преобразователь длины волны CrossGain Modulation (XGM) с выходным сигналом y (t), отпечатанным на λ3. Впоследствии OF на λ3 использовался для фильтрации выходного сигнала блока активации, затем использовался волоконный усилитель, легированный эрбием (EDFA), для компенсации оптических потерь, в то время как другой OF на λ3 использовался для фильтрации усиленного спонтанного излучения (ASE) EDFA.. Затем сигнал подавался в оптоволоконный контур обратной связи общей длиной 61 м, включая компоненты с оптоволоконным соединением, а также все необходимые оптоволоконные соединения между ними. Индуцированная временная задержка была равна D = T × (N × K-1), где T обозначает период символа, N – положительное целое число, а K – количество символов для каждого периода сигнала. Индуцированная задержка D была точно настроена с помощью оптической линии задержки (ODL), чтобы позволить версии y (t1) временно совпадать с новой записью x0 (t). Как уже упоминалось выше, оптоволоконный контур обратной связи имеет длину, равную 61 м, вызывая временную задержку 305 нс, что соответствует N, равному 190. VOA использовался для обеспечения весового значения wr, в то время как оптический разветвитель был использован для создания двух идентичных копий сигнала, первая из которых перенаправлялась в осциллограф (OSC) для отслеживания выходного сигнала нейрона.

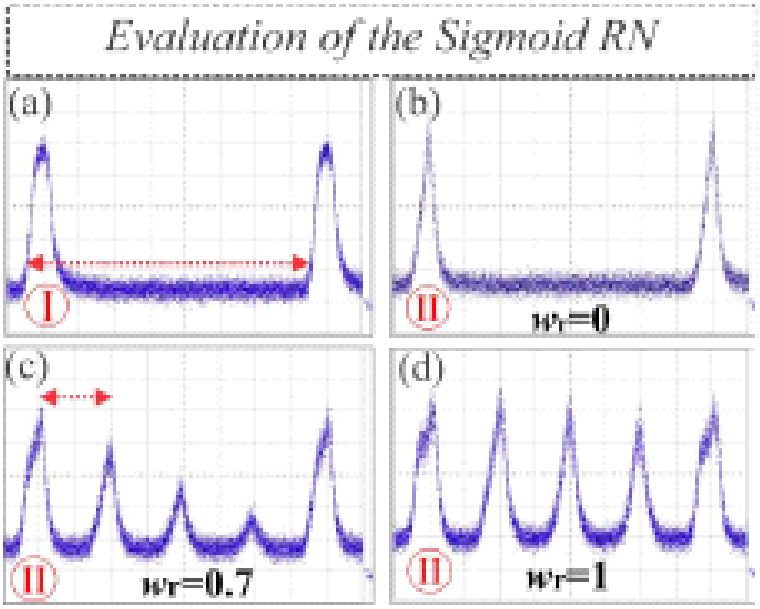


Рис. 3. Экспериментальная часть. временные трассы (200 пс / дел), полученные для разметки single-λ. (a) входной сигнал и (b) – (d) выход RNN для коэффициентов ослабления: (b) wr = 0, (c) wr = 0,7 и (d) wr = 1. Масштаб по оси Y: (3,5 мВ / дел).

Вторая копия была объединена с взвешенным входным сигналом w0 · x0 (t) перед повторным вводом в блок активации сигмовидной кишки. Чтобы экспериментально проверить работу RNN, исследуя различные уровни мощности рекуррентного сигнала, использовались несколько весовых значений (wr), при этом входной вес w0 постоянно был установлен на «1», следовательно, соответствующий нулевому затуханию. Рисунок 3 (b) – (d) иллюстрирует полученные временные кривые после запуска в RNN входного сигнала x0 (t), показанного на рисунке 3 (a). Установка максимально возможного значения затухания, равного 30 дБ, соответствует wr = 0. Результирующий выходной сигнал РНС показан на рис. 3 (б). С учетом значения затухания 1,55 дБ (wr = 0,7) выходной сигнал RNN формирует три последовательных импульса с периодом T, как показано на рисунке 3 (c). Пиковая мощность этих импульсов периодически уменьшалась в пределах окна периода T, таким образом подтверждая особенности памяти предполагаемого нейрона. Наконец, без применения какого-либо ослабления к рекуррентному сигналу (wr = 1), эти три импульса были почти уравновешены по мощности относительно начального входного импульса, как ясно видно на рис. 3 (d). Условия эксплуатации используемых устройств во время проведенных экспериментов кратко описаны ниже: Для SOA-MZI, SOA1 и SOA2 приводились в действие постоянным током 260 мА и 290 мА, соответственно. Входной (порт C) и вспомогательный лазерный луч (порт D) составляли 6,2 дБм и 7 дБм соответственно, в то время как оптическая мощность управляющих сигналов на портах A и H составляла 10,6 дБм и 9,8 дБм соответственно. Ток возбуждения выходного SOA составлял 300 мА, в то время как инжектированный лазерный луч имел оптическую мощность -9,8 дБм.

IV. ВСЕ-ОПТИЧЕСКАЯ РЕКУРРЕНТНАЯ НЕЙРОННАЯ СЕТЬ WDM

Чтобы всесторонне оценить возможности предлагаемой RNN по поддержанию рекуррентности и, таким образом, функции памяти, мы провели дополнительный эксперимент, в котором четыре взвешенных входа WDM были запущены на входе нейрона.

После своевременной синхронизации, а не наоборот. На рисунке 4 показана экспериментальная установка, использованная для обоих сценариев. Четыре лазерных луча на λ0 = 1546,3 нм, λ1 = 1547,1 нм, λ2 = 1547,9 нм и λ3 = 1548,7 нм модулировались четырьмя отдельными модуляторами LiNbO3, управляемыми AWG, но несли тот же сигнал xi (t). Временная синхронизация или десинхронизация четырех входных сигналов была выполнена с использованием ODL перед мультиплексированием.

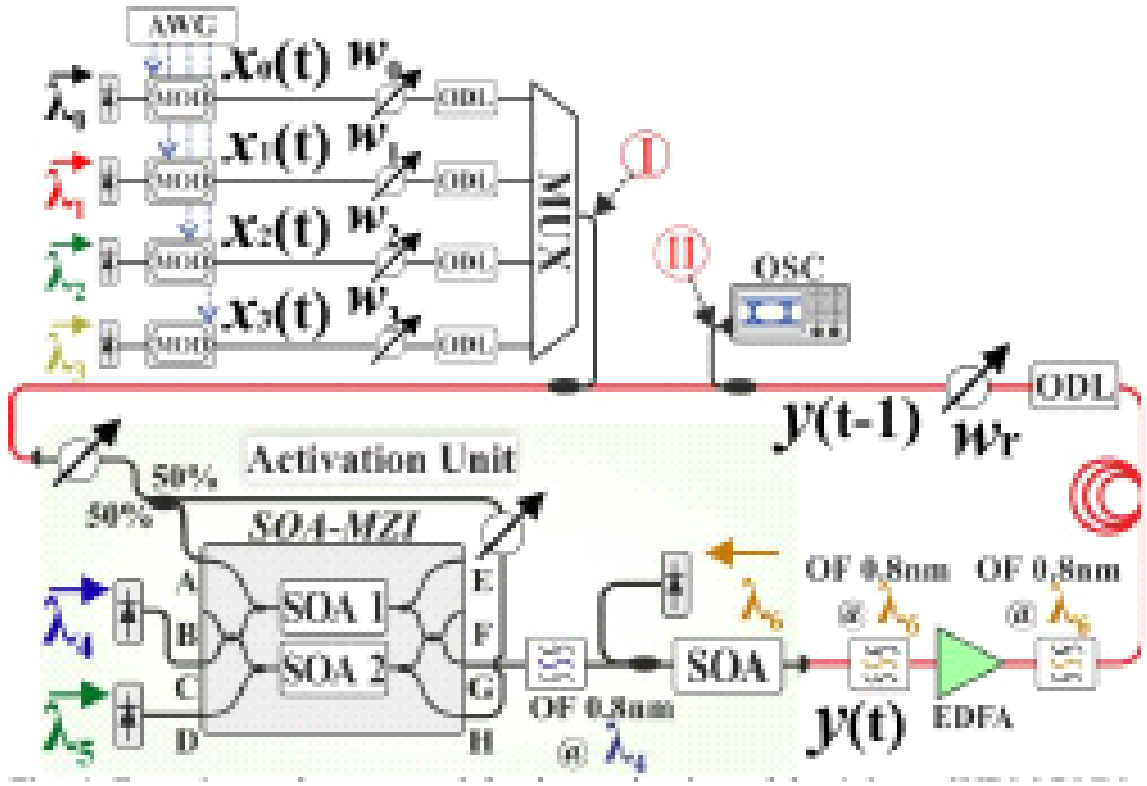


Рис. 4. Экспериментальная установка, использованная для оценки 4-входной WDM RNN. Рис. 2. (a) Экспериментальная установка, используемая для оценки рекуррентных возможностей предлагаемой RNN, (б) Экспериментальная и теоретическая передаточная функция сигмовидного блока активации [16]. Разрешенное лицензионное использование ограничивается:

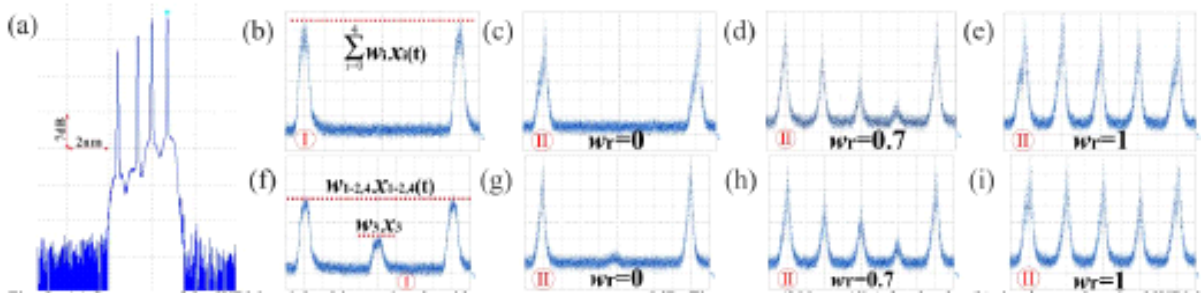


Рис. 5. (a) Спектр взвешенных входных сигналов WDM с контрастом мощности до 8 дБ. Временные диаграммы (200 пс / дел) для (b) своевременно синхронизированного входа WDM, а также для выхода RNN с коэффициентом затухания: (c) wr = 0, (d) wr = 0,7 и (e) wr = 1. Соответствующие результаты после наведения задержки на два временных интервала для сигнала w3x3 (t) показаны в (i) – (f). Масштаб по оси Y: (3,5 мВ / дел).

На рисунке 5 (a) показан спектр сигнала WDM на выходе мультиплексора (MUX). Когда все сигналы своевременно синхронизированы, сигнал WDM изображен на рис. 5 (b), а пиковая мощность суммированных сигналов обозначена красной пунктирной линией. На рисунках 5 (c) – (e) показан выходной сигнал RNN для значений wr 0, 0,7 и 1, которые соответствуют уровням затухания 30, 1,55 и 0 дБ для рекуррентного

сигнала, соответственно. RNN поддерживала повторяемость для всех исследованных значений весов, ясно показывая стабильную работу модуля также и для сигналов WDM. На рисунке 5 (f) показан асинхронный входной сигнал WDM, когда сигнал w3x3 (t) задерживается на 2T (800 пс) по отношению к w1x1 (t), w2x2 (t) и w4x4 (t), которые остались синхронизированными друг с другом. Уровень мощности w3x3 (t) и w1-2,4x1-2,4 (t) отмечен соответствующими красными пунктирными линиями на временной шкале на рис. 5 (f). Как ясно видно на фиг. 5 (g), путем принудительного применения wr = 0 и установки порога блока активации сигмовидной формы выше уровня мощности w3x3 (t), w3x3 (t) подавляется до нуля. Выходные сигналы RNN для повторяющихся весовых значений 0,7 и 1 изображены на фиг. 5 (h) и (i) соответственно. В этом случае также предлагаемая RNN сохраняет функции памяти, поддерживая повторяемость. Экспериментальная оценка RNN с 4 входами WDM проводилась с использованием лазерных лучей, настроенных на λ4 = 1550,1 нм, λ5 = 1551,1 нм и λ6 = 1152,5 нм, в то время как оптическая мощность управляющих сигналов на портах A и H была равна до 10 дБм и 9,2 дБм соответственно. Условия эксперимента для остальной части экспериментальной установки были точно такими же, как при оценке RRN в разделе III.

V. ПОВТОРЯЮЩАЯСЯ НЕЙРОННАЯ СЕТЬ GATED SIGMOID

Проникая глубже в архитектуры закрытых RNN, стремясь напоминать полностью функциональный GRU, мы включили в используемую RNN механизм стробирования, позволяющий гибко реконфигурировать функции «забыть». Поэтому мы разместили коммутатор на основе SOA-MZI перед используемой RNN. На рис. 6 (a) показана развернутая архитектура на основе рис. 1 (b). Входной сигнал x0 (t), отпечатанный на λ0, взвешивается с коэффициентом выигрыша перед входом во входной порт SOA-MZI, который работает как оптический переключатель с синусоидальным откликом [34]. Кроме того, непрерывный лазерный луч на λ1 вводится во входной порт синусоидального переключателя, выполняющего роль вспомогательного сигнала. Выигрыш x0 (t) возникает на коммутируемом порте синусоидального переключателя до входа в сигмовидную RNN без внесения каких-либо изменений в ранее описанные рабочие условия. Выходной сигнал RNN y (t-1) взвешивается с коэффициентом wf перед достижением порта управления синусоидального переключателя. Тогда выходной сигнал переключателя будет пропорционально уменьшаться по отношению к амплитуде рекуррентного сигнала y (t-1). Следовательно, когда оптической мощности управляющего

сигнала достаточно, чтобы вызвать сдвиг π-фазы в SOA1 синусоидального переключателя, суммированный сигнал будет полностью «забыт». Экспериментальная установка показана на рис. 6 (a). Лазерный луч с λ0 = 1546,3 нм вводился в оптический модулятор LiNbO3, управляемый AWG, который создавал периодический сигнал x0 (t) с битовой комбинацией «1110», K = 4 и T = 400 пс. Выигрыш веса реализуется через VOA до того, как взвешенный сигнал достигает порта B синусоидального переключателя. Вспомогательный сигнал с λ1 = 1547,1 нм вводился в порт C. Выходной сигнал синусоидального переключателя, переносимый λ0, появлялся на порту G. После фильтрации с использованием OF с полосой пропускания 3 дБ 0,8 нм был получен контраст мощности до 8дБ. Временные диаграммы (200 пс / дел) для (b) своевременно синхронизированного входа WDM, а также для выхода RNN с коэффициентом затухания: (c) wr = 0, (d) wr = 0,7 и (e) wr = 1. Соответствующие результаты после наведения задержки на два временных интервала для сигнала w3x3 (t) показаны в (i) – (f). Масштаб по оси Y: (3,5 мВ / дел).

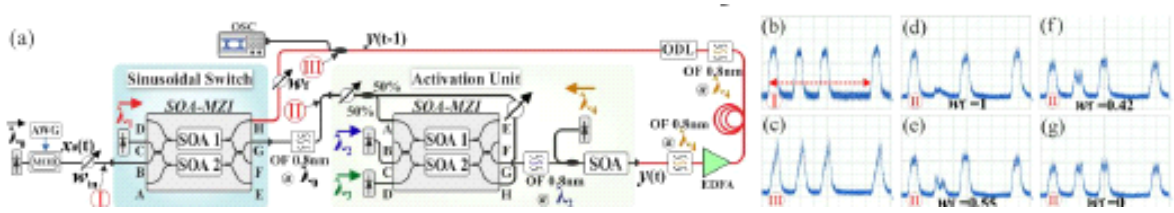


Рис. 6. (a) Экспериментальная установка, используемая для проверки механизма стробирования, включенного в полностью оптическую RNN с помощью синусоидального переключателя на основе SOA-MZI. Временные кривые (200 пс / дел), полученные (б) на входе и (в) на выходе RNN (III) для wf = 0. Выход переключателя (II) для (d) wf = 1, (e) wf = 0,55, (f) wf = 0,42 и (g) wf = 0. Масштаб по оси Y: (3,5 мВ / дел).

переадресовано в RNN. Оптоволоконный контур обратной связи, включающий оптоволоконные выводы синусоидального переключателя, имеет общую длину 76 м, в результате чего N равно 237. Рекуррентный сигнал y (t-1), отпечатанный на λ4, разделяется на 2 идентичные копии. Первый вводится в OSC для целей оценки, а второй взвешивается с коэффициентом wf до достижения порта H синусоидального переключателя. Рисунки 6 (b) – (g) иллюстрируют полученные временные кривые, оценивающие механизм стробирования используемого устройства с использованием различных значений wf. Временной ход входного сигнала x0 (t) показан на рис. 6 (b), а выходной сигнал переключателя для wf = 0 и win = 1 показан на рис. 6 (c). Регулируя значение взвешивания до wf = 1, что соответствует нулевому затуханию, выходной сигнал переключателя преобразуется в «1010», как можно увидеть на рис. 6 (d). На рисунках 6 (e) – (g) показано постепенное преобразование выхода «1010» RNN в исходный входной шаблон «1110» путем постепенного уменьше-

ния значений wf до 0,55, 0,42 и 0 соответственно. Весовые значения wf были реализованы путем применения оптического ослабления 2,9, 3,2 и 30 дБ соответственно. Обратите внимание, что на рис. 6 (d) – (g), где постепенное проявление второго импульса дает сигнал с изменяемым рабочим циклом, можно наблюдать небольшое колебание амплитуды импульса. В каждом случае, когда сигнал вводится в EDFA оптоволоконного контура обратной связи, в результате получается усиленный сигнал с различным коэффициентом усиления, который, в свою очередь, вызывает эти колебания амплитуды после прохождения вентилей SOA. Условия эксплуатации используемых устройств кратко описаны ниже: SOA1 и SOA2 синусоидального переключателя приводились в действие постоянным током 240 мА и 270 мА, соответственно. Оптическая мощность входного (порт B) и вспомогательных сигналов (порт C) составляла -1 и 1 дБм соответственно, в то время как оптическая мощность управляющего сигнала на порту H составляла 2,1 дБм.

VI. ПОЛНОСТЬЮ ОПТИЧЕСКИЕ RNNS ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАДАЧ

Помимо возможности их аппаратной реализации, фотонные RNN могут считаться полезными только при проверке их способности обучаться более сложным схемам NN, чтобы подтвердить их экспериментально полученную производительность по процедурам обучения, используемым для реального приложения. задачи. С этой целью сложная архитектура NN, использующая предложенную фотонную RNN, показанную на рис. 1 (a), использовалась для запуска эталонного теста для задачи прогнозирования в области финансов, в то время как производительность была дополнительно увеличена за счет использования фотонной gatedRNN на рис. 1 (c). Стандартные инструменты и методики глубокого обучения использовались для обучения сети с использованием набора данных книги заявок на ограничение частоты (сокращенно «FI-2010») [35]. Набор данных FI-2010 содержит данные (более 4,5 миллионов лимитных заказов), собранные от 5 финских компаний. Задача прогнозирования касается предсказания направления будущего движения средней цены в привязанной установке оценки (вверх, вниз или стационарно) после 10 временных шагов, как описано в [36]. 10 последних векторов признаков были загружены в модель для каждого шага прогнозирования, в то время как первые 4 дня набора данных использовались для проведенной оценки. Используемая архитектура нейронной сети основана на 32 повторяющихся нейронах, образующих слой RNN. Обратите внимание, что выходной сигнал каждого рекуррентного нейрона подавался не только на сам нейрон, но и на оставшиеся рекуррентные

нейроны в соответствии с современными архитектурами рекуррентного глубокого обучения [37]. Затем за рекуррентным слоем следуют 512 нейронов прямой связи и 3 выходных нейрона (по одному для каждой категории прогнозирования).

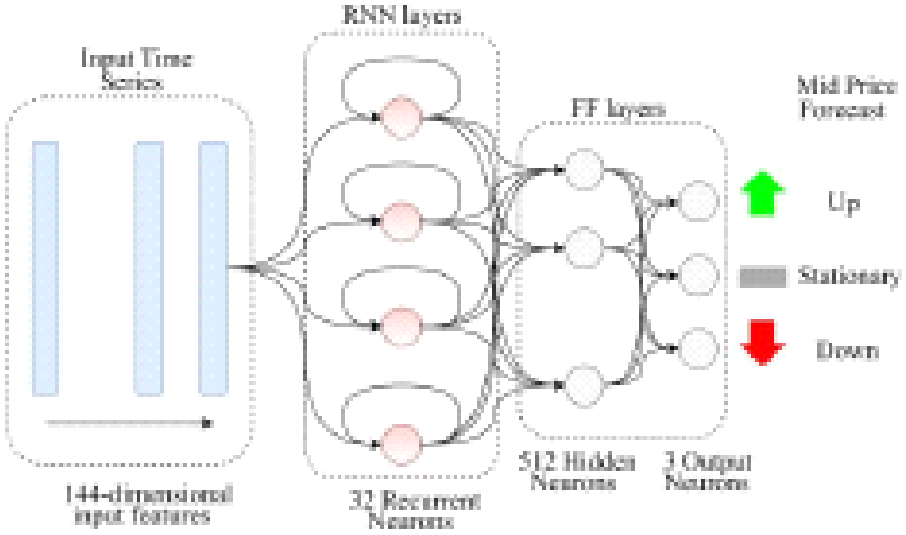


рис. 7. Архитектура нейронной сети, используемая для прогнозирования временных рядов книги лимитных ордеров.

TABLE I
F1-2010 DATASET EVALUATION

Model	Avg. F1	Cohen's κ
MLP	35.27±1.05	0.1058±0.0108
RNN	40.44±1.77	0.1648±0.0184
Photonic RNN	**41.68±2.73**	**0.1693±0.0300**
Photonic Gated-RNN	**41.85±1.64**	**0.1699±0.0238**

ТАБЛИЦА I F1-2010 ОЦЕНКА НАБОРА ДАННЫХ Модель Сред. K MLP F1 Коэна 35,27 ± 1,05 0,1058 ± 0,0108 RNN 40,44 ± 1,77 0,1648 ± 0,0184 Photonic RNN 41,68 ± 2,73 0,1693 ± 0,0300 Photonic Gated-RNN 41,85 ± 1,64 0,1699 ± 0,0238 <6

Полная архитектура изображена на рис. 7. Оптимизация выполнялась в течение 20 периодов с использованием алгоритма RMSprop со скоростью обучения η = 10-4 [38]. Фотонные модели были соответствующим образом инициализированы с учетом экспериментального поведения передаточной функции синусоидального переключателя и модуля сигмовидной функции активации. В частности, для обоих компонентов был использован процесс подбора экспериментально полученных передаточных функций [17], [34]. Стоит отметить, что наш подход фундаментально отличается от методов резервуарных вычислений, поскольку мы оптимизируем всю архитектуру сквозным образом, а не просто устанавливаем классификатор поверх фиксированного случайного преобразования. Экспериментальные результаты для набора данных FI-2010 представлены в таблице I. Производительность предложенного фотон-

ного рекуррентного нейрона также сравнивалась с базовым уровнем MLP, который работает с последним вектором признаков (отбрасывая потенциально полезную временную информацию), а также с Модель PHC, использующая обычную сигмовидную функцию активации вместо фотонной. Предложенная фотонная RNN достигает показателя F1 41,68% и показателя каппа 0,1693%, что значительно превосходит модель MLP и подтверждает ее способность захватывать временную информацию, содержащуюся во входных данных. Предложенная фотонная стробируемая PHC достигает показателя F1 41,85% и показателя каппа 0,1699%. И предлагаемая фотонная неуправляемая и стробируемая PHC, кажется, работает немного лучше, чем PHC с той же архитектурой и регулярными сигмовидными активациями, демонстрируя, что предлагаемая архитектура, основанная на не стробированных или стробированных рекуррентных нейронах, может эффективно обучаться использование инструментов глубокого обучения, а также то, что его производительность конкурентоспособна по сравнению с существующими повторяющимися моделями. В потенциальной аппаратной реализации предложенных схем фотонной несвязанной и стробируемой RNN соответствующие модули на основе SOA будут иметь энергоэффективность 180 пДж / символ и 300 пДж / символ, соответственно. Однако более сложная технология, такая как фотонный анализ III / V-на-Si. кристаллы [39]. предлагает нелинейные функции, аналогичные SOA, но с энергоэффективностью для соответствующих макетов 43,2fJ / Symbol и 72fJ / Symbol соответственно. Возможные недостатки оборудования могут быть уменьшены с помощью передовых методов обучения устойчивости к помехам и вывода, описанных в [40] – [42].

VII. ЗАКЛЮЧЕНИЕ

Мы экспериментально продемонстрировали первую фотонную RNN и полностью оптическую рекурсивную схему стробирования, позволяющую осуществлять логический вывод с помощью оптических импульсов длительностью 100 пс и задержек по времени пролета. Оба модуля основаны на полностью оптическом блоке активации сигмоида, находящемся в рекурсивной структуре петли на основе волокна, при этом в механизме стробирования используется дополнительный оптический вентиль после стадии суммирования. Предлагаемая RNN была проверена экспериментально с использованием четырех входных потоков данных WDM с оптическими импульсами длительностью 100 пс, в то время как стробируемый рекурсивный модуль подтвердил свои полномочия для выборочного «забывания» входных данных в зависимости от выхода RNN и значений рекурсивного веса. Наконец, потенциал предлагаемой полностью оптической невязанной и стробированной-

RNN для проникновения в приложения AI через входные данные временных рядов был подтвержден путем включения 32 RNN в архитектуру NN и обучения ее для успешного решения задач прогнозирования на наборе финансовых данных FI-2010.. Учитывая, что в приложениях для анализа и прогнозирования временных рядов преобладают варианты LSTM и GRU со стробированными RNN, где как RNN, так и механизмы стробируемых схем составляют жизненно важные строительные блоки, наша работа обеспечивает многообещающий полностью оптический инструментарий для передачи преимуществ скорости и задержки фотоны в приложениях для последовательного изучения данных.

REFERENCES

[1] J. Park et al., “Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications,” 2018.

[2] N. P. Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit,” ACM SIGARCH Comput. Archit. News, vol. 45, no. 2, pp. 1–12, 2017.

[3] M. Davies et al., “Loihi: A Neuromorphic Manycore Processor with OnChip Learning,” IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.

[4] S. B. Furber et al., “Overview of the SpiNNaker system architecture,” IEEE Trans. Comput., vol. 62, no. 12, pp. 2454–2467, 2013.

[5] F. Akopyan et al., “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip,” IEEE Trans. Comput. Des. Integr. Circuits Syst., vol. 34, no. 10, pp. 1537–1557, 2015.

[6] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” Nat. Photonics, vol. 11, no. 7, pp. 441–446, Jun. 2017.

[7] T. Ferreira de Lima et al., “Machine Learning with Neuromorphic Photonics,” J. Light. Technol., vol. PP, no. X, pp. 1–1, 2019.

[8] H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, “Neuromorphic Photonic Integrated Circuits,” IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 6, pp. 1–15, 2018.

[9] M. A. Nahmias, T. F. De Lima, A. N. Tait, H. T. Peng, B. J. Shastri, and P. R. Prucnal, “Photonic Multiply-Accumulate Operations for Neural Networks,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, p. 1, 2020.

[10] B. Huval et al., “An Empirical Evaluation of Deep Learning on Highway Driving,” pp. 1–7, 2015.

[11] D. Avenport, A. L. T. Orres, P. A. P. Intus, and J. O. H. N. B. Owers, “Heterogeneous silicon photonics sensing for autonomous cars,” vol. 27, no. 3, pp. 3642–3663, 2019.

[12] A. Totovic, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, “Femtojoule per MAC Neuromorphic Photonics : An Energy and Technology Roadmap,” J. Sel. Toplics Quantum Electron., early access.

[13] A. N. Tait et al., “Neuromorphic photonic networks using silicon photonic weight banks,” Sci. Rep., vol. 7, no. 1, pp. 1–10, 2017.

[14] A. N. Tait, A. X. Wu, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Two-pole microring weight banks," Opt. Lett., vol. 43, no. 10, p. 2276, May 2018.

[15] B. Shi, N. Calabretta, and R. Stabile, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, pp. 1–11, Jan. 2020.

[16] G. Mourgias-Alexandris et al., "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," J. Light. Technol., vol. 38, no. 4, pp. 811–819, Feb. 2020.

[17] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," Opt. Express, vol. 27, no. 7, p. 9620, 2019.

[18] M. Miscuglio et al., "All-optical nonlinear activation function for photonic neural networks [Invited]," Opt. Mater. Express, vol. 8, no. 12, p. 3851, Dec. 2018.

[19] A. N. Tait et al., "Silicon Photonic Modulator Neuron," Phys. Rev. Appl., vol. 10, no. 1, p. 1, 2019.

[20] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature, vol. 569, no. 7755, pp. 208–214, 2019.

[21] M. A. Nahmias et al., "A TeraMAC Neuromorphic Photonic Processor," in 31st Annual Conference of the IEEE Photonics Society, IPC 2018, 2018, pp. 1–2.

[22] J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," Advanced Photonics, vol. 1, no. 04. p. 1, 2019.

[23] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-Chip Optical Convolutional Neural Networks," pp. 1–18, 2018.

[24] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," Sci. Rep., vol. 6, no. March, pp. 1–12, 2016.

[25] K. Vandoorne et al., "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat. Commun., vol. 5, pp. 1–6, 2014.

[26] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," IEEE Trans. Neural Networks, vol. 22, no. 9, pp. 1469–1481, 2011.

[27] F. Laporte, A. Katumba, J. Dambre, and P. Bienstman, "Numerical demonstration of neuromorphic computing with photonic crystal cavities," Opt. Express, vol. 26, no. 7, p. 7955, 2018.

[28] A. Katumba, J. Heyvaert, B. Schneider, S. Uvin, J. Dambre, and P. Bienstman, "Low-Loss Photonic Reservoir Computing with Multimode Photonic Integrated Circuits," Sci. Rep., vol. 8, no. 1, pp. 1–10, 2018.

[29] H. Peng, T. F. de Lima, M. A. Nahmias, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Autaptic Circuits of Integrated Laser Neurons," in Conference on Lasers and Electro-Optics, 2019, vol. 1, no. c, p. SM3N.3, doi: 10.1364/CLEO_SI.2019.SM3N.3.

[30] B. Romeira, R. Avo, J. M. L. Figueiredo, S. Barland, and J. Javaloyes, "Regenerative memory in time-delayed neuromorphic photonic resonators," Sci. Rep., vol. 6, no. January, pp. 1–13, 2016.

[31] L. Larger et al., "Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing," Opt. Express, vol. 20, no. 3, p. 3241, Jan. 2012.

[32] L. Larger, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, and M. Jacquot, "High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification," Phys. Rev. X, vol. 7, no. 1, pp. 1–14, 2017.

[33] G. Zhou, J. Wu, C.-L. Zhang, and Z.-H. Zhou, "Minimal Gated Unit for Recurrent Neural Networks," International Journal of Automation and Computing 13, no. 3, 226-234, Mar. 2016.

[34] G. Mourgias-alexandris, A. Tsakyridis, T. Alexoudi, K. Vyrsokinos, and N. Pleros, "Optical Thresholding Device with a Sigmoidal Transfer Function," Photonics Switch. Comput., pp. 6–8, 2018.

[35] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods," J. Forecast., vol. 37, no. 8, pp. 852– 866, 2018.

[36] P. Nousi et al., "Machine learning for forecasting mid-price movements using limit order book data," IEEE Access, vol. 7, pp. 64722–64736, 2019.

[37] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 25,2020 at 11:06:08 UTC from IEEE Xplore. Restrictions apply. 1077-260X (c) 2020 IEEE.

Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_ standards/publications/ rights/index.html for more information. This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2020.2995830, IEEE Journal of Selected Topics in Quantum Electronics > REPLACE THIS LINE WITH YOUR PAPER IDENTIFICATION NUMBER (DOUBLE-CLICK HERE TO EDIT) < 7 Networks Learn. Syst., vol. 28, no. 10, pp. 2222–2232, 2017.

[38] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural networks for machine learning, vol. 4, no. 2, pp. 26–31, 2012.

[39] T. Alexoudi et al., "III-V-on-Si Photonic Crystal Nanocavity Laser Technology for Optical Static Random Access Memories," IEEE J. Sel. Top. Quantum Electron., vol. 22, no. 6, pp. 295–304, 2016.

[40] N. Passalis, G. Mourgias-alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, "Training Deep Photonic Convolutional Neural Networks with Sinusoidal Activations," in IEEE Transactions on Emerging Topics in Computational Intelligence.

[41] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," Opt. Express, vol. 27, no. 10, p. 14009, 2019.

САМОНАСТРАИВАЮЩАЯСЯ МИКРОВОЛНОВАЯ МАСКИРОВКА С ПОДДЕРЖКОЙ ГЛУБОКОГО ОБУЧЕНИЯ БЕЗ ВМЕШАТЕЛЬСТВА ЧЕЛОВЕКА

Цянь Ч., Чжан Б., Шэнь И., Цзин Л., Ли Э., Шэнь Л., Чэнь Х.
https://doi.org/10.1038/s41566-020-0604-2.

Становление невидимым по своему желанию очаровывало человечество на протяжении веков, а в последнее десятилетие оно привлекло большое внимание в связи с появлением метаматериалов. Однако современные плащи-невидимки обычно работают в детерминированной системе или в сочетании с внешней помощью для достижения активной маскировки. Здесь мы предлагаем концепцию интеллектуальной (то есть самоадаптирующейся) маскировки, основанной на глубоком обучении, и представляем маскировку метаповерхности в качестве примера реализации. В эксперименте маскировка метаповерхности показывает время отклика в миллисекундах на постоянно меняющуюся волну и окружающую среду без какого-либо вмешательства человека. Наша работа приближает доступные стратегии маскировки к широкому спектру приложений реального времени на месте, таких как движущиеся малозаметные машины. Этот подход открывает путь к упрощению использования других интеллектуальных метаустройств в микроволновом режиме и в более широком электромагнитном спектре и, в более общем плане, позволяет автоматически решать электромагнитные обратные задачи проектирования.

Rendering an object invisible with an invisibility cloak is a long-standing dream for humanity, with potential in many interesting applications. For example, it would be very excit-

Чтобы ездить по разным средам, стеллажам и холмам – для езды на разных холмах, стилей и холмов. не будучи обнаруженным. Такой сценарий был немыслим, пока не было предложено преобразование оптики1 и метаматериалов 2.

Плащ на основе трансформирующей оптики создает невидимость за счет точной конструкции его основных параметров, чтобы направлять поток света вокруг скрытого объекта, но экспериментальное достижение составов объемных материалов с анизотропией и неоднородностью представляет собой серьезную проблему. Другие методы невидимости, такие как подавление рассеяния 7,8 и градиентное покрытие метаповерхности 9–11, также были предложены в форме электромагнитных (ЭМ) 1–11, звуковых 12 и упругих волн13 или тепловых потоков14, в зависимости от сценария применения. В ответ на внешний раздражитель или нестационарную среду идеальный плащ-невидимка должен быстро и автоматиче-

ски настраивать свою внутреннюю структуру – свои активные компоненты – так, чтобы он всегда оставался невидимым, как если бы плащ наделен способностью хамелеона. Это очень желательное, но нелегко достижимое свойство – интеллектуальная или самоприспособляемость – должно быть развито так, чтобы его можно было применять в широком диапазоне приложений реального времени, связанных с движущимися объектами и средами на месте. Для этого очень важно уметь интерпретировать сложные и не интуитивные отношения между маскирующей структурой, входящим светом и окружающей средой, а затем передавать эти знания / опыт плащу. Однако стратегии, доступные в настоящее время для обработки взаимодействий света и материи, обычно включают в себя полноволновое моделирование в каждом конкретном случае и восходящий дизайн, что ограничивает область применения несколькими способами. Во-первых, поскольку полноволновое моделирование включает итерационные и длительные вычисления уравнений Максвелла, они по своей сути являются длительными процедурами15,16; это неизбежно снижает эффективность маскировки и ограничивает ее способность реагировать в короткие сроки. Во-вторых, в погоне за желаемой наноструктурой будет много неудачных попыток; они обычно выбрасываются, что приводит к неэффективной трате ресурсов. Наконец, используемая стратегия должна быть способной быстро реагировать на постепенные изменения сценария (например, в оптической среде); такие изменения в отклике плаща нетривиальны из-за сильной нелинейной и неуловимой природы взаимодействий света с веществом18. Программируемые и реконфигурируемые метаматериалы и так называемые адаптивные ЭМ-устройства были тщательно изучены и лежат в основе производительности некоторых установленных технологий19–21, но они также должны работать в тандеме с внешним вводом (зависимым, неавтономным) и в процессе испытаний и - Режим ошибки для соответствия требованиям заказчика, сильно затрудняющий приложения в реальном времени (дополнительное примечание 11). Учитывая все эти факторы, очень востребован эффективный подход, позволяющий избежать отнимающего много времени повторяющегося ручного труда и развить способность к саморегулированию в мантии-невидимке, но его сложно реализовать. В этой статье мы предлагаем концепцию интеллектуального плаща-невидимки, основанного на глубоком обучении, и строим пример такого плаща с настраиваемой метаповерхностью (рис. 1). Свойство отражения каждого элемента внутри метаповерхности можно независимо настраивать, подавая различные напряжения смещения постоянного тока (d.c.) через нагруженный варакторный диод 19 на микроволновых частотах. Встроенная предварительно обученная искусственная нейронная сеть (ИНС), фундаментальный строительный блок глубокого обучения, метаповерхность может быстро реагировать в милли-

секундном масштабе времени на постоянно меняющуюся падающую волну и окружающий фон без какого-либо вмешательства человека.

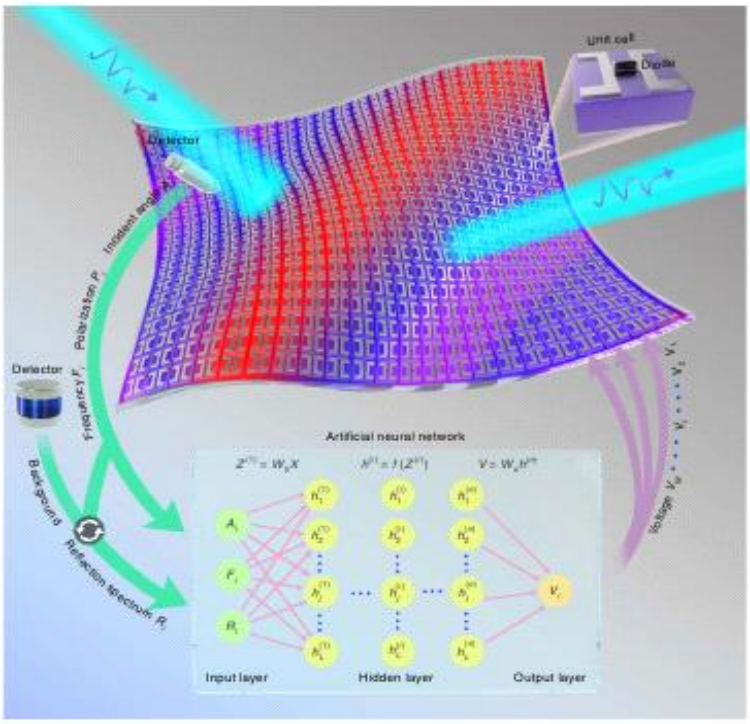


Рис. 1 | Схема самоадаптирующейся метаповерхности с поддержкой глубокого обучения. Маскировка метаповерхности состоит из ультратонкого слоя активных метаатомов, каждый из которых содержит варакторный диод, который независимо управляется постоянным током. напряжение смещения (верхний правый угол). В ответ на обнаруженную падающую волну и окружающий фон все напряжения смещения автоматически рассчитываются предварительно обученной ИНС и мгновенно выдаются платформой источника питания. Обратите внимание, что обнаруженная фоновая информация преобразуется в требуемый спектр отражения для каждого метаатома на основе метода восстановления волны 9, а затем подается в предварительно обученную ИНС в качестве входных данных. См. Дополнительное примечание 7 для описания принципа работы.

Все напряжения смещения рассчитываются автоматически и мгновенно поступают на маскировку. Мы используем программу конечных разностей во временной области (FDTD) 22,23 с предварительно обученной ИНС для имитации реальной сцены, а затем оцениваем производительность с помощью экспериментального эксперимента. Это подтверждает, что маскировка метаповерхности демонстрирует эффективную и устойчивую самоадаптируемость в ответ на случайную падающую волну и фон. Наша работа приближает доступные стратегии маскировки к практическим сценариям, включающим изменение фона, движущихся объектов, мультистатическое обнаружение и многое другое, и это может быть легко масштабируется до более высоких частот (дополнительное примечание 12). Более того, предложенная концепция приведет к новому роду

интеллектуальных метаустройств и неконтролируемому манипулированию электромагнитными волнами24. Архитектура интеллектуальной маскировки-невидимки Архитектура интеллектуальной метаповерхностной маскировки представлена на рис. 1. Она состоит из пяти основных частей: настраиваемого реконфигурируемого метаповерхностного включения, двух детекторов (для падающей волны и окружающего фона), предварительно обученной ИНС и платформа электроснабжения. Ультратонкое метаповерхностное включение состоит из множества активных элементарных ячеек / метаатомов, каждая из которых обеспечивает свой локальный спектр отражения и, таким образом, создает рассеянную волну, подобную той, которая генерируется оголенным окружением без скрытого объекта. Спектр локального отражения, содержащий как амплитуду, так и фазу, выводится из окружающего фона, местоположения маскировки и падающей волны с помощью метода реконструкции волны9 (дополнительное примечание 7). В каждый метаатом (верхний правый угол, рис. 1) встроен варакторный диод, и его спектр отражения может динамически изменяться путем подачи различных источников постоянного тока. напряжения смещения (выход архитектуры) на нем на микроволновых частотах19,20. Наряду с этим активным подходом, электрическое стробирование, механическое срабатывание, материалы с фазовым переходом и магнитооптическое управление также могут использоваться на их собственных частотах, в диапазоне от микроволновых до оптических и даже видимых частот25. Маскировка метаповерхности (и другие маскировки из метаматериалов) в значительной степени зависит от точной локальной манипуляции фазой – электрического или магнитного отклика – резонансных элементов для восстановления рассеянного света10. Таким образом, для конкретного типа структуры или состояния такая маскировка обычно работает в узкой полосе пропускания, с ограниченным углом падения и со стационарным фоном11. При изменении падающей волны или фона такая метаповерхностная маскировка может стать неэффективной и даже выйти из строя. Здесь факторы, влияющие на маскировку, обнаруживаются в реальном времени и преобразуются в требуемый спектр локального отражения в качестве одного входа в ИНС (дополнительный рисунок 12). Между тем, обнаруженная падающая волна устанавливается в качестве другого входа из-за сильной частотной дисперсии и нелокального эффекта метаповерхности19,26; эта падающая волна характеризуется углом падения, частотой и поляризацией. Используя глубокое обучение, мы стремимся использовать и обобщать сложную взаимосвязь {падающая волна и спектр отражения} → {напряжение смещения} для отдельного метаатома (дополнительный рис. 12). Предварительно обученная ИНС используется параллельно со многими такими метаатомами внутри метаповерхностного плаща, закрывающего большой объект (дополнительный рис. 13); таким образом мы применя-

ем модель ИНС к сложному приложению. В последнее время глубокое обучение использовалось для значительного улучшения современных основных приложений, таких как распознавание изображений / речи, перевод и здравоохранение27,28, но оно имеет также проник в другие дисциплины, такие как материаловедение и квантовая механика29,30. Как показано на рис. 1, ИНС обычно состоит из набора входных искусственных нейронов, которые взаимосвязаны с несколькими скрытыми слоями и одним выходным слоем. Информация распространяется вперед посредством линейной операции, такой как умножение матриц, за которой следует нелинейная функция активации.

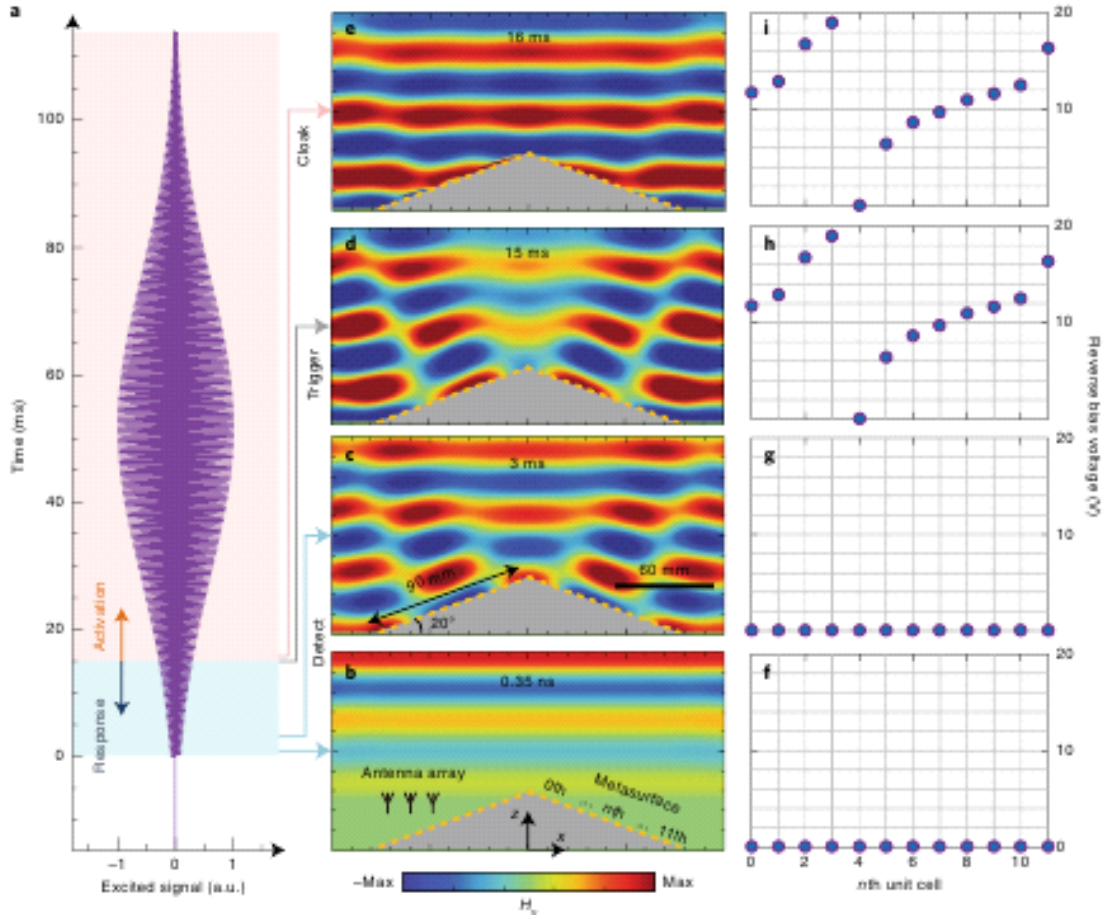
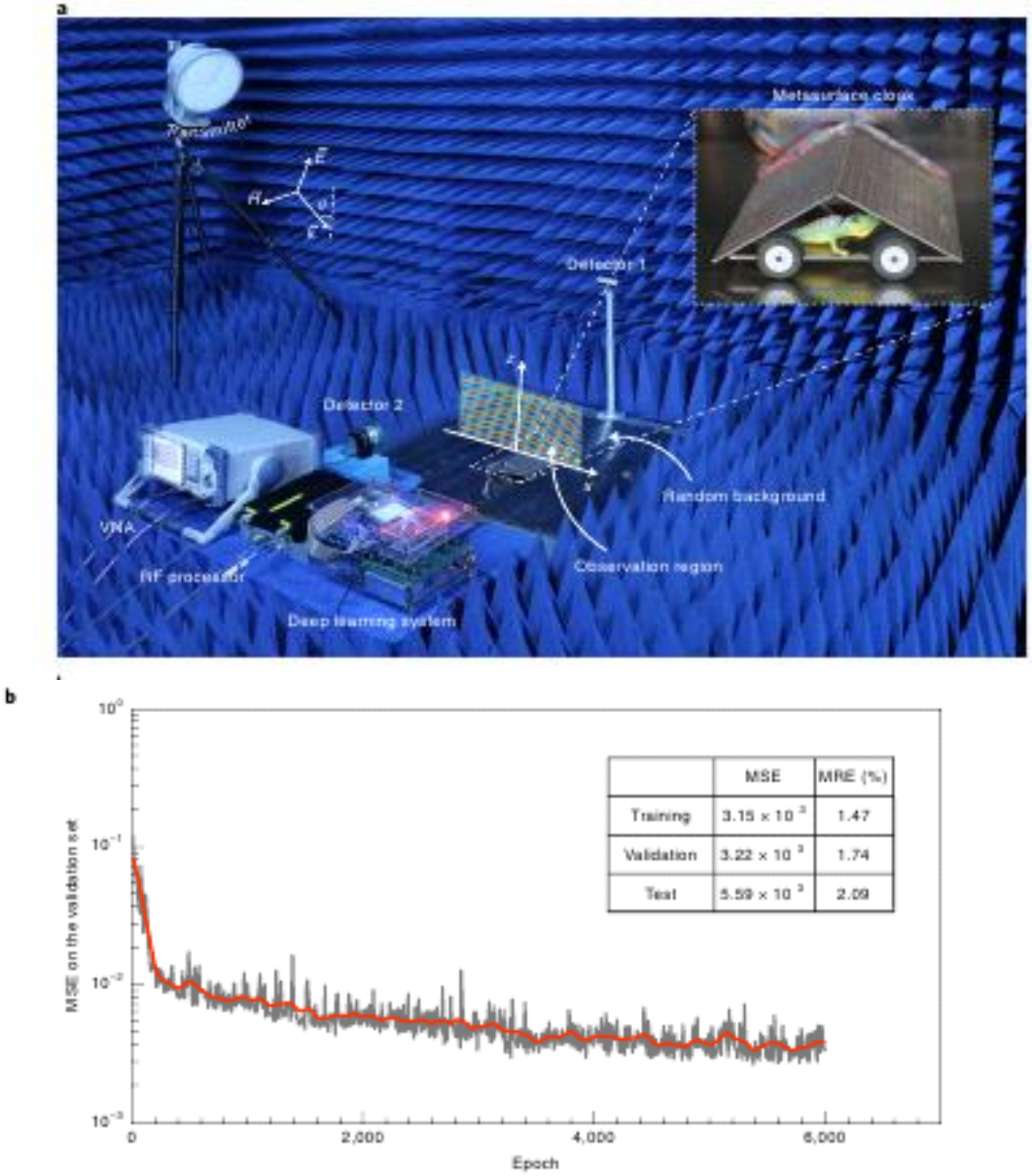


Рис. 2 Переходный отклик самоадаптивной маскировки в симуляциях FDtD. a – Гауссов импульс, излучаемый с z = 120 мм, падает на треугольный выступ tPEC, покрытый метаповерхностью; его форма волны удовлетворяет требованиям с центральной частотой fc = 8 ГГц, τ = 33 мс и 0 = 52 мс. b – e: временная эволюция (0,35 нс (b), 3 мс (c), 15 мс (d) и 16 мс (e)) магнитного поля Hy в области наблюдения (120 мм × 240 мм). Во время этого процесса информация об инциденте обнаруживается антенной решеткой и затем передается в предварительно обученную ИНС вместе с гипотетически известным фоном. При t = 15 мс (d) срабатывает маскировка метаповерхности, которая впоследствии делает выпуклость невидимой. f – i, распределения обратных напряжений смещения, соответствующие b – e, соответственно. Благодаря геометрической симметрии напряжения смещения вдоль двух скошенных кромок идентичны при нормальном падении.

Перед тем, как использовать ИНС, синаптические силы между каждым слоем должны быть адекватно обучены с помощью алгоритма обратного распространения, такого как градиентный спуск или адаптивный оптимизатор. переходный отклик интеллектуального плаща с использованием моделирования FDtD. Для концептуальной ясности мы рассматриваем двумерный (2D) ковровый плащ с метаповерхностью при поперечно-магнитном (TM или p-поляризованном) поляризованном освещении. Плащ с метаповерхностью тщательно разработан, чтобы скрыть выступ идеального электрического проводника (PEC) или другой объект в свободном пространстве. Для каждой элементарной ячейки обратный постоянный ток напряжение смещения в диапазоне от 0 до 20 В обеспечивается для управления емкостью варакторного диода и настройки спектра отражения (фаза отражения покрывает почти от -180 ° до 180 ° в рабочем диапазоне). Подробности об элементарной ячейке и ее модели эквивалентной схемы представлены в дополнительных примечаниях 1 и 2. Чтобы показать, как работает интеллектуальная маскировка, алгоритм FDTD, вставленный с предварительно обученной ИНС, используется для наблюдения за переходной реакцией интеллектуальной маскировки (см. Дополнительное примечание. 3 для моделирования). В сочетании с антенной решеткой для обнаружения падающих волн (см. Методы) и гипотетически известной окружающей средой, полный набор программы моделирования (то есть численный алгоритм ЭМ для маскировки, предварительно обученная ИНС и алгоритм обнаружения падающей волны) волны) создан для имитации реальной сцены. При моделировании на маскировку падает плоская TM-волна со случайной косинус-модулированной гауссовой формой волны; форма волны и физические изображения в разное время представлены на рис. 2. В исходном состоянии все напряжения смещения установлены на ноль. По прошествии времени падающая волна взаимодействует с выступом, сильно искажая рассеянные волны, как показано на рис. 2c. Через ~ 15 мс (практическое время отклика) детектор получает информацию о падающей волне и передает ее предварительно обученной ИНС. Плащ метаповерхности срабатывает (рис. 2d) и делает выпуклость невидимой (рис. 2e). Этот пример демонстрирует работу интеллектуальной маскировки во временной области. Эксперимент и обучение Для эксперимента мы выбрали материал F4B с диэлектрической постоянной 3,5 и тангенс угла потерь 0,003 в качестве подложки. Фотограф изготовленного прототипа (изготовленного по традиционной технологии печатных плат) представлена на рис. 3a. Размеры метаповерхностного включения 180 × 210 × 2 мм3 (24 × 28 элементарных ячеек); он разделен на две идентичные части. Каждый столбец (28 элементарных ячеек) вдоль направления y имеет одинаковое напряжение смещения, при этом варакторные диоды приварены между

металлическими разделенными I-образными островками на диэлектрической подложке.



	MSE	MRE (%)
Training	3.15×10^{-3}	1.47
Validation	3.22×10^{-3}	1.74
Test	5.59×10^{-3}	2.09

Рис. 3 | Экспериментальная установка и результаты обучения сети. а, Гауссов пучок с TM поляризацией от линзовой антенны с высокой направленностью падает на рассеиватель с углом падения θ. Область наблюдения (200 мм × 400 мм) находится на высоте 60 мм над землей, а геометрия маскировки соответствует изображенной на рис. 2. Окружающий фон и падающая волна обнаруживаются в реальном времени детекторами 1 и 2. соответственно (подробности см. в дополнительном примечании 5), подаваемые в систему глубокого обучения, и все напряжения смещения мгновенно вычисляются и передаются в маскировку метаповерхности. Этот процесс занимает ~ 15 мс. Врезка: фотография изготовленного прототипа. ВАЦ, векторный анализатор цепей; РФ, радиочастота. б, среднеквадратичная ошибка (MSE) проверки, установленная за период, показана серой кривой. Оранжевая кривая показывает сглаженную MSE. В таблице-вставке указаны окончательная MSE и средняя относительная ошибка (MRE) для трех перечисленных наборов

Вместо результатов моделирования мы экспериментально измерили спектр отражения для плоской метаповерхности при различных падающих волнах и напряжениях смещения19 в качестве обучающих выборок (дополнительное примечание 4). Подавленная амплитуда отражения метаатома на дополнительном рис. 4 в основном вызвана внутренним сопротивлением варакторного диода, которое может быть дополнительно улучшено с помощью микроволновой усиливающей среды31. Мы обучили ИНС соотношением {падающая волна и спектр отражения} → {напряжение смещения} для метаатомов с использованием ~ 10 000 измеренных выборок (дополнительное примечание 8), которые были разделены на наборы для обучения, проверки и тестирования (80%, 10% и 10% соответственно). Входные данные были нормализованы, перемешаны и затем переданы в сеть, так как такая обработка может ускорить сходимость алгоритма. Среднеквадратичная ошибка (MSE) использовалась для представления функции потерь между нормализованным и желаемым выходом, а потеря обучающего набора использовалась для генерации градиентов (чистое обучение). Гиперпараметры (например, количество скрытых слоев, нейронов и скорость обучения) были установлены в соответствии с производительностью на проверочном наборе18; В конечном итоге была выбрана полностью связная сеть с пятью скрытыми слоями и 60 нейронами на слой (дополнительное примечание 8).

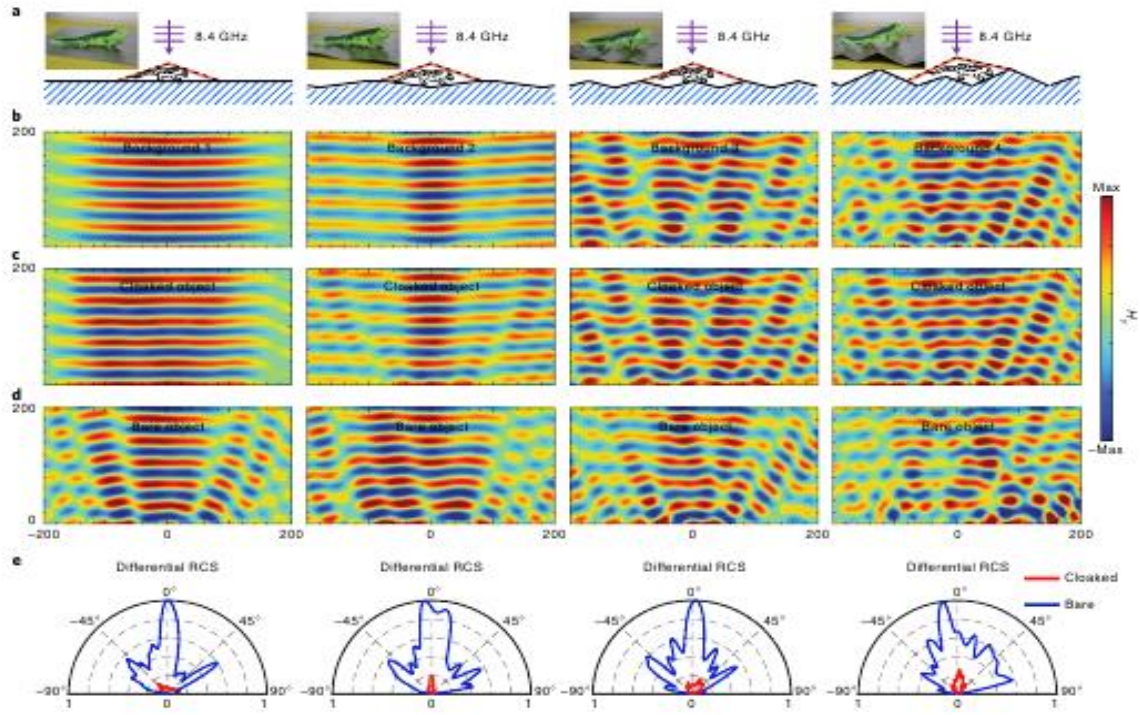


Рис. 4 Демонстрация самоадаптивного отклика маскировки на случайный фон для нормального падения волн на частоте 8,4 ГГц. а, Схема четырех фонов 1–4. Серая пунктирная линия, прикрепленная к красной линии, очерчивает плащ метаповерхности. Объект, используемый для демонстрации маскирующей способности, представляет собой пластиковую модель в форме

хамелеона, которую можно заменить на любой объект аналогичного размера. b – d: Распределение ближнепольного магнитного поля Ну фона (b), замаскированного объекта (c) и голого объекта (d), соответственно (в мм). д. Дифференциальная RCS в дальней зоне для скрытых и голых объектов для четырех случаев. Подробности измерений представлены в дополнительном примечании 6.

В иллюстративных целях средняя относительная ошибка (MRE) определяется как средний процент отклонения на точку напряжения смещения (путем нахождения ошибки между выходным сигналом и желаемым значением с последующей нормализацией по величине). Как указано во вставке таблицы на рис. 3b, MRE всех трех наборов меньше 2,1% и близко совпадают друг с другом, предполагая, что обученная ИНС является гладкой и почти не переобученной. Таким образом, мы можем с уверенностью заключить, что предварительно обученная ИНС заслуживает доверия и применима в этом сценарии. На рис. 3a представлена полная экспериментальная система, использованная для проверки работоспособности интеллектуальной маскировки, соответствующей схеме на рис. 1. Предварительно обученная ИНС и силовой модуль интегрированы вместе (обозначены как «система глубокого обучения»). Падающая волна (включая частоту и угол падения) и окружающая среда обнаруживаются в реальном времени высоконаправленной антенной Вивальди и камерой глубины, соответственно (подробности конструкции см. В дополнительном примечании 5). Сначала мы выводим требуемые спектры отражения для всех метаатомов (время преобразования настолько мало, что им можно пренебречь), а затем вводим их параллельно в предварительно обученную ИНС (дополнительный рис. 13). Время отклика, включая обнаружение (часть, которая занимает больше времени), время вычислений и инструкций ИНС, составляет ~ 15 мс. Это может быть дополнительно уменьшено путем использования радара с фазовой решеткой, например, с помощью алгоритма направления прибытия, такого как MUSIC или ESPRIT32,33. В нашем первом эксперименте мы исследовали сценарии с TM-поляризованным гауссовым лучом от линзовой антенны, падающим нормально на рассеивателе на четырех случайных фонах / формах рельефа (синие заштрихованные области) (рис. 4a). Распределение магнитного поля в ближней зоне по рельефу (рис. 4b, обозначено как «фон»), замаскированному объекту (рис. 4c) и голому объекту (рис. 4d) были сканированы (дополнительный рис. 9a). Объект, используемый для демонстрации маскирующей способности (здесь пластиковая модель в форме хамелеона), может быть изменен на любой произвольный объект аналогичного размера. Картины поля, создаваемые замаскированным объектом, аналогичны тем, которые генерируются окружающей средой, в резком контрасте с таковыми

для голого объекта, где рассеянные волны сильно искажены. Небольшие расхождения на крайних правых панелях на рис. 4b, c могут возникнуть из-за того, что детектор, который мы использовали, недостаточно точен, чтобы охарактеризовать сложный окружающий ландшафт; это можно улучшить, используя более совершенные детекторы, такие как лазерный радар. Чтобы количественно продемонстрировать характеристики маскировки, дифференциальное радиолокационное сечение (RCS) было определено как

$$\sigma_{\text{cloaked/bare}} = 2\pi\rho\left|H_y^{c/b} - H_y^{g}\right|^2, \text{ where } H_y^{c}, H_y^{b} \text{ and } H_y^{g} \text{ are the mag-}$$

соответственно11, а ρ установлено равным 1,3 м для приближения дальнего поля34 (дополнительный рисунок 9b). На рисунке 4e показано, что дифференциальная RCS замаскированного объекта сильно подавлена для всех четырех форм рельефа, при этом общее RCS уменьшено более чем на 85%. Мы протестировали плащ интеллекта во втором эксперименте (рис. 5), в котором одновременно менялись угол падения, частота и форма рельефа. Показано, что замаскированный объект успешно компенсирует внешнее освещение.

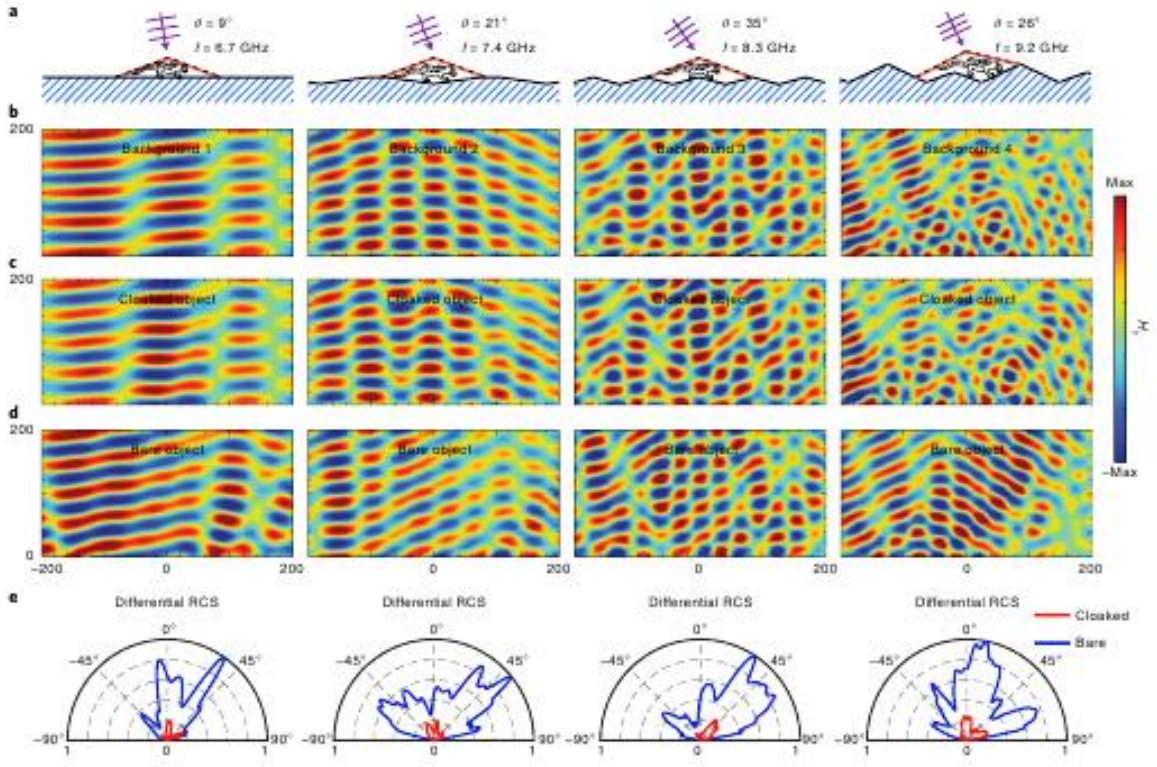


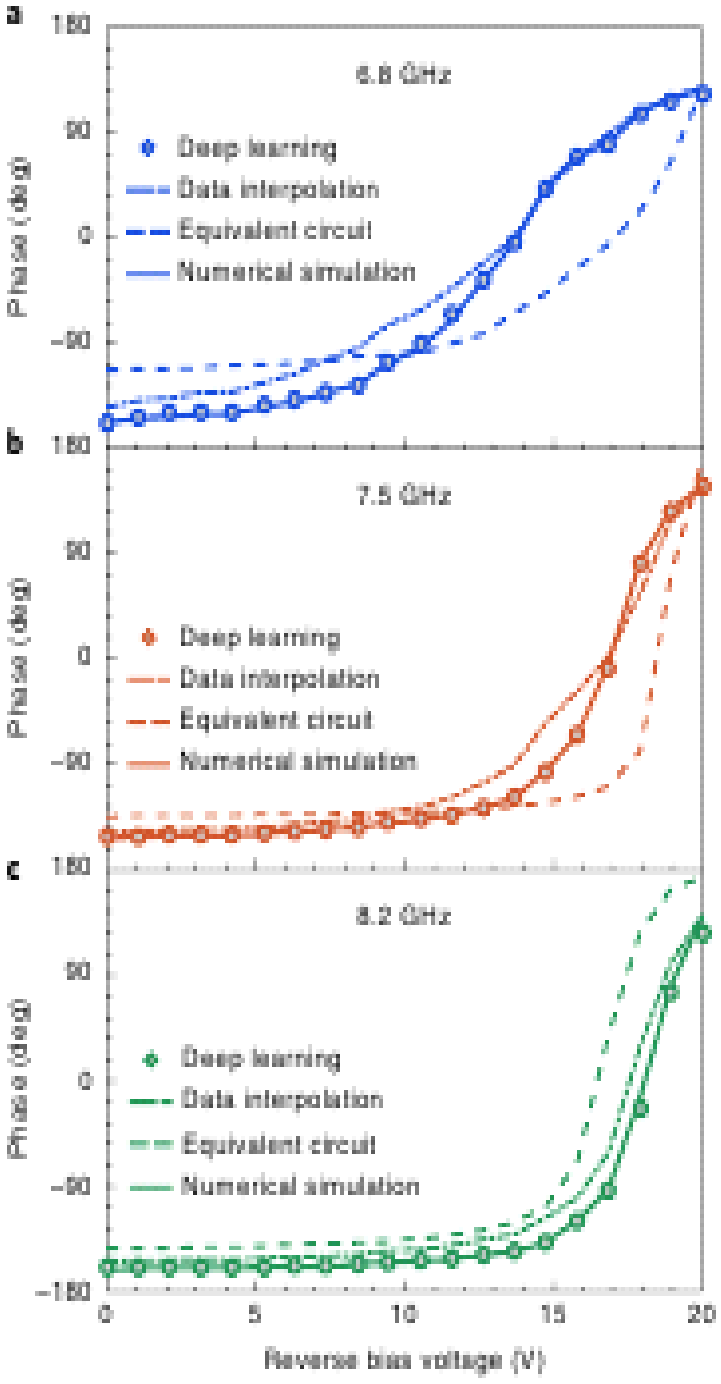
Рис. 5 | Демонстрация самоадаптивного отклика маскировки на случайные и одновременные изменения падающей волны и фона. a, Схема фона и падающей волны. b – d, Распределение ближнепольного магнитного поля Hy фона (b), замаскированного объекта (c) и голого объекта (d), соответственно (в мм). e, Дифференциальная RCS в дальней зоне для скрытого и открытого объекта для четырех случаев.

и вписывается в окружающую среду. В каждом эксперименте реакция завершается за миллисекунды, тогда как хамелеону требуется ~ 6 с. Эти эксперименты с множеством переменных убедительно подтверждают надежность нашего интеллектуального плаща. Дополнительное примечание 10 показывает больше результатов, а дополнительное видео 1 демонстрирует характеристики замаскированного транспортного средства, проезжающего через случайный рельеф, по сравнению с незакрытым случаем.

Обсуждение

На практике фоны постоянно меняются, что является большой проблемой для обобщения на ограниченных выборках, особенно для приложений распознавания и классификации изображений. Тем не менее, для маскировки метаповерхности фоновая информация может быть воплощена в серии спектров локального отражения, предоставляемых метаатомами (дополнительное примечание 7). Спектры локального отражения обобщены, и все случаи проиллюстрированы в нашем собранном наборе данных. Если мы будем рассматривать фон или индуцированные глобальные спектры отражения как входные данные ИНС, то сам сбор данных будет очень громоздким и обременительным, поскольку потребуются многочисленные образцы. Вот почему мы строим ИНС для отображения {падающей волны и спектра отражения} → {напряжение смещения} для отдельных элементарных ячеек и применяем это к каждой элементарной ячейке внутри маскировки метаповерхности (дополнительный рис. 13). В каждом эксперименте коллективное действие всех элементарных ячеек (в нашем случае 672), а не отдельной ячейки или нескольких элементарных ячеек, приводит к эффекту маскировки. 672 входа могут одновременно содержать данные из наборов для обучения, проверки и тестирования. Это нормально для глубокого обучения; например, в задаче перевода в тестовом наборе могут быть части фраз, которые существуют в обучающем наборе. Важно отметить, что, хотя все входные данные взяты из тестового набора, как и в эксперименте на крайней левой панели рис. 5c, эффекты маскировки также вполне удовлетворительны. Рабочий диапазон метаповерхностной маскировки с полосой пропускания 6,7–9,2 ГГц по своей природе ограничен диапазоном емкостей варакторного диода и резонансной структурой, а не ограничениями интеллектуальной системы. Как моделирование, так и экспериментальные результаты подтверждают этот факт. Стратегия активной маскировки35, которая использует массив элементарных источников для компенсации рассеянных полей, вызванных скрытым объектом, может пойти дальше, потенциально реализуя интеллектуальную маскировку в широкой полосе пропускания. Между тем концепция интеллектуальной

маскировки может масштабироваться до более высоких частот. На оптических частотах мы можем создавать маски с пространственной фазовой модуляцией с настраиваемыми затвором проводящими оксидными метаповерхностями, такими как оксид индия и олова36, или веществами с фазовым переходом, такими как германий-сурьма теллур37 и диоксид ванадия38 (фактические стратегии см. В дополнительном примечании 12). Кроме того, ограничение этой работы на одну поляризацию может быть преодолено путем гибридизации множества диодов с непараллельной или ортогональной ориентацией 39 вместе с более сложной конфигурацией сети. При проектировании наноструктур обычное полноволновое моделирование в значительной степени основывается на итерационных и длительных вычислениях уравнений Максвелла под руководством эксперта по оптике, пока не будет достигнуто локально оптимизированное решение. Альтернативный подход, модель эквивалентной схемы40, быстр, но используется эмпирически только для грубого анализа и для простых структур и случаев; таким образом, это невозможно для приложений, требующих высокой точности (дополнительное примечание 2), таких как нынешняя метаповерхностная маскировка. Глубокое обучение снимает эти ограничения, автоматически обнаруживая мощные комбинации нелинейных функций с непревзойденной точностью и эффективностью, как показано на рис. 6 (построено путем поворота предварительно обученной ИНС назад; дополнительное примечание 9). Мы также отмечаем, что преимущества глубокого обучения будут заметно проявляться в высокопроизводительных сценариях реального времени, на месте и сложных сценариях, таких как практическое 3D-маскирование и мультистатическое обнаружение. Для создания данных для ИНС может потребоваться некоторое время, но есть как минимум две причины, по которым этот метод все еще полезен. Во-первых, оборудование недорогое, и для экономии времени генерацию данных можно легко выполнять параллельно на разных машинах. В нашей работе 10000 образцов (в среднем, выборка только 10 раз по частотам 6–10 ГГц, в 20 раз по углам падения -90 ° -90 ° и в 50 раз по фазе отражения 0–2π) легко доступны с небольшими усилиями, получение предварительно обученной ИНС с высокой точностью. В дополнительном примечании 8 поясняется, что 10 000 выборок достаточно для нашей работы, а MRE неизменно низок (3,9%), даже несмотря на то, что обученные выборки сокращены вдвое (дополнительный рис. 14).



www.nature.com/naturephotonics

Рис. 6 спектр отражения метаповерхности, полученный разными методами. a – c: Фаза отражения метаповерхности в зависимости от приложенного напряжения смещения при нормальном падении на трех разных частотах: 6,8 ГГц (a), 7,5; ГГц (b) и 8,2 ГГц (c). Прогнозы ИНС (выполненные в миллисекундной шкале времени) хорошо совпадают с реальным спектром, полученным путем моделирования (полученным во второй или даже минутной шкале времени). Напротив, фазовые кривые, построенные с помощью линейной интерполяции данных и модели эквивалентной схемы, далеки от реальности, хотя время выполнения в миллисекундах является многообещающим. Более подробные обсуждения интерполяции данных и модели эквивалентной схемы приведены в дополнительных примечаниях 2 и 9.

Во-вторых, как показано на рис. 6, ошеломляющее повышение точности и эффективности стоит компромисса. Помимо этого, после того, как ИНС предварительно обучена, ее можно применять к другим функциям, таким как оптимизация структур и их обратное направление для прогнозирования предпочтительного для пользователя электромагнитного ответа.

Заключение

Мы предложили концепцию интеллектуальной маскировки, основанной на глубоком обучении, и реализовали маскировку метаповерхности в качестве примера. Плащ с метаповерхностью, встроенный в предварительно обученную ИНС, демонстрирует эффективную и надежную самонастраивающуюся способность реагировать на постоянно меняющуюся волну и фон без вмешательства человека. Одно вычисление с прямой связью в ИНС экономит значительное количество времени по сравнению с повторяющимися итерационными вычислениями, необходимыми в адаптивной антенне 41 и адаптивной оптике 42 (дополнительное примечание 11). Эта концепция была подтверждена симуляциями FDTD и экспериментальным микроволновым экспериментом, который переводит исследования маскировки на следующий этап – интеллектуальные маскировки. Предлагаемая концепция не только прокладывает путь к радикально новым метаустройствам, но и к другим областям больших данных, таким как оптимизация проектирования наноструктур по требованию15,16,18, решение обратной задачи ЭМ43 и использование скрытых физических знаний 44. В свою очередь, классическая электромагнетизм и оптика будут способствовать развитию глубокого обучения. Например, по сравнению с традиционными реализациями микроэлектроники, оптические ИНС, использующие нанофотонные схемы 45 и метаповерхности 46, могут предложить повышение скорости вычислений и энергоэффективности. Мы уже наблюдаем взаимодействие между этими двумя, казалось бы, не связанными друг с другом дисциплинами и ожидаем более интересных работ.

онлайн-контент

Любые методы, дополнительные ссылки, резюме отчетов Nature Research, исходные данные, расширенные данные, дополнительная информация, благодарности, информация экспертной оценки; сведения об авторском вкладе и конкурирующих интересах; заявления о доступности данных и кода доступны по адресу https://doi.org/10.1038/s41566-020-0604-2.

Поступило: 17 апреля 2019 г.; Принята в печать: 12 февраля 2020 г.; Опубликовано в Интернете: 23 марта 2020 г.

Ссылки

1. Пендри, Дж. Б., Шуриг, Д. и Смит, Д. Р. Управление электромагнитными полями. Science 312, 1780–1782 (2006).

2. Энгета Н. и Циолковски Р. В. Метаматериалы: физика и инженерные изыскания (Wiley, 2006).

3. Schurig, D. et al. Электромагнитная маскировка из метаматериала в микроволновом диапазоне Science 314, 977–980 (2006).

4. Цай В., Четтиар У. К., Кильдишев А. В., Шалаев В. М. Оптическая маскировка с использованием метаматериалов. Nat. Фотон. 1. С. 224–227 (2007).

5. Лэнди, Н. и Смит, Д. Р. Полнопараметрическая однонаправленная оболочка из метаматериала для микроволн. Nat. Mater.12. С. 25–28 (2013).

6. Лю Р. и др. Широкополосная маскировка заземляющего слоя. Science 323, 366–369 (2009).

7. Алё, А., Энгета, Н. Достижение прозрачности с помощью плазмонных и метаматериальных покрытий. Phys. Ред. E 72, 016623 (2005).

8. Эдвардс, Б., Алё, А., Сильвейринья, М. Г., Энгета, Н. Экспериментальная проверка плазмонной маскировки на микроволновых частотах с использованием метаматериалов. Phys. Rev. Lett. 103, 153901 (2009).

9. Yu, N. et al. Распространение света с фазовыми разрывами: обобщенные законы отражения и преломления. Science 334, 333–337 (2011).

10. Ni, X., Wong, Z.J., Mrejen, M., Wang, Y., Zhang, X. Ультратонкий плащ из кожи-невидимки для видимого света. Science 349, 1310–1314 (2015).

11. Yang, Y. et al. Полнополяризационная трехмерная метаповерхностная маскировка с сохраненными амплитудой и фазой. Adv. Mater. 28. С. 6866–6871 (2016).

12. Зигоняну Л., Попа Б. И. и Каммер С. А. Широкополосная всенаправленная акустическая наземная маскировка. Nat. Mater. 13. С. 352–355 (2014).

13. Фархат М., Гено С. и Энох С. Сверхширокополосная эластичная маскировка в тонких пластинах. Phys. Rev. Lett. 103, 024301 (2009).

14. Han, T. et al. Экспериментальная демонстрация двухслойного теплового плаща. Phys. Rev. Lett. 112, 054302 (2014).

15. Ма, В., Ченг, Ф. и Лю, Ю. Глубокое обучение позволило проектировать хиральные метаматериалы по запросу. АСУ Нано 12, 6326–6334 (2018).

16. Malkiel, I. et al. Дизайн и характеристика плазмонных наноструктур с помощью глубокого обучения. Light Sci. Appl. 7, 60 (2018).

17. Raccuglia, P. et al. Обнаружение материалов с помощью машинного обучения с помощью неудачных экспериментов. Nature 533, 73–76 (2016).

18. Peurifoy, J. et al. Моделирование нанофотонных частиц и обратный дизайн с использованием искусственных нейронных сетей. Sci. Adv. 4, eaar4206 (2018).

19. Li, H. et al. Двухдиапазонная зонная пластинчатая антенна Френеля с независимо управляемыми лучами. IEEE Trans. Антенны Propag. 66, 2113–2118 (2018).

20. Цуй Т.Дж., Ци М.К., Ван Х., Чжао Дж. И Ченг К. Кодирование метаматериалов, цифровых метаматериалов и программируемых метаматериалов. Light Sci. Appl. 3, e218 (2014)

21. Сюй К., Штюбиану Г. Т., Городецкий А. А. Адаптивные инфракрасные отражающие системы, вдохновленные головоногими моллюсками. Science 359, 1495–1500 (2018).

22. Талов А. и Хагнесс С. С. Вычислительная электродинамика: метод конечных разностей во временной области (Artech House, 2000).

23. Qian, C. et al. Переходный отклик сигнала через дисперсионную маску-невидимку. Опт. Lett. 41. С. 4911–4914 (2016).

24. Желудев Н.И., Кившар Ю.С. От метаматериалов к метаустройствам. Nat. Mater. 11, 917–924 (2012).

25. Шалтоут А. М., Шалаев В. М., Бронгерсма М. Л. Пространственно-временное управление светом с активными метаповерхностями. Наука 364, eaat3100 (2019).

26. Квон, Х., Сунас, Д., Кордаро, А., Полман, А., Алу, А. Нелокальные метаповерхности для обработки оптических сигналов. Phys. Rev. Lett. 121, 173004 (2018).

27. Шмидхубер, Дж. Глубокое обучение в нейронных сетях: обзор. Нейронная сеть. 61, 85–117 (2015).

28. Hinton, G. et al. Глубокие нейронные сети для акустического моделирования в распознавании речи: общие взгляды четырех исследовательских групп. Сигнальный процесс IEEE. Mag. 2012. Т. 29. С. 82–97.

29. Рампрасад, Р., Батра, Р., Пилания, Г., Манноди-Канаккитоди, А. и Ким, К. Машинное обучение в информатике материалов: недавние приложения и перспективы. npj Comput. Mater. 3, 54 (2017).

30. Карлео Г. и Тройер М. Решение квантовой задачи многих тел с помощью искусственных нейронных сетей. Science 355, 602–606 (2017).

31. Е, Д., Чанг, К., Ран, Л., Синь, Х. Усиливающая среда для микроволнового излучения с отрицательным показателем преломления. Nat. Commun. 5, 5841 (2014).

32. Константин, А. Б. Теория антенн: анализ и конструкция (Wiley, 2005).

33. Zhang, Y. & Ng, B. P. МУЗЫКАЛЬНАЯ оценка DOA без оценки количества источников. IEEE Trans. Сигнальный процесс. 58, 1668–1676 (2010).

34. Qian, C. et al. Экспериментальное наблюдение сверхрассеяния. Phys. Rev. Lett. 122, 063901 (2019).

35. Селванаягам М. и Элехериадес Г. В. Экспериментальная демонстрация активной электромагнитной маскировки. Phys. Ред. X 3, 041011 (2013).

36. Парк Дж., Канг Дж., Ким С., Лю Х. и Бронгерсма М. Фаза динамического отражения и управление поляризацией в метаповерхностях. Nano Lett. 17. С. 407–413 (2017).

37. Wang, Q. et al. Оптически реконфигурируемые метаповерхности и фотонные устройства на основе материалов с фазовым переходом. Nat. Фотон. 10. С. 60–65 (2016).

38. Tian, Z. et al. Реконфигурируемые наномембраны и микротрубки из диоксида ванадия с регулируемыми температурами фазовых переходов. Nano Lett. 18. С. 3017–3023 (2018).

39. Ma, X. et al. Активный метаматериал для манипулирования поляризацией. Adv. Опт. Mater. 2, 945–949 (2014).

40. Сарабанди К. и Бехдад Н. Частотно-избирательная поверхность с миниатюрными элементами. IEEE Trans. Антенны Propag. 55, 1239–1245 (2007).

41. Видроу Б., Мэнти П. Э., Гритс Л. Дж. И Гуд Б. Б. Адаптивные антенные системы. Proc. IEEE 55, 2143–2159 (1967).

42. Aeschlimann, M. et al. Адаптивное субволновое управление нанооптическими полями. Nature 446, 301–304 (2007).

43. Колтон Д. и Кресс Р. Теория обратного акустического и электромагнитного рассеяния (Springer, 1997).

44. Родригес-Ниева, Дж. Ф. и Шерер, М. С. Определение топологического порядка с помощью машинного обучения без учителя. Nat. Phys. 15. С. 790–795 (2019).

45. Хьюз, Т. В., Минков, М., Ши, Ю. и Фан, С. Обучение фотонных нейронных сетей с помощью обратного распространения распространения на месте и измерения градиента. Оптика 5, 864–871 (2018).

46. Lin, X. et al. Полностью оптическое машинное обучение с использованием глубоких нейронных сетей. Science 361, 1004–1008 (2018).

ПОСЛЕСЛОВИЕ

Одним из основных направлений современных когнитивных исследований является искусственный интеллект (ИИ). Под нейроморфными системами понимаются модели искусственных нейронных сетей, архитектура и дизайн которых основаны на особенностях структуры и принципах работы реальных нейробиологических систем.

Нейроморфное моделирование находится на пересечении нескольких областей исследований, в том числе нейробиологии, теории нейронных сетей, математического моделирования, электронной техники, кремниевой фотоники, создании адаптивных метаматериалов.

Связь между фотоникой и вычислениями является столпом современной оптики и предметом передовых исследований.

Основополагающая работа Хинтона и др. в 2006 году, они ввели название техники «Глубокое обучение». (Точностью распознавания цифр (> 98%)). Примеры такого же подхода включают в себя нейроморфную сеть Цурикова А.Н., см. патент РФ, и обучение импульсных нейронных сетей. При этом были получены хорошие результаты для разного количества нейронных сетей в ансамбле: 1, 4, 16 и 64. Получены вероятности правильного распознавания от 92.7% до 99.42% [56], а также математическая модель спайковых нейронов Ижикевича (Izhikevich) и крупномасштабное моделирование Элиасмита (Eliasmith).

После чтения статей сборника лишний раз убеждаешься в том, что ИТ – это практическая дисциплина, нацеленная на **результат**.

Ещё раз следует отметить, что данный сборник стал возможен благодаря личному энтузиазму Геннадия Семеновича Мельникова.

АВТОРЫ СБОРНИКА

Мельников Геннадий Семёнович, с.н.с. Государственного Оптического Института 1970-2018 г.г, с.н.с. АО «НПП «АМЭ», 2018-2019 г.г, Россия · Санкт-Петербург, E-mail: gmelnikov HYPERLINK "mailto:gmelnikov@list.ru"@ HYPERLINK "mailto:gmelnikov@list.ru"list HYPERLINK "mailto:gmelnikov@list.ru". HYPERLINK "mailto:gmelnikov@list.ru"ru

Мельникова Элеонора Ильинична, инженер ВНИИТУ 1969-1997, в настоящее время-пенсионер, Россия · Санкт-Петербург, E-mail: melnikovaeleonora HYPERLINK "mailto:melnikovaeleonora@list.ru"@ HYPERLINK "mailto:melnikovaeleonora@list.ru"list HYPERLINK "mailto:melnikovaeleonora@list.ru". HYPERLINK "mailto:melnikovaeleonora@list.ru"ru

Самков Владимир Михайлович, нач.лаб. Государственного Оптического Института 1970-2018 г.г, в.н.с. АО «НПП «АМЭ», 2018-2021 г.г, Россия · Санкт-Петербург, E-mail: samkovvm HYPERLINK "mailto:samkovvm@yandex.ru"@ HYPERLINK "mailto:samkovvm@yandex.ru"yandex HYPERLINK "mailto:samkovvm@yandex.ru". HYPERLINK "mailto:samkovvm@yandex.ru"ru

Тупиков Владимир Алексеевич, д.т.н., директор Государственного Оптического Института 2008-2011 г.г., зам. Генерального директора АО «НПП «АМЭ», 2011-2021 г.г, Россия · Санкт-Петербург, E-mail: tupikov HYPERLINK "mailto:tupikov@nppame.ru"@ HYPERLINK "mailto:tupikov@nppame.ru"nppame HYPERLINK "mailto:tupikov@nppame.ru". HYPERLINK "mailto:tupikov@nppame.ru"ru

Бондаренко Владимир Александрович, начальник центра интеллектуальной обработки изображений, АО «НПП «АМЭ», Россия · Санкт-Петербург, E-mail: bondarenko@nppame.ru

Чураков Вадим Сергеевич, к.ф.н., доцент, Шахтинский филиал Южно-Российского гуманитарного института. E-mail: v.s.chur@mail.ru

Тотович Ангелина, факультет информатики Университета Аристотеля в Салониках;

Дабос Джордж, факультет информатики Университета Аристотеля в Салониках;

Пассалис Николаос, факультет информатики Университета Аристотеля в Салониках;

Тефас Анастасиос, факультет информатики Университета Аристотеля в Салониках;

Плерос Никос, факультет информатики Университета Аристотеля в Салониках;

Маркес Бики А., кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Филипович Мэтью Дж., кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Ховард Эмма Р., кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Бангари Вирадж, кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Гуо Чжиму, кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Морисон Хью Д., кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада;

Феррейра де Лима, кафедра электротехники, Принстонский университет, Принстон, Нью-Джерси 08544, США;

Тейт Александр Н., кафедра электротехники, Принстонский университет, Принстон, Нью-Джерси 08544, США; отдел прикладной физики, Национальный институт стандартов и технологий, Боулдер;

Прунал Пол Р., кафедра электротехники, Принстонский университет, Принстон, Нью-Джерси 08544, США;

Шастри Бхавин Дж., кафедра физики, инженерной физики и астрономии, Королевский университет, Кингстон, ON KL7 3N6, Канада; кафедра электротехники, Принстонский университет, Принстон, Нью-Джерси 08544, США;

Вэн Цзинкай, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр совместных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Дин Юйцзян, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр совместных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Ху Чэнбо, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр совместных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Чжу Сюэ-Фэн, школа физики и инновационный институт Хуачжунского университета науки и технологий, 430074 Ухань, Хубэй, П. Р. Китай;

Лян Бинь, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр совместных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Ян Цзин, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр совместных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Чэн Цзяньчунь, ключевая лаборатория современной акустики, Министерство энергетики, Институт акустики, Департамент физики, Центр сов-

местных инноваций передовых микроструктур, Нанкинский университет, 210093 Нанкин, Китай;

Цзо Вэйвэнь, Государственная ключевая лаборатория передовых систем и сетей оптической связи, Интеллектуальный центр интеграции микроволнового светового излучения (iMLic), Департамент электронной инженерии, Шанхайский университет Цзяо Тонг, Шанхай 200240, Китай;

Ма Боуэн, Государственная ключевая лаборатория передовых систем и сетей оптической связи, Интеллектуальный центр интеграции микроволнового светового излучения (iMLic), Департамент электронной инженерии, Шанхайский университет Цзяо Тонг, Шанхай 200240, Китай;

Сю Шаофу, Государственная ключевая лаборатория передовых систем и сетей оптической связи, Интеллектуальный центр интеграции микроволнового светового излучения (iMLic), Департамент электронной инженерии, Шанхайский университет Цзяо Тонг, Шанхай 200240, Китай;

Цзо Сютин, Государственная ключевая лаборатория передовых систем и сетей оптической связи, Интеллектуальный центр интеграции микроволнового светового излучения (iMLic), Департамент электронной инженерии, Шанхайский университет Цзяо Тонг, Шанхай 200240, Китай;

Ван Синцзюнь, Государственная ключевая лаборатория передовых систем и сетей оптической связи, Школа электроники и компьютерных наук, Пекинский университет, Пекин 100871, Китай;

Малашин Роман О., Группа нейронных сетей и искусственного интеллекта, Институт физиологии им. Павлова РАН; Государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия;

Ву Алли Х., Принстонский университет, Принстон, Нью-Джерси, США;

Нахмиас Митчелл А., Принстонский университет, Принстон, Нью-Джерси, США;

Пэн Суан-Тунг, факультет электротехники, Принстонский университет, Принстон, штат Нью-Джерси, 08544 США;

Хуан Чаоран, факультет электротехники, Принстонский университет, Принстон, штат Нью-Джерси, 08544 США;

Санни Фебин П., госуниверситет Колорадо;

Тахери Эбадолла, госуниверситет Колорадо;

Никдаст Махди, госуниверситет Колорадо;

Пасрича Судеп, госуниверситет Колорадо;

Панда Сумьяшри С., Индийский технологический институт, Гандинагар, Гуджарат, Индия, 382355;

Хегде Рави С., Индийский технологический институт, Гандинагар, Гуджарат, Индия, 382355;

Бянь Тунь, Школа наук, Китайский университет геолого-геофизических исследований, Пекин, 100083, Китай; Школа информационной инженерии Китайского университета геолого-геофизических исследований, Пекин, 100083, Китай;

Йи Юйсюань, Школа информационной инженерии Китайского университета геолого-геофизических исследований, Пекин, 100083, Китай;

Ху Цзяле, Школа информационной инженерии Китайского университета геолого-геофизических исследований, Пекин, 100083, Китай;

Чжан Инь, Школа дистанционного зондирования и информационной инженерии, Уханьский университет, Ухань 430072, Китай;

Ван Иде, Школа информационной инженерии Китайского университета геолого-геофизических исследований, Пекин, 100083, Китай;

Гао Лю, Школа наук, Китайский университет геолого-геофизических исследований, Пекин, 100083, Китай;

Расмуссен Торстен С., Технический университет Дании, DK-2800 Kongens Lyngby, Дания;

Йи Ю., Технический университет Дании, DK-2800 Kongens Lyngby, Дания;

Морк Джеспер, Технический университет Дании, DK-2800 Kongens Lyngby, Дания;

Мургиас-Александрис Джордж, факультет информатики Университета Аристотеля в Салониках.

Научное издание

НЕЙРОМОРФНЫЕ ФОТОННЫЕ СИСТЕМЫ

Научн. ред. Г.С.Мельников

Ростов-на Дону, редакция научной литературы.
Подписано в печать 11.03.2019 г.
Формат 60x84 1/16. Бумага офсетная. Печать цифровая.
Печ. л. 8,7. Тир. 500 экз.