Декабрь 2023 г. *Том 193, № 12*

УСПЕХИ ФИЗИЧЕСКИХ НАУК

ОБЗОРЫ АКТУАЛЬНЫХ ПРОБЛЕМ

Подходы фотоники для реализации нейроморфных вычислений

А.И. Мусорин, А.С. Шорохов, А.А. Чежегов, Т.Г. Балуян,
К.Р. Сафронов, А.В. Четвертухин, А.А. Грунин, А.А. Федянин

Физические ограничения скорости работы электронных устройств заставляют искать альтернативные способы обработки информации. Последние несколько лет развивается нейроморфная фотоника — раздел фотоники, объединяющий физику оптических и оптоэлектронных устройств с математическими алгоритмами искусственных нейронных сетей. Подобный симбиоз позволяет решать определённые классы вычислительных задач, в частности некоторые задачи искусственного интеллекта, с большей скоростью и высокой энергоэффективностью по сравнению с электронными устройствами на архитектуре фон Неймана. Рассмотрены оптические аналоговые вычисления, фотонные нейронные сети, способы перемножения матриц оптическими средствами, обсуждаются преимущества и недостатки существующих подходов.

Ключевые слова: нейроморфная фотоника, искусственный интеллект, машинное обучение, резервуарные вычисления, матрично-векторное умножение, фотонные вычисления, нейронные сети, оптический сопроцессор, фотонные тензорные вычисления, оптическое фурье-преобразование, интегральная фотоника, интерферометр Маха – Цендера, кольцевые резонаторы, волноводы

PACS numbers: 07.05.Mh, 42.79.Hp DOI: https://doi.org/10.3367/UFNr.2023.07.039505

Содержание

1. **Введение (1284).**
2. **Оптические вычисления на платформе интегральной фотоники (1285).**
   2.1. Вектор-матричные операции и их реализация методами интегральной фотоники. 2.2. Фотонная машина Изинга на интегральной платформе.
3. **Оптические вычисления в свободном пространстве (1292).**
   3.1. Дифракционные нейронные сети. 3.2. Оптическое преобразование Фурье.
4. **Заключение (1296).**

Список литературы (1297).

1. Введение

XXI век неразрывно связан с индустрией информационных технологий (ИТ). Взрывной рост количества информации, обусловленный появлением социальных сетей, тенденцией к облачному типу хранения данных, развитием интернет-ресурсов и сферы развлечений, повышением безопасности банковских и финансовых операций, стимулировал развитие новых способов и алгоритмов обработки и передачи данных, в том числе с помощью света [1]. Стремительная экспансия ИТ во все сферы деятельности требует внушительного количества вычислительных ресурсов и мощностей, что в свою очередь делает актуальными вопросы ускорения работы процессоров, разработки новых компьютерных архитектур, уменьшения энергопотребления, миниатюризации систем [2 – 5]. Решения ряда упомянутых проблем можно достичь с помощью фотоники [6]. Она привлекает внимание учёных благодаря высокой частоте электромагнитных волн, широкой полосе пропускания, возможности распараллеливания. Кроме того, прогресс в технологиях промышленного изготовления микропроцессоров и оптоэлектронных компонентов привёл к появлению частных компаний, готовых создавать интегральные фотонные чипы. Востребованность таких устройств обусловлена формированием рынка оптических вычислений из-за увеличения числа заказчиков, заинтересованных в модернизации своих оптоволоконных линий связи и центров обработки данных. Увеличение числа потребителей приводит к росту экономики в данной сфере и удешевляет производство фотонных элементов за счёт массовости. В связи с этим выполнение математических операций не электронным образом в цифровом виде, а при помощи аналоговой оптической обработки сигнала может быть реализовано быстрее, проще и дешевле, что стимулирует развитие нейроморфной фотоники [7].

Методы фотоники могут успешно применяться для дополнения электронных процессоров. Световой сигнал, особенно в свободном пространстве, обладает ценными

А.И. Мусорин, А.С. Шорохов, А.А. Чежегов, Т.Г. Балуян, К.Р. Сафронов, А.В. Четвертухин, А.А. Грунин, А.А. Федянин [(a)]
Московский государственный университет им. М.В. Ломоносова, физический факультет,
Ленинские горы 1, стр. 2, 119991 Москва, Российская Федерация
E-mail: [(a)] fedyanin@nanolab.phys.msu.ru

Статья поступила 8 ноября 2022 г., после доработки 4 июля 2023 г.

для вычислений физическими свойствами. С помощью оптических систем, используя линзы, можно осуществлять преобразование Фурье, использовать явление интерференции для сложения комплексных величин, использовать явление дифракции для преобразования исходного сигнала, производить нелинейное квадратичное преобразование при детектировании и складывать интенсивности [8 – 10]. Кроме того, при работе с широкими неоднородными параллельными пучками, представляющими в сечении многомерную матрицу, возможны операции над всей матрицей со скоростью, не зависящей от её размера. Указанные свойства оптического сигнала, несущего информацию, позволяют ускорить в первую очередь операцию вектор-матричного и матрично-матричного умножения, составляющую основу исполнения и обучения многослойных нейронных сетей.

Узкими местами применения оптического сопроцессора являются операции ввода и вывода данных [11]. Цифро-аналоговые и аналого-цифровые преобразования (ЦАП и АЦП) вместе с обвязкой зачастую являются ограничивающим фактором и по скорости, и по энергопотреблению. Преимущество оптики состоит в возможности параллельной обработки информации, в работе с изображениями высокого разрешения.

Цель настоящего обзора — описать разработанные методы фотонных вычислений, способы умножения матриц оптическими средствами, рассмотреть современные подходы к реализации фотонных нейронных сетей.

2. Оптические вычисления на платформе интегральной фотоники

2.1. Вектор-матричные операции и их реализация методами интегральной фотоники

Интегральная фотоника является одной из наиболее перспективных платформ для реализации оптических вычислений в промышленном масштабе. Ключевым фактором является то, что подобные устройства могут быть созданы на существующей базе микроэлектронной промышленности с использованием хорошо отработанных методов полупроводниковой технологии. Согласно исследованиям Yole Group [12, 13], рынок интегральной кремниевой фотоники увеличится почти до 4 млрд долларов к 2025 г., при этом аппаратные ускорители и фотонные интерпозеры (кремниевый чип, основная роль которого состоит в электрическом соединении дорожек между памятью и процессором) могут занять в нём значительную часть благодаря таким крупным игрокам, как Nvidia. На данный момент существует целый ряд компаний, выпускающих на заказ интегральные фотонные чипы на основе кремния на изоляторе и нитрида кремния, однако ключевыми сложностями до сих пор остаются лазеры на чипе (совмещение кремния с III – V полупроводниками), а также объединение фотоники и КМОП-электроники (комплементарная структура металл – оксид – полупроводник; англ. CMOS, complementary metal – oxide – semiconductor) в едином устройстве. Многие производители уже достигли существенного прогресса, о чём можно судить по недавним заявлениям, например, Global Foundries [14]. Среди компаний, специализирующихся на решениях в области аппаратных ускорителей на основе интегральной фотоники, можно выделить американские LightMatter и Lightelligence, Fathom computing, а также британскую Salience labs. Наиболее распространённой операцией для ускорения является вектор-матричное произведение, однако также существуют примеры для решения комбинаторных задач [15]. Фотонные аппаратные ускорители на платформе интегральной фотоники могут быть встроены непосредственно в сетевые решения, например в крупных центрах обработки данных, где связь между вычислительными кластерами осуществляется за счёт оптических межсоединений. Это даёт дополнительное преимущество по сравнению с аналогами в задачах облачных вычислений, а также в задачах обеспечения безопасности и предотвращении кибернетических атак [16].

2.1.1. Различные архитектуры фотонных аппаратных ускорителей и их особенности. Чаще всего отмечают три основные архитектуры для реализации вектор-матричных вычислений с помощью интегральной фотоники: а) фотонная сеть на основе интерферометров типа Маха – Цендера (ИМЦ) — аналогичные фотонные матрицы используются в реализациях оптических интегральных квантовых компьютеров; б) фотонная сеть на основе микрокольцевых резонаторов; в) фотонная сеть по типу кроссбар-массива с использованием оптических мемристоров — топологии полносвязной коммутационной матрицы на основе матричного коммутатора (от англ. crossbar). Полносвязной матрица называется потому, что любой входной порт может быть связан с любым выходным портом по аналогии с мемристивными электрическими цепями. Первая архитектура выделяется своей универсальностью и слабой чувствительностью к погрешностям изготовления, однако требует большого числа элементов и занимает много пространства на чипе, что негативно сказывается на возможностях её масштабирования. Вторая архитектура на основе микрокольцевых резонаторов может быть реализована в гораздо более компактном исполнении, она изначально оптимизирована для параллельной работы с несколькими каналами по длине волны (мультиплексирование по длине волны, от англ. wavelength division multiplexing, WDM), но при этом очень чувствительна к отклонениям при производстве и требует специальной подстройки (пассивной или активной). Кроссбар-массив также обладает преимуществами работы с WDM, может быть реализован в компактном формфакторе и не страдает от разброса параметров при изготовлении, но ограничивается эффективностью кросс-соединений и делителей, которые привносят большие оптические потери и сокращают возможности для масштабирования фотонной цепи для реальных приложений. Рассмотрим каждую из указанных архитектур подробнее.

2.1.2. Фотонная матрица на основе интерферометров Маха – Цендера. Первой и наиболее распространённой архитектурой на данный момент является фотонная матрица интерферометров типа Маха – Цендера, которая может реализовать любую заранее заданную матрицу (например, матрицу весов для полносвязной нейронной сети или матрицу ядра для свёрточных нейронных сетей). Известно, что любую матрицу с помощью сингулярного разложения можно представить в виде произведения двух унитарных и одной диагональной

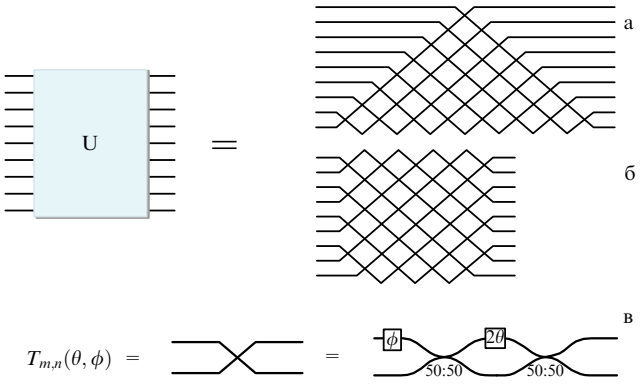

Рис. 1. Представление унитарной матрицы в виде сетки ИМЦ. Каждый узел сетки представляет собой 2×2 интерферометр, свойства которого определяются внешней и внутренней фазой в одном из плеч (φ и θ соответственно) [17].

матриц:

$$M = UZV^T,$$

где M — исходная матрица, U и V — унитарные матрицы, Z — диагональная матрица. В свою очередь, унитарная матрица может быть представлена в виде произведения матриц поворота (рис. 1), реализуемых с помощью сетки интегральных ИМЦ [17]. Данный подход широко применяется также в задачах оптических квантовых вычислений (например, при квантовой нормализации) как на чипе, так и в трёхмерном исполнении на оптическом столе.

Каждый интерферометр 2×2 может быть описан матрицей поворота вида

$$T = \mathrm{i}\exp\left(\frac{\mathrm{i}\theta}{2}\right)\begin{pmatrix} \exp(\mathrm{i}\varphi)\sin\theta/2 & \cos\theta/2 \\ \exp(\mathrm{i}\varphi)\cos\theta/2 & -\sin\theta/2 \end{pmatrix}.$$

Следуя процедуре, описанной в работе [17], можно составить фотонную сеть из подобных интерферометров, соответствующую любой унитарной матрице размерности $N \times N$ [18]. В общей сложности для этого потребуется $N(N-1)/2$ ИМЦ элементов (рис. 2). Для реализации любой матрицы по принципу сингулярного разложения потребуется две такие сетки, а также средний промежуточный слой амплитудных модуляторов, которые реализуют диагональную матрицу. Дополнительный слой фазовых модуляторов нужен для коррекции первой унитарной матрицы. Это необходимо делать в случае физической реализации всей матрицы M, однако на практике часто используют одну сеть ИМЦ для последовательной реализации матриц U и V (в таком случае фазовая коррекция не обязательна).

Помимо самой фотонной матрицы ИМЦ требуется также устройство для формирования входного вектора, закодированного в амплитуде световой волны в каждом из каналов сетки. Для этого внешний лазерный источник непрерывной генерации разделяется на нужное число входов, в каждом из которых устанавливается амплитудный модулятор, управляемый внешним электрическим драйвером, задающим с помощью ЦАП последовательность данных из памяти. Фазы ИМЦ фотонной матрицы также задаются с помощью аналогичных драйверов по схожей процедуре. Прошедший через фотонную матрицу свет детектируется фотодиодами, которые через трансимпедансные усилители передают далее сигнал на АЦП. Преобразователи возвращают его в цифровое представление и сохраняют данные во внешнюю память.

Основным ограничением частоты работы устройства становится предельная частота драйверов с ЦАП – АЦП. Сами по себе модуляторы могут работать на частотах в десятки, а в перспективе даже сотни ГГц. Поэтому часто используют быстрые и энергоэффективные ЦАП – АЦП с низким разрешением. Существенного выигрыша в скорости и уменьшении потребляемой энергии удаётся достичь только для задач, где возможно многократное использование одной и той же матрицы весов (как в случае задач свёрточных нейронных сетей, где ядро свёртки фиксировано), а также параллельное вычисление на нескольких длинах волн. Из-за погрешностей при изготовлении, а также шума и отклонений при работе устройства решения, существующие на данный момент, работают с низкоразрядными числами (до 8 бит), однако для широкого круга задач этого оказывается достаточно.

Впервые представленная архитектура в аппаратных ускорителях для задач глубокого машинного обучения

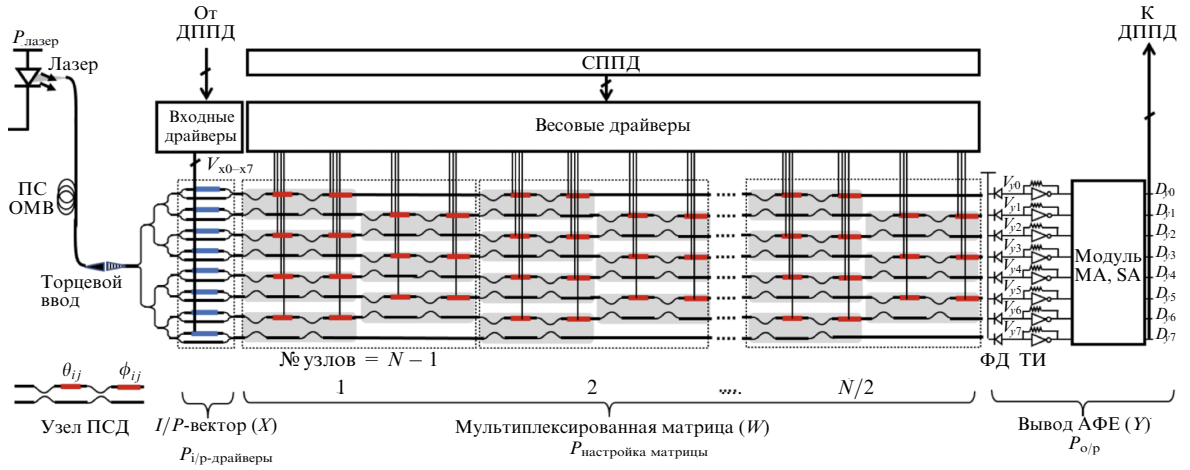

Рис. 2. Схематическое изображение архитектуры фотонного ускорителя на основе ИМЦ. Представленная сеть интерферометров реализует одну унитарную матрицу. Единичный ИМЦ выделен серым фоном, модуляторы, задающие φ и θ, показаны красным. Синим отмечены амплитудные модуляторы, формирующие входной вектор [18].

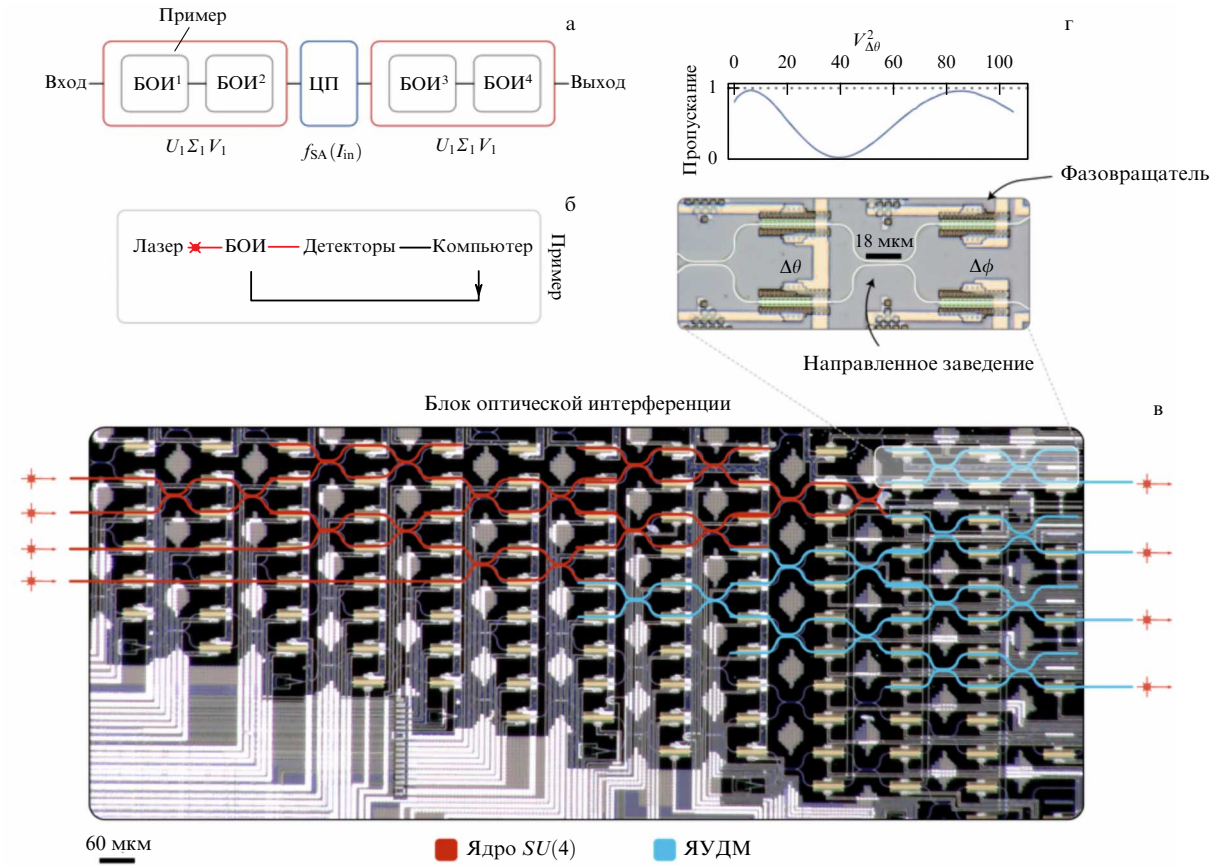

Рис. 3. Пример использования ИМЦ-архитектуры для реализации полносвязной нейронной сети для задач классификации [19].

была предложена группой Марина Сольячича (M. Soljacic) в статье [19]. В качестве модели была рассмотрена полносвязная нейронная сеть для классификации входных данных (кодированные в световом сигнале звуковые последовательности, соответствующие разным буквам). Сеть состояла из двух слоёв, отвечающих умножению входного вектора на матрицы весов, с промежуточным слоем нелинейной активации, реализованным отдельно с помощью стандартного микропроцессора. Вектор-матричное произведение происходило с помощью фотонной матрицы, при этом одна сеть ИМЦ использовалась дважды для последовательного задания двух унитарных матриц (БОИ1 и БОИ2, рис. 3). Сеть была предобучена стандартным методом обратного распространения ошибки на компьютере, после чего по найденным весам были рассчитаны фазы φ и θ для каждого ИМЦ, формирующего фотонную матрицу (по методу, описанному в работе [17]). Точность решения задачи классификации в такой системе составила порядка 77 % по сравнению с 91 %, полученным с помощью стандартного центрального процессора. Низкая эффективность может быть связана с рядом факторов, среди которых стоит отметить тепловые наводки в элементах ИМЦ со стороны соседних ячеек, конечную точность задания фазовых значений в разных плечах интерферометров, шумы детекторов при считывании, а также отклонение геометрических параметров структуры при изготовлении от модельных. Можно проводить корректировку ошибок в уже изготовленной структуре [20], что потенциально может улучшить точность решения задачи классификации в подобных устройствах.

Интерферометры Маха – Цендера, проект LightMatter. Один из примеров коммерческой реализации данной архитектуры был представлен американской компанией LightMatter, которая была основана бывшими выпускниками группы Марина Сольячича. На данный момент компания зарегистрировала более 30 патентов, связанных с фотонными аппаратными ускорителями на основе интегральных оптоэлектронных устройств. Актуальным продуктом является процессор Envise [21], представленный на официальном сайте компании, со спецификациями и характерными результатами тестирования производительности (рис. 4). Основной частью устройства являются две фотонные матрицы на основе ИМЦ, изготовленные по кремниевой технологии стандарта 90 нм. Источником света выступает внешний лазер, на чип излучение заводится по массиву оптических волокон. Оптическое ядро совмещается с электронной логической КМОП-микросхемой, изготовленной по технологии 12 нм. Данная микросхема основана на RISC-V-архитек-

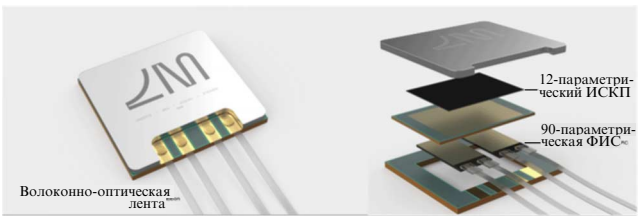

Рис. 4. Оптоэлектронный аппаратный ускоритель ENVISE от LightMatter [21].

туре и содержит до 256 вычислительных ядер. Также на чипе размещается SRAM-память объёмом до 500 Мб. Данные могут быть представлены в различном формате, есть возможность работы с 8-битным и 16-битным представлением. Кроме того, существуют возможности для интеграции с наиболее популярными программными платформами (PyTorch, TensorFlow). Такой аппаратный ускоритель может использоваться для серверных решений, а ключевыми операциями являются вектор-матричные и тензорные вычисления. Высокая скорость работы и малое энергопотребление достигается за счёт ряда факторов, среди которых стоит выделить WDM-мультиплексирование (до восьми параллельных каналов), а также микромеханические интегральные модуляторы NOEMS (nano-opto-electro-mechanical system) [22 – 24], существенно уменьшающие потребление энергии по сравнению с термооптическими модуляторами, а также модуляторами на основе инжекции носителей, при этом обеспечивая высокую частоту переключения до 1 МГц. Кроме того, данный модулятор позволяет существенно увеличить плотность ИМЦ на чипе за счёт своего компактного размера (порядка 25 мкм). Стоит также отметить, что для обеспечения высокой частоты операций (сэмплинга) используются скоростные модули ЦАП – АЦП со средним разрешением: 8- или 16-битным представлением данных. Это может быть оправдано в большинстве приложений машинного обучения и искусственного интеллекта, для которых более грубое представление данных не сказывается на итоговой точности решения задачи [25, 26].

Интерферометры Маха – Цендера, проект Lightelligence. Аналогичное аппаратное решение разрабатывается компанией Lightelligence [27], которая также аффилирована с группой Сольячича. Как и в случае Envise, используется массив ИМЦ для реализации фотонной матрицы 64×64. Переход между цифровым электронным и аналоговым фотонным доменами происходит с помощью микроэлектронного КМОП-чипа. Технологией для объединения чипов является так называемый flipchip bonding, позволяющий размещать все необходимые элементы на единой платформе. Актуальным продуктом является аппаратный ускоритель PACE (Photonic Arithmetic Computing Engine) [28], работающий с системной частотой синхронизации до 1 ГГц (рис. 5). Как и в случае с Envise, основной операцией является вектор-матричное произведение, однако среди ключевых задач компания выделяет Max-Cut, Min-Cut и задачу Изинга, в которой

Рис. 5. Оптоэлектронный аппаратный ускоритель PACE от Lightelligence [28].

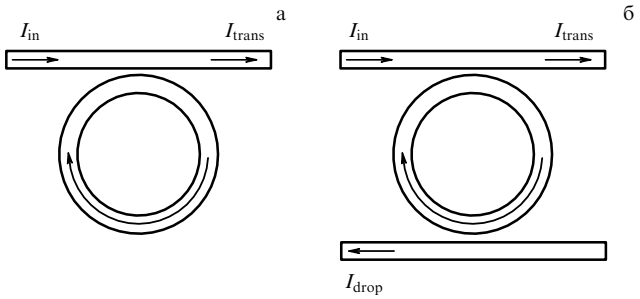

Рис. 6. Две базовые конфигурации микрокольцевых резонаторов: (а) односвязное микрокольцо, (б) двухсвязное микрокольцо [30].

авторы обещают ускорение до трёх порядков величины по сравнению со стандартными решениями на основе графических процессоров. Одним из возможных путей для масштабирования технологии, предложенным в Lightelligence, является использование многомодовых интерферометров вместо стандартных ИМЦ, в том числе спроектированных с помощью метода обратной оптимизации их формы [29].

2.1.3. Интегральная оптика: микрорезонаторы. В противовес оптическим вычислительным схемам с логическими элементами на интерферометрах Маха – Цендера, те же схемы на резонаторах позволяют не только управлять сигналом при помощи мощности и/или фазы исходного (или управляющего) сигнала, но и обеспечивать в том или ином виде контроль и селективность за счёт длины волны. Зачастую такие схемы получаются более компактными, чем интерферометрические, потому что позволяют использовать один волновод для разных каналов, закодированных длиной волны излучения.

Рисунок 6 иллюстрирует типичные схемы расположения включения кольцевого резонатора в оптическую схему [30]. Прямые волноводы принято называть шинами, количество шин определяет, в каком режиме может работать устройство. Одна шина превращает микрорезонатор в частотный фильтр, который не пропускает сигнал в окрестности длин волн, соответствующих собственным значениям резонатора. Вторая шина, в свою очередь, позволяет либо выводить сигнал из резонатора и передавать его дальше, либо заводить в кольцо дополнительный сигнал, что влияет на значение фазы и мощности на выходе, т.е. сделать своего рода оптический транзистор.

Для того чтобы такие схемы работали эффективно и на больших масштабах, резонаторы должны обладать высокими (порядка $10^7 – 10^8$) значениями добротности, а шины — обеспечивать малые потери излучения при распространении. Добротность резонатора, как и пропускная способность шин, зависят от их геометрических параметров (высота, ширина, радиус) и используемых материалов. Выбор материалов широк: от кремния и его соединений до полимерных веществ. Эффективность заведения и вывода сигнала из шины в резонатор и из резонатора в шину определяется константой связывания, значением которой можно управлять, изменяя расстояние между кольцом и волноводом. Все описанные выше параметры также зависят от длины волны излучения, посредством которого осуществляется передача сигнала. Таким образом, при проектировании резонаторных схем

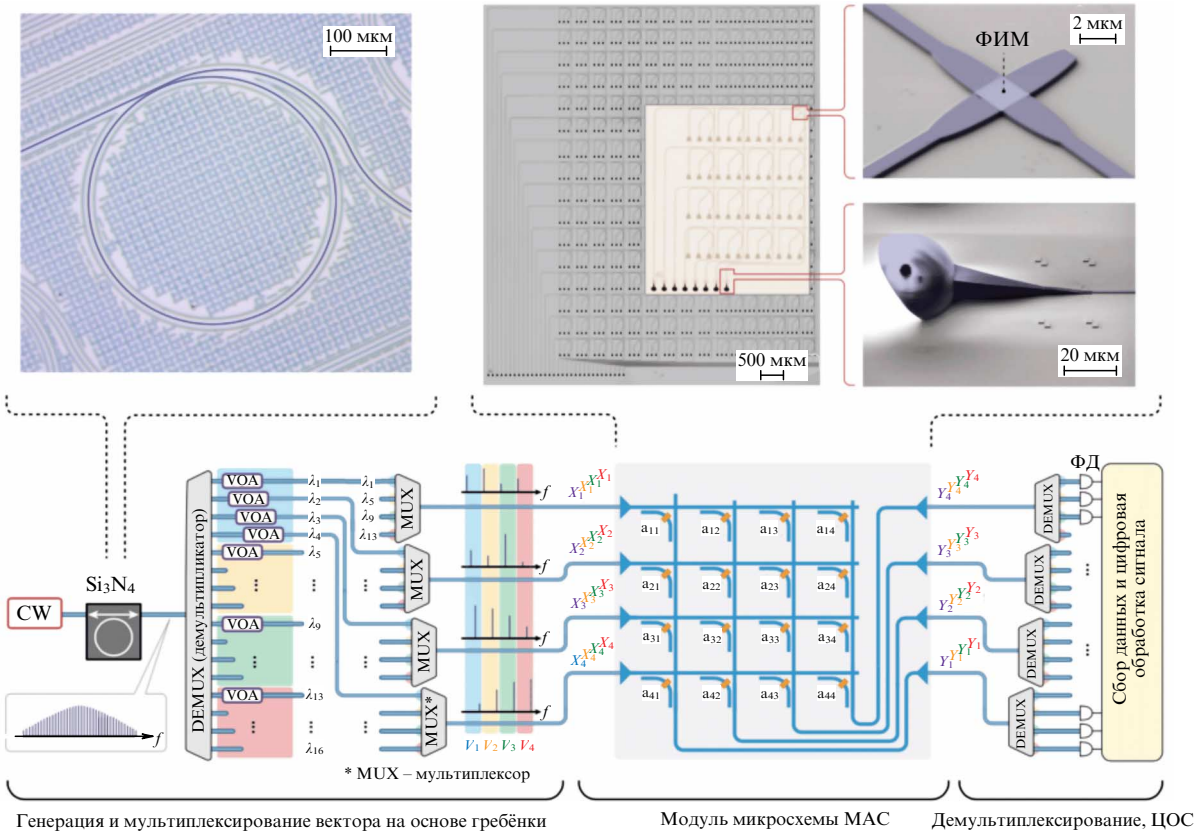

Рис. 7. Схематическое изображение фотонного процессора на основе кроссбар-массива. Оранжевые области $a_{11} - a_{44}$ соответствуют элементам оптической памяти на основе PCM-материала (реальное РЭМ-изображение представлено на вставке справа сверху), кодирующим пропускание в каждом отдельном канале и соответствующим весам матрицы ядра, с которым необходимо провести свёртку. С помощью WDM мультиплексирования возможна параллельная обработка сразу нескольких входных векторов ($V_1 - V_4$, показаны разным цветом). Для этого в качестве источника излучения используется микрогребёнка, полученная с помощью кольцевого микрорезонатора из нитрида кремния (реализовано на отдельном чипе) [31].

требуется решить задачу по оптимизации параметров и выбору длины волны сигнального излучения.

2.1.4. Фотонная кроссбар-матрица на основе материалов с фазовым переходом для модовых переключателей. Третьим вариантом микроархитектуры является фотонная матрица по типу кроссбар-массива. Впервые такой вариант фотонного ускорителя был представлен в работе [31] в группе Хариша Баскарана. Ключевым элементом структуры является энергонезависимая фотонная память на основе материала с фазовым переходом (от англ. phase-change material — PCM) [32]. С помощью такой памяти можно фиксировать элементы фотонной матрицы (т.е. коэффициенты пропускания и отражения), что позволяет существенно снизить энергопотребление в случае многократного её использования без необходимости обновления, например в задачах свёртки с известным ядром (в свёрточных нейронных сетях). Схематически структура фотонного процессора представлена на рис. 7. Подобная архитектура используется британской компанией Salience labs, которая была основана выпускниками упомянутой выше научной группы. Похожая кроссбар-архитектура упоминалась также в патенте [22].

В качестве источника излучения в данной работе выступала оптическая гребёнка, полученная на интегральном чипе на основе нитрида кремния при накачке микрокольцевого резонатора. Аналогично можно было использовать несколько лазеров на разных длинах волн и мультиплексировать их излучение перед заведением на оптический чип. Каждый канал модулируется независимо по амплитуде (в согласии со значениями, передаваемыми из внешней памяти через ЦАП), после чего в каждый входной волновод фотонной матрицы заводится излучение сразу на нескольких длинах волн. Затем входной вектор, закодированный с помощью амплитуды световой волны в каждом из горизонтально ориентированных каналов фотонной матрицы ($X_1 - X_4$ на рис. 7), разбивается в равных пропорциях между вертикальными каналами. Для этого коэффициент деления DC-делителей увеличивается от левого края к правому, обеспечивая необходимую пропорцию. Стоит отметить, что DC-делители являются наиболее проблемным элементом фотонной матрицы, потому что, во-первых, они обладают существенной дисперсией и неравномерно делят излучение на разных длинах волн, а во-вторых, обладают большой чувствительностью к разбросу геометрических параметров при изготовлении. Последнее нужно учитывать для корректной работы ускорителя, в частности, использовать специальную нормализацию элементов фотонной матрицы уже созданного устройства. Новые эффективные делители, разработанные с помощью подходов машинного обучения и генетической оптимизации, также могли бы помочь в решении подобных проблем.

После деления входных векторов происходит их взвешивание с помощью элементов PCM. В разных фазовых состояниях данные материалы обладают разным

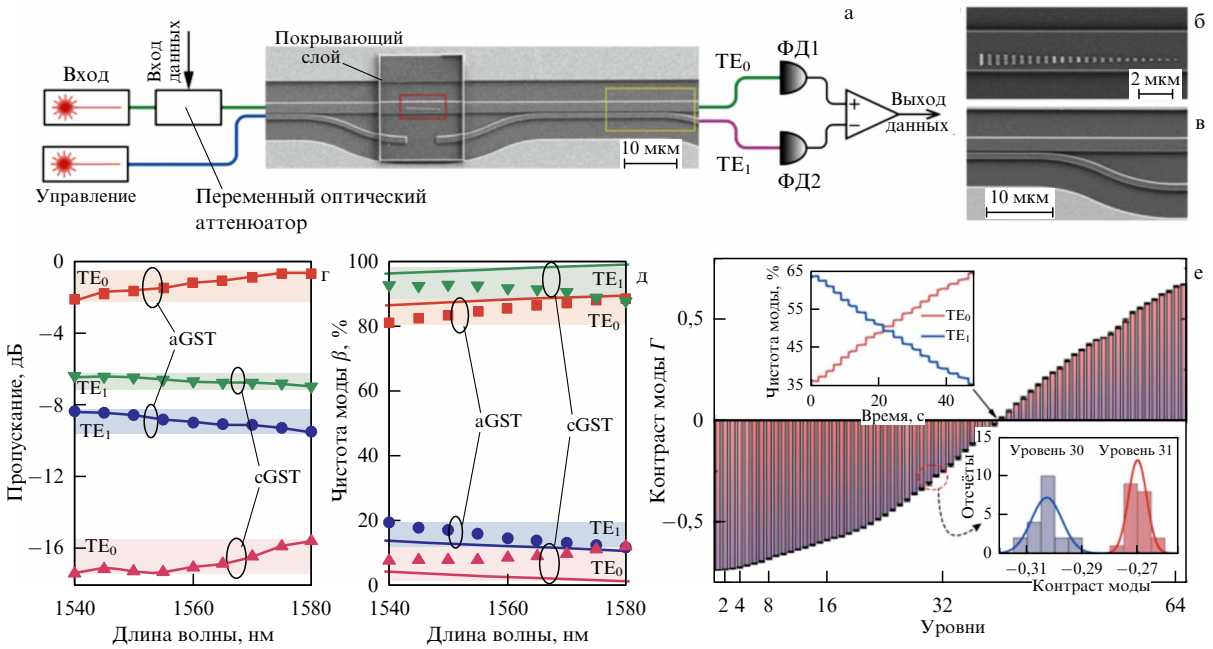

Рис. 8. Интегральная метаповерхность на основе PCM-материала для модовой конверсии на чипе. При полностью кристаллическом состоянии PCM проходящее излучение эффективно преобразуется из TE_0- в TE_1-моду, при полностью аморфном почти никаких изменений не происходит. Справа сверху вставкой показано РЭМ-изображение метаповерхности на многомодовом волноводе из нитрида кремния [33].

поглощением: как правило, в аморфном оно ниже, чем в кристаллическом. Кроме того, за счёт оптической перестройки, т.е. локального нагрева при распространении наносекундных импульсов через элемент, можно добиться устойчивых промежуточных состояний с частичной кристаллизацией или аморфизацией, что позволяет регулировать пропускание в каждом волноводе с определённой точностью через фиксированные уровни (в данной работе — с дискретизацией до 5 бит). Таким образом, каждый элемент входного вектора испытывает определённое ослабление в согласии с весами предполагаемой матрицы-фильтра. После этого излучение из разных горизонтальных каналов смешивается в вертикальных волноводах и поступает на детектор, предварительно пройдя процедуру демультиплексирования при параллельной работе на нескольких длинах волн. Сигнал с детекторов после трансимпедансного усилителя поступает на устройство АЦП, и далее результат передаётся во внешнюю память устройства. Как и в случае предыдущих архитектур, точность и скорость работы АЦП–ЦАП является ограничивающим фактором производительности фотонного ускорителя. По оценкам авторов работы [31], подобное устройство способно производить до 10^{15} MAC-операций в секунду при энергопотреблении около 20 фДж на операцию (учитывая только оптические потери). Однако необходимо учитывать перечисленные выше проблемы и дополнительную внешнюю обвязку, которые способны существенно снизить ожидаемую производительность и увеличить энергопотребление всей системы.

Продолжением представленной архитектуры может служить подход с использованием не только разных длин волн при WDM-мультиплексировании, но и разных модовых состояний в интегральных оптических волноводах. Данная идея была предложена в работе [33]. Ключевым отличием от предыдущего случая является использование наноструктурированной метаповерхности из материала с фазовым переходом вместо сплошного слоя на волноводе (рис. 8), что позволяет не только задействовать дополнительную степень свободы, но и достичь большего числа устойчивых уровней при перестройке (дискретизация 6 бит против 5 бит в предыдущей работе). Размер отдельных элементов метаповерхности подбирается таким образом, чтобы в кристаллическом состоянии материала согласовывать эффективные показатели преломления для TE_0 - и TE_1-мод многомодового волновода из нитрида кремния. При изменении фазового состояния материала согласование ухудшается, что приводит к сохранению модового состава в проходящей световой волне. Таким образом, изменяя через промежуточные уровни состояние метаповерхности, можно варьировать модовый контраст излучения. Это используется для задания элементов фотонной матрицы вместо обычного пропускания, как было реализовано в предыдущей работе.

Объединение таких метаповерхностей в кроссбар-массив позволяет рассматривать его как аналогичный фотонный процессор для вектор-матричных вычислений (рис. 9). Здесь также используется подход WDM-мультиплексирования, а сама схема работает по похожим принципам. По оценкам авторов статьи [31], подобное устройство способно совершать до 164 тераопераций в секунду, однако стоит учитывать, что все обозначенные выше сложности для фотонного кроссбар-массива сохраняют свою актуальность и в этом случае. Кроме того, метаповерхность также будет обладать чувствительностью к длине волны излучения, что скажется на возможности масштабирования системы по спектру (сколько реально каналов WDM можно будет использовать параллельно без деградации точности вычислений).

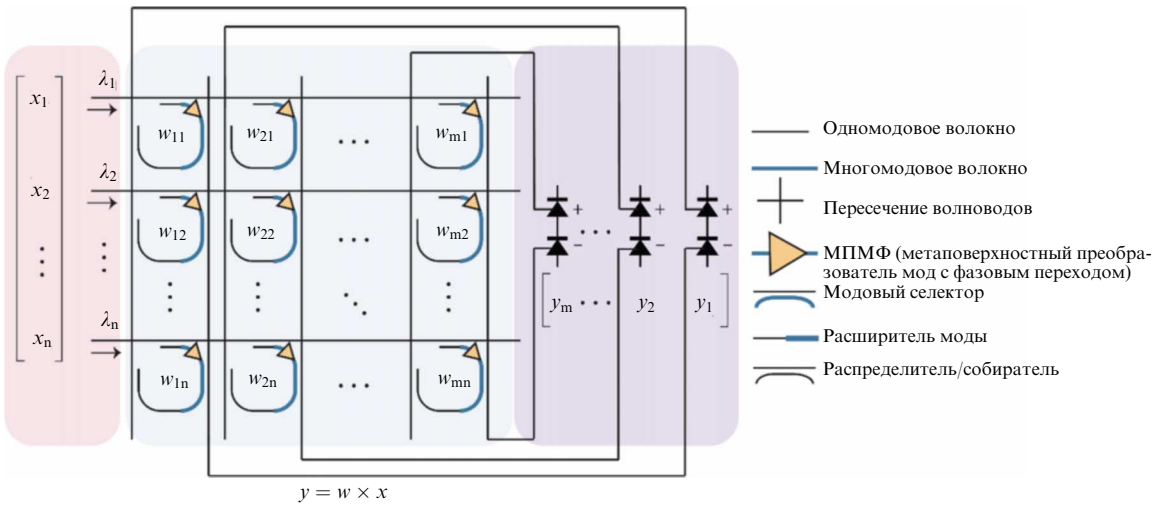

Рис. 9. Фотонная матрица по типу кроссбар-массива с использованием интегральных PCM-метаповерхностей (отмечены как МПМФ). В отличие от предыдущей работы, необходимо использовать балансные фотодетекторы, что немного усложняет схему [33].

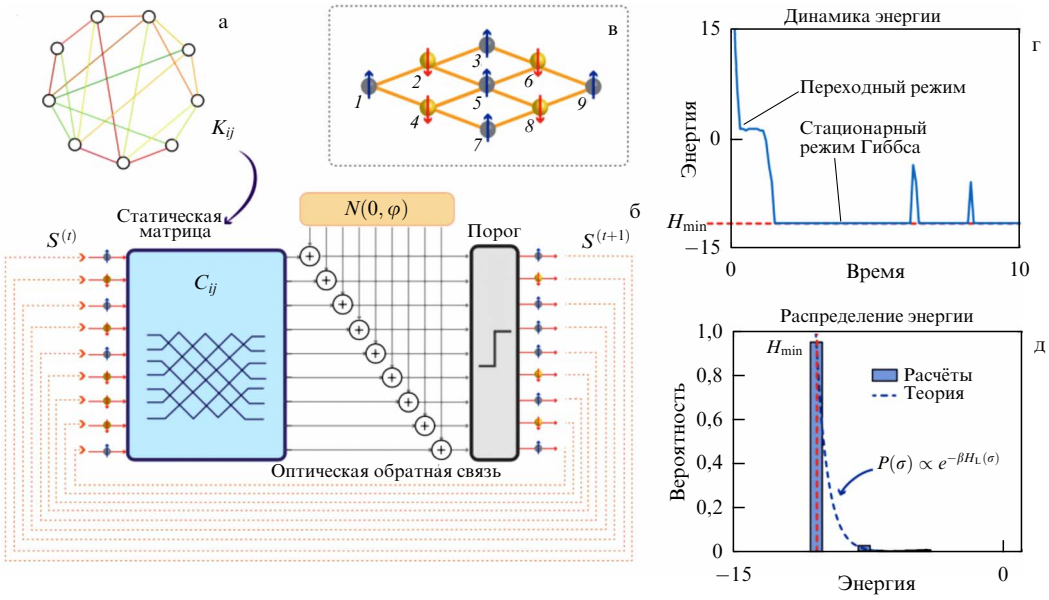

Рис. 10. Фотонный ускоритель для задачи Изинга [34].

2.2. Фотонная машина Изинга на интегральной платформе

Примеры фотонных ускорителей для высокопроизводительных вычислений не ограничиваются MAC-операциями. Ещё одним актуальным приложением может быть решение комбинаторных задач, например определение максимального разреза графа (MAX-CUT). В частности, в работе [34] был продемонстрирован интегральный фотонный ускоритель для задачи Изинга на основе уже рассмотренной выше архитектуры сетки ИМЦ. Особый интерес к указанной проблеме возникает в связи с возможностью свести многие распространённые комбинаторные проблемы к задаче Изинга [35]. Принцип работы фотонного ускорителя для задачи Изинга схематически представлен на рис. 10. На каждом алгоритмическом шаге на вход фотонного ускорителя поступает вектор спинового состояния, кодируемый с помощью амплитудных модуляторов в каждом волноводном канале. Излучение, соответствующее данному вектору, проходит через фотонную матрицу, состоящую из ИМЦ, реализующих матрицу C_{ij}, связанную с исходной матрицей спиновой связи K_{ij}. На выходе матрицы происходит зашумление результата гауссовым шумом со стандартным отклонением φ, которое может быть реализовано как в оптическом, так и в электронном виде. После этого с помощью пороговой функции происходит формирование нового вектора состояния, вновь поступающего на вход фотонной матрицы. Независимо от начального вектора через множество итераций результат сходится к распределению Гиббса для конкретной матрицы спиновой связи. По оценкам авторов статьи [34], при работе на частоте порядка 1 ГГц энергия на операцию масштабируется как $9/N$ пДж, где N — число спинов в задаче, в то время как стандартный графический процессор даёт значение порядка 2,2 пДж на операцию. Однако стоит отметить, что, как и в случае фотонной матрицы для перемножения матриц, предельное значение N ограничивается оптическими потерями,

резко возрастающих при увеличении размерности матрицы. На данный момент есть упоминания об экспериментальной реализации 64 входных каналов, но дальнейшее масштабирование может быть затруднено в рамках представленной архитектуры и существующих элементов на фотонном чипе.

3. Оптические вычисления в свободном пространстве

3.1. Дифракционные нейронные сети

Распространение электромагнитного излучения в пространстве описывается волновым уравнением [36]. При этом в ходе распространения излучение претерпевает пространственные изменения — происходит дифракция. Если известно распределение электромагнитного поля $U(x,y)$ в плоскости P, то распределение поля $U'(x',y')$ в плоскости P', которая расположена на расстоянии z от плоскости P, может быть найдено согласно принципу Гюйгенса – Френеля [37]:

$$U'(x',y') = \frac{z}{\mathrm{i}\lambda}\iint_P U(x,y)\,\frac{\exp(\mathrm{i}kr)}{r^2}\,\mathrm{d}x\,\mathrm{d}y\,, \qquad (1)$$

где $r = [z^2 + (x-x')^2 + (y-y')^2]^{1/2}$. Формула (1) может быть переписана в виде

$$U'(x',y') = \iint_P U(x,y)\,f(x-x',y-y')\,\mathrm{d}x\,\mathrm{d}y\,,$$

где $f(x-x',y-y') = z\exp(\mathrm{i}kr)/(\mathrm{i}\lambda r^2)$. Последнее выражение эквивалентно формуле свёртки двух функций. Таким образом, распространение распределения поля $U(x,y)$ в свободном пространстве на расстояние z равносильно применению операции свёртки с фиксированным ядром.

Это уникальное свойство электромагнитного излучения позволило разработать полностью оптическую нейронную сеть — дифракционную нейронную сеть (ДНС) [38]. Основная идея такой нейронной сети показана на рис. 11. Сеть состоит из амплитудной маски на входе (входной уровень на рис. 11а) и нескольких фазовых (амплитудно-фазовых) масок. Когерентное излучение освещает амплитудную маску, которая задаёт входное распределение поля. В данной статье в качестве изображений использовались рукописные цифры из набора данных MNIST (от англ. Modified National Institute of Standards and Technology dataset) [39]. Далее, согласно принципу Гюйгенса – Френеля, каждая точка амплитудной маски является точечным источником вторичных волн. Их интерференция формирует распределение поля на первой фазовой маске, расположенной на некотором расстоянии. Фазовые маски состояли из отдельных пикселей, вносивших некоторую фазовую задержку в каждой точке пространства. Затем каждый пиксель фазовой маски снова выступает как точечный источник вторичных волн, и так продолжается на протяжении всей ДНС. Фазовые маски позволяют управлять условием интерференции вторичных волн, что в итоге позволяет получить на выходе ДНС желаемое распределение поля. В частности, авторы решали задачу классификации рукописных цифр из набора данных MNIST (рис. 11б). Для этого на выходе ДНС помещались 10 детекторов, соответствующих цифрам от 0 до 9. В качестве предсказания ДНС выступал номер детектора, на котором наблюдается максимум интенсивности.

Для того чтобы понять, какие фазовые маски нужно взять для успешного решения задачи классификации, авторы реализовали и обучили ДНС на компьютере. После этого авторы экспериментально реализовали ДНС из 5 фазовых масок для электромагнитного излучения с частотой 0,4 ТГц. Каждая фазовая маска состояла из 200×200 нейронов размером 400 мкм каждый, а расстояние между масками составило 3 см. Маски изготавливались на 3D-принтере, а необходимая задержка фаз вносилась толщиной материала. В итоге экспериментально реализованная ДНС показывает точность 88 % при расчётной точности 91,7 %.

От стандартных глубоких нейронных сетей ДНС отличается тем, что это физическая и полностью оптическая сеть. Кроме того, она имеет некоторые архитектурные особенности. Во-первых, входные сигналы для нейронов являются комплекснозначными, определяемыми волновой интерференцией и комплексным коэффициен-

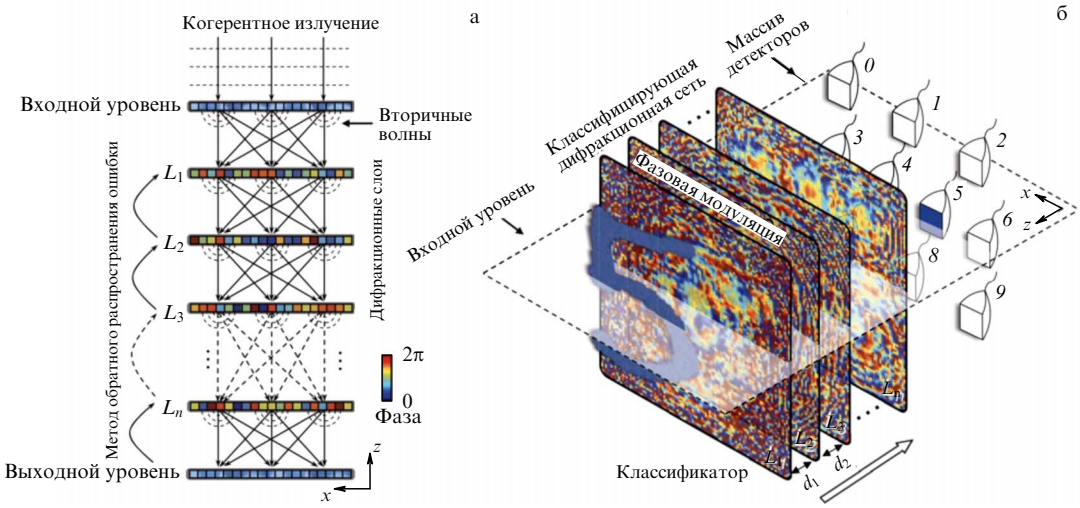

Рис. 11. (а) Схематичное изображение принципа работы ДНС. (б) Схема работы ДНС для задачи классификации цифр из набора данных MNIST [38].

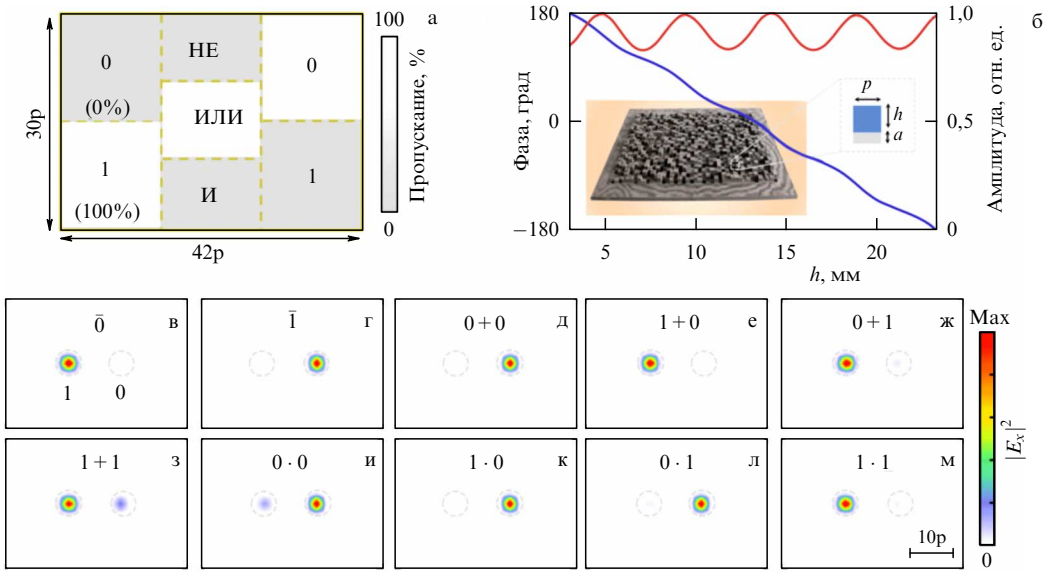

Рис. 12. (а) Амплитудная маска на входе ДНС для кодирования логической операции. (б) Зависимость фазы и амплитуды пропускания одного пикселя метаповерхности от высоты пикселя. (в – м) Примеры распределения поля на выходе ДНС при выполнении различных логических операций [40].

том пропускания/отражения всех масок. Во-вторых, индивидуальной функцией нейрона является фазовая и амплитудная модуляция его входного сигнала для генерации вторичной волны, в отличие от других нелинейных функций нейрона (сигмоида, ReLU, гиперболический тангенс), используемых в современных глубоких нейронных сетях. В-третьих, выходные сигналы каждого нейрона соединяются с нейронами следующего слоя (фазовой маски) посредством распространения волн и когерентной (или частично когерентной) интерференции, обеспечивая уникальную форму взаимосвязи внутри сети. Например, хорошо известно, что в современных свёрточных нейронных сетях поле восприятия (receptive field) регулируется размерностью ядра свёртки и числом последовательных операций свёртки. В ДНС же поле восприятия зависит от осевого расстояния между различными фазовыми масками и свойствами пространственной и временно́й когерентности источника освещения. Вторичная волна каждого нейрона теоретически будет рассеиваться под всеми углами, воздействуя в принципе на все нейроны следующего слоя. Однако при заданном расстоянии между последовательными слоями ДНС интенсивность волны от нейрона будет затухать ниже минимального уровня шума обнаружения после определённого расстояния, которое фактически задаёт поле восприятия ДНС и может быть физически отрегулировано путём его изменения между слоями сети, интенсивностью входного оптического излучения или длиной когерентности и диаметра источника освещения.

Помимо задач классификации, ДНС успешно решают задачи оптических вычислений, например для реализации управляемого оптического гейта (логической операции — от англ. gate) [40]. Для этого авторы использовали ДНС, состоящую из двух амплитудно-фазовых масок. На входе ДНС располагалась амплитудная маска (рис. 12а), задающая необходимую логическую операцию и операндов. Для демонстрации были выбраны операции НЕ, ИЛИ, И и частота электромагнитного излучения 17 ГГц. В качестве амплитудно-фазовых масок выступали метаповерхности — структурированные поверхности, параметры отражения от которых определяются коллективными эффектами специально созданной структуры. Для них пропускание каждого пикселя определялось толщиной материала (рис. 12б). Далее в зависимости от распределения поля, прошедшего через входную амплитудную маску, излучение фокусировалось в одной из двух областей выходного экрана, соответствующих логическим 0 и 1. Таким образом, выполнение логических операций по своей сути было сведено с бинарной классификации входного распределения поля. Экспериментально реализованная ДНС успешно справлялась с вычислением результата всех логических операций, а контраст между сигналами из областей, соответствующих правильному и неправильному результату операции, не опускался ниже 9,6 дБ. Предложенная авторами идея может быть расширена на выполнение всех основных бинарных логических операций, а диапазон работы может быть перенесён в видимый или инфракрасный. Преимуществом такого подхода является то, что многофункциональный логический гейт реализован не как набор отдельных гейтов, а как один гейт, которым можно легко управлять с помощью входного распределения поля.

ДНС способны обрабатывать оптический сигнал с широким спектром, т.е. параллельно для многих длин волн [41]. Авторы показали, что, изменяя функцию потерь нейронной сети, можно добиться желаемой функциональности, например фильтрации по спектру или (де)мультиплексирования. В качестве примера был взят импульс с начальной частотой 0,25 ТГц и конечной 1 ТГц, проходящий через трёхслойную ДНС. Авторы показали, что можно реализовать узкополосный фильтр с заранее заданной центральной частотой и добротностью, определяемой функцией потерь ДНС. Для достижения указанной цели ДНС была обучена фокусировать излучение с заданной частотой в выходную апертуру размером 2 мм. Эта идея может быть расширена на несколько частот, которые могут фокусироваться либо в одну

выходную апертуру (фильтр с несколькими полосами пропускания), либо в разные выходные апертуры в зависимости от длины волны (аналог спектрального демультиплексора). Кроме того, изменяя расстояния между слоями ДНС, можно добиться смещения центральной частоты фильтрации для уже изготовленных слоёв ДНС. Такая реализация ДНС может быть востребована в задачах оптической обработки информации.

Всесторонний анализ качества работы ДНС в задачах классификации изображений представлен в работе [42]. Во-первых, авторы показывают влияние нескольких параметров на итоговую точность ДНС. В частности, продемонстрировано, что функция потерь, используемая в процессе обучения ДНС, оказывает решающее влияние на итоговую точность классификации. Например, замена функции ошибок со средней квадратичной ошибки на кросс-энтропию позволяет повысить точность работы пятислойной ДНС на данных MNIST с 91 % до 97 %, а на данных Fashion MNIST [43] с 81 % до 89 %. Однако такое увеличение точности сопровождается падением интенсивности излучения на выходе из ДНС, так как при использовании кросс-энтропии сеть учится преобразовывать входящее изображение таким образом, чтобы на желаемый детектор (который соответствует заданному классу, например цифре 1) попадало больше излучения, чем на остальные детекторы. При этом не накладывается никаких условий на излучение, не попадающее на детекторы. Авторы также исследовали зависимость точности ДНС от расстояния между слоями сети и установили, что при уменьшении расстояния с 40 длин волн до 4 точность на наборе данных MNIST падает на 3 %. Во-вторых, авторы показали, что ДНС можно соединить с простой "цифровой" нейронной сетью и повысить точность классификации наборов данных MNIST и Fashion MNIST до 99 % и 90 %, что сравнимо с точностью современных полностью цифровых нейронных сетей. Стоит отметить, что энергопотребление современных цифровых свёрточных нейронных сетей составляет порядка $10^{-3}-10^{-4}$ Дж/изображение (для сети ResNet [44, 45]), в то время как гибридная ДНС потребляет порядка 10^{-9} Дж/изображение.

Зависимость качества работы ДНС от числа слоёв теоретически исследована в работе [46]. В ней исследуется вопрос сложности ДНС и её способности производить произвольные преобразования падающего излучения. Показано, что матрица, задающая оптическое преобразование, имеет ранг, определяемый произведением числа пикселей на входе и на выходе ДНС. При этом ДНС, состоящая из конечного числа слоёв, в общем случае соответствует матрице с меньшим рангом. При увеличении числа слоёв ранг матрицы ДНС растёт линейно с числом слоёв, пока не достигнет предельно допустимого значения. Таким образом, для реализации ДНС с наибольшей способностью обучения, т.е. с наибольшей сложностью, требуется увеличивать как число слоёв, так и число пикселей (нейронов) на входе и на выходе ДНС.

Для реализации ДНС в видимом диапазоне в работе [47] предлагается использовать метаповерхности — диэлектрические структуры, состоящие из элементов, имеющих субволновые размеры. В данной работе рассматривается метаповерхность, состоящая из параллелепипедов из материала TiO_2. В силу того что параллелепипед несимметричен относительно ортогональных осей, проходящих вдоль его сторон, отклик по поляризации оказывается анизотропным. Кроме того, изменяя размеры параллелепипеда, можно менять амплитуду и фазу излучения, которое рассеивается на нём. Два вышеуказанных фактора позволяют реализовать ДНС, способную выполнять различные задачи в зависимости от поляризации падающего излучения, т.е. можно реализовать ДНС с мультиплексированием излучения по поляризации. В частности, авторы показали ДНС, способную одновременно классифицировать изображения из данных MNIST и Fashion MNIST. Для этого изображения каждого из наборов данных попадают на вход ДНС со своей поляризацией, а далее в зависимости от класса изображения излучение фокусируется в область, соответствующую заданному классу, например, при подаче на вход цифры 1 излучение фокусируется в левом верхнем углу камеры, а при подаче цифры 3 — в правом нижнем. В зависимости от набора данных и от класса излучение должно фокусироваться в разные места матрицы. Ещё одним нововведением данной статьи является интеграция ДНС с детектирующей системой. Для этого метаповерхность изготавливается прямо над поверхностью камеры на расстоянии 100 мкм, что позволяет легко масштабировать производство таких ДНС.

3.2. Оптическое преобразование Фурье

Известно [37], что тонкая линза в параксиальном приближении вносит задержку фаз, определяемую уравнением $t=\exp\left[-\mathrm{i}(k/2f)(x^2+y^2)\right]$, где k — волновой вектор излучения, f — фокус линзы, x, y — координаты относительно оси линзы. Пусть на линзу падает некоторое распределение поля $U(x,y)$, тогда после линзы наблюдается распределение поля $U(x,y)*t$, а распределение поля в фокальной плоскости линзы может быть найдено по формуле (1)

$$U'(x',y')=\frac{z}{\mathrm{i}\lambda}\iint_P U(x,y)\exp\left[-\mathrm{i}\,\frac{k(x^2+y^2)}{2f}\right]\times \\ \times\frac{\exp(\mathrm{i}kr)}{r^2}\,\mathrm{d}x\,\mathrm{d}y\,.$$

При использовании линзы в параксиальном приближении выполняется условие $z\gg x, y$, и r можно разложить в ряд Тейлора:

$$r=\sqrt{z^2+(x-x')^2+(y-y')^2}\approx \\ \approx z\left(1+\frac{(x-x')^2}{2z^2}+\frac{(y-y')^2}{2z^2}\right).$$

Тогда распределение поля в фокальной плоскости линзы $(z=f)$ может быть найдено по формуле

$$U'(x',y')=\frac{\exp(\mathrm{i}kf)}{\mathrm{i}\lambda f}\iint_P U(x,y)\exp\left[-\frac{\mathrm{i}k(x^2+y^2)}{2f}\right]\times \\ \times\exp\left[\frac{\mathrm{i}k\left((x-x')^2+(y-y')^2\right)}{2f}\right]\mathrm{d}x\,\mathrm{d}y= \\ =\frac{\exp(\mathrm{i}kf)\exp\left[\mathrm{i}k(x'^2+y'^2)/(2f)\right]}{\mathrm{i}\lambda f}\times \\ \times\iint_P U(x,y)\exp\left[\frac{\mathrm{i}k(xx'+yy')}{f}\right]\mathrm{d}x\,\mathrm{d}y\,.$$

Таким образом, $U'(x',y')$ пропорционально фурье-преобразованию падающего на линзу распределения поля.

На основе линз можно построить фурье-дифракционную нейронную сеть (ФДНС) [48]. Принцип работы такой сети следующий. Входное распределение поля попадает на линзу, формирующую фурье-образ этого распределения поля в своей фокальной плоскости. Далее в данном месте размещается оптический элемент, модулирующий излучение (фазовая маска, пространственный модулятор света (англ. spatial light modulator — SLM) или массив микрозеркал (digital micromirror device — DMD)). Такая модуляция равносильна перемножению фурье-образа входного сигнала на некоторую комплексную матрицу. Модулирующий элемент, в свою очередь, находится в фокусе второй линзы, которая производит обратное преобразование Фурье уже для модулированного излучения. Как известно, фурье-образ произведения двух величин (в данном случае фурье-образ входного излучения и матрицы модулирующего элемента) соответствует операции свёртки. Таким образом, принцип работы ФДНС похож на принцип работы свёрточных нейронных сетей. В качестве модулирующего элемента в работе [48] был использован массив микрозеркал с разрешением 1920×1080 пикселей, обновляющихся с частотой 20 кГц. Для достижения высокой скорости работы ФДНС на вход поступало не единичное изображение из набора данных, а 16 изображений, объединённых в матрицу 4×4, причём каждое изображение имело размер 208×208 пикселей, т.е. итоговое составное изображение имело размер 832×832 пикселей. Благодаря такому объединению все 16 изображений могут подвергаться операции свёртки одновременно. Скорость операции ограничена только частотой обновления пикселей массива микрозеркал (20 кГц) и частотой считывания сигнала камерой (1 кГц). Авторы показали работоспособность концепции ФДНС на примере однослойной сети, для которой происходило обучение 16 ядер свёртки, т.е. для ФДНС задавались 16 конфигураций массива микрозеркал, и эти конфигурации применялись последовательно. После ФДНС следовала однослойная полносвязная цифровая нейронная сеть с нелинейной функцией активации. После обучения ФДНС продемонстрировала точность классификации, равную 98 % для MNIST и 63 % для CIFAR-10 [49]. По результатам исследований авторы показали, что их модель ФДНС способна вычислять результат операции свёртки для больших матриц в 10 раз быстрее, чем современные графические ускорители. Скорость вычислений может быть ещё выше при использовании передовых разработок в области массивов микрозеркал.

Продолжением данной идеи является реализация операций свёртки оптическими методами, а всех остальных операций нейронной сети цифровыми методами [50]. Авторы предлагают использовать массив микролинз так, чтобы в фокусе каждой микролинзы располагалась своя амплитудно-фазовая маска. Такой подход позволяет одновременно вычислить результат операции свёртки для нескольких ядер. В качестве примера авторы попытались заменить первый свёрточный слой хорошо известной свёрточной нейронной сети AlexNet [51] на слой, реализованный оптически. Остальные слои нейронной сети при этом были реализованы на компьютере. Было установлено, что при замене цифрового свёрточного слоя на оптический точность классификации для набора данных Kaggle's Cats and Dogs [52] падает с 96 % до 87 %. Скорее всего, такое падение вызвано отсутствием нелинейной функции активации для оптического свёрточного слоя и маленьким размером тренировочного набора данных. На более известном наборе данных MNIST точность нейронной сети с оптическим свёрточным слоем отличается от точности полностью цифровой нейронной сети менее чем на 0,5 %. Основной мотивацией замены цифрового свёрточного слоя на оптический является ускорение времени работы нейронной сети. Время T_{latency} выполнения операции свёртки с помощью оптической схемы определяется суммой времён: $T_{\text{latency}} = T_{\text{source}} + T_{\text{prop}} + T_{\text{detect}} + T_{\text{data}}$, где T_{source} — время, необходимое для генерации входного изображения, T_{prop} — время, за которое свет проходит через оптическую схему, T_{detect} — время, необходимое для детектирования сигнала, T_{data} — время, необходимое для передачи детектированного сигнала в программное обеспечение, реализующее оставшуюся часть нейронной сети. Время T_{source} определяется быстродействием системы генерации изображений и в данной работе составляло 1 мс (частота обновления 1 кГц). Время T_{prop} составляет несколько пикосекунд, и его можно не учитывать. Время T_{detect} определяется быстродействием детектирующего элемента и в случае CCD-камеры составляет порядка 1 мс. Время T_{data} определяется временем передачи сигнала от детектирующего элемента к компьютеру. Для соединения USB 3.0 с пропускной способностью 2500 Мбит $с^{-1}$ и изображения весом 100 КБ T_{data} будет равно 0,32 мс. Таким образом, итоговое время вычисления результата операции свёртки будет занимать 2,32 мс. Важно отметить, что это время не чувствительно к изменению разрешения изображения. В итоге может быть получен график (рис. 13) сравнения скорости работы свёрточных слоёв, реализованных полностью в цифровом виде и с использованием оптики. Вычисление результата работы одного свёрточного слоя, реализованного с помощью оптической схемы, происходит быстрее, чем вычисление с помощью графического ускорителя, при размере изображения более 500×500 пикселей.

Авторы также оценили энергопотребление свёрточного слоя, реализованного оптически, и свёрточного слоя, размещённого на графическом ускорителе. При оптической реализации энергопотребление не зависит

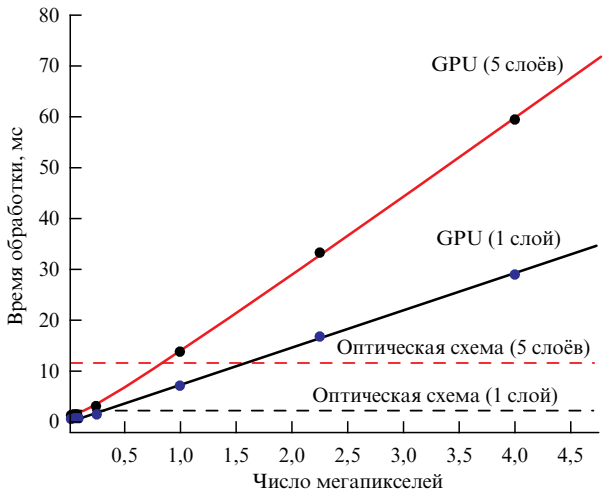

Рис. 13. Зависимость скорости работы свёрточных слоёв, размещённых на графическом ускорителе (GPU) и реализованных с помощью оптической схемы [50].

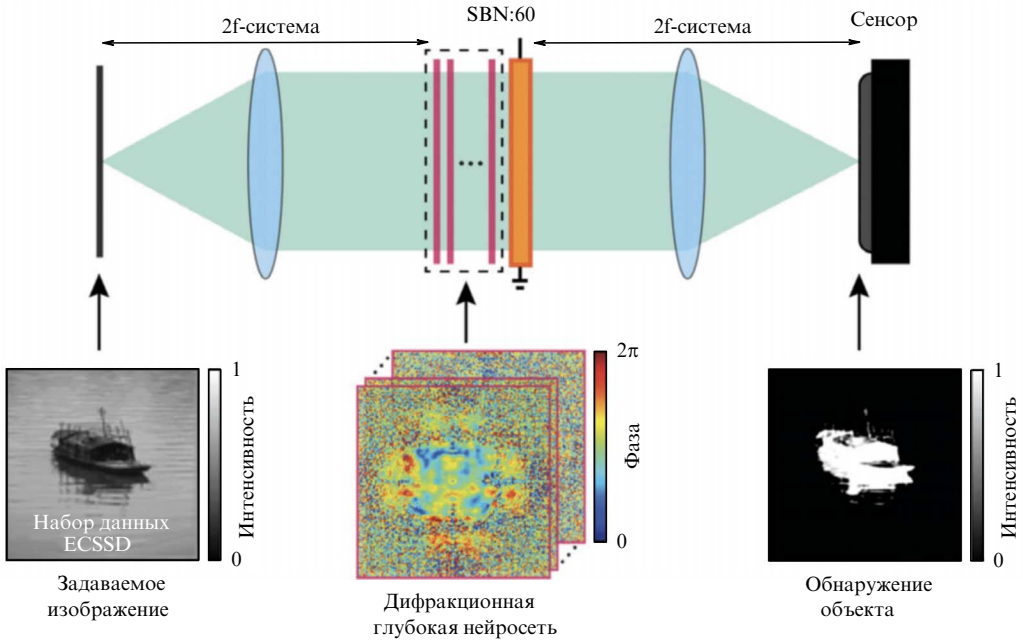

Рис. 14. Схема ФДНС с пластинкой фоторефрактивного материала SBN:60 [53].

от размерности ядра свёртки, в то время как энергопотребление свёрточного слоя на графическом ускорителе растёт пропорционально числу пикселей в ядре свёртки. Можно сделать вывод, что оптическая реализация даже одного свёрточного слоя востребована только при работе с изображениями с высоким разрешением и при применении ядер свёртки большой размерности.

Одним из главных препятствий в использовании ДНС является отсутствие лёгкого способа реализации оптической нелинейной функции активации. В работе [53] для этого предлагается использовать тонкую пластинку фоторефрактивного материала SBN:60, установленную в конце многослойной ФДНС (рис. 14). Основной мотивацией использования именно материала SBN:60 является его большой нелинейный отклик. Показатель преломления данного материала меняется в зависимости от интенсивности падающего излучения согласно закону

$$n = \kappa E_{\text{app}} \frac{\langle I \rangle}{1 + \langle I \rangle} ,$$

где n — изменение показателя преломления, κ — константа, определяемая показателем преломления материала, электро-оптическим коэффициентом и интенсивностью $\langle I_0 \rangle$ однородной засветки материала, E_{app} — величина приложенного статического электрического поля, $\langle I \rangle$ — возмущение интенсивности относительно равномерной засветки $\langle I_0 \rangle$. Было показано, что изменением n можно добиться изменения фазы проходящего излучения от 0 до π, если взять пластинку толщиной 1 мм и приложить к ней статическое напряжение 972 В при интенсивности падающего излучения 0,1 мВт мм$^{-2}$. Внедрение нелинейного слоя позволило авторам реализовать ФДНС, способную решать задачи сегментации и детектирования объектов на изображениях (рис. 14 внизу). Такая ФДНС может работать в режиме реального времени и детектировать объекты прямо на видео.

4. Заключение

В настоящем обзоре рассмотрены передовые подходы к реализации аналоговых фотонных вычислений и фотонных интеллектуальных систем.

Первая группа подходов основана на использовании элементов интегральной фотоники для реализации вектор-матричных операций, позволяющих внедрить фотонные нейронные сети в традиционные вычислительные системы в качестве сопроцессора для ускорения вычислений и повышения энергоэффективности. Данный подход уже находит практическое применение и имеет значительное преимущество в энергоэффективности, вплоть до 20 фДж на операцию, что на два порядка превышает существующую энергоэффективность стандартных центральных процессоров. Рассмотрены также проблемы данного подхода, связанные с эффективностью и предельной частотой оптоэлектронных преобразований, а также разрядностью чисел, с которыми происходят вычислительные операции.

Вторая группа подходов основана на оптических вычислениях в свободном пространстве, позволяющих производить математические операции за один вычислительный такт со всем массивом данных за счёт использования физических свойств оптического излучения. Данные подходы пока не нашли своего практического применения в вычислительной технике, в том числе по причине сложности их интеграции с текущими вычислительными алгоритмами, но, безусловно, заслуживают внимания и имеют значительный потенциал для работы с массивами данных большой размерности.

В качестве преимуществ фотонных вычислителей над электронными можно указать выигрыш в энергоэффективности на два порядка за счёт возможности параллельной работы сразу на нескольких длинах волн, используя физически один и тот же массив, при одинаковом количестве весовых элементов в фотонном и электронном кроссбар-массивах, повышенную пропускную способ-

ность благодаря модуляции сигнала с гораздо большей частотой, пониженными требованиями на отведение тепла от фотонного чипа. Преимуществами электронных компонентов над оптическими является компактность, т.е. возможность разместить больше элементов и компенсировать разницу в производительности. Однако вопрос дрейфа сопротивления остаётся открытым, что снижает точности расчётов аппаратного ускорителя.

Оптические вычисления для определённых задач открывают перспективы не только повышения энергоэффективности, но и скорости работы за счёт использования низкоразрядных высокоскоростных ЦАП – АЦП при низких энергозатратах, уменьшения числа электрооптических преобразований при работе с оптическими свёрточными нейронными сетями или если сигнал изначально представлен в оптическом домене.

Исследование выполнено в рамках научной программы Национального центра физики и математики (проект "Национальный центр исследования архитектур суперкомпьютеров") и при поддержке Некоммерческого фонда развития науки и образования "Интеллект".

Список литературы

1. Xu R et al. *Opt. Laser Technol.* **136** 106787 (2021)
2. Marković D et al. *Nat. Rev. Phys.* **2** 499 (2020)
3. Xu M et al. *Adv. Funct. Mater.* **30** 2003419 (2020)
4. Zhang J et al. *Adv. Intell. Syst.* **2** 1900136 (2020)
5. Sunny F P et al. *ACM J. Emerg. Technol. Comput. Syst.* **17** (4) 61 (2021)
6. Yao K, Unni R, Zheng Y *Nanophotonics* **8** 339 (2019)
7. Ferreira de Lima Th et al. *Nanophotonics* **6** 577 (2017)
8. Goodman J W, Dias A R, Woody L M *Opt. Lett.* **2** 1 (1978)
9. Zuo Y et al. *Optica* **6** 1132 (2019)
10. Liu J et al. *PhotoniX* **2** 5 (2021)
11. Shastri B J et al. *Nat. Photon.* **15** 102 (2021)
12. Silicon Photonics. From Technologies to Markets. Market and Technology Report 2021. Yole Group, https://s3.i-micronews.com/uploads/2021/05/YINTR21175-Silicon-Photonics-2021-Sample.pdf
13. Silicon Photonics 2022. Market and Technology Trends. Yole Group, https://www.yolegroup.com/product/report/silicon-photonics-2022/
14. GlobalFoundries Announces Next Generation in Silicon Photonics Solutions and Collaborates with Industry Leaders to Advance a New Era of More in the Data Center. March 7, 2022. GlobalFoundries Press Releases, https://gf.com/gf-press-release/globalfoundries-announces-next-generation-silicon-photonics-solutions-and/
15. Huang C et al. *Adv. Phys. X* **7** 1981155 (2022)
16. Kirtas M et al. "Early detection of DDoS attacks using photonic neural networks", in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop* (*IVMSP*) (Piscataway, NJ: IEEE, 2022) https://doi.org/10.1109/IVMSP54334.2022.9816178
17. Clements W R et al. *Optica* **3** 1460 (2016)
18. Al-Qadasi M A et al. *APL Photon.* **7** 020902 (2022)
19. Shen Y et al. *Nat. Photon.* **11** 441 (2017)
20. Bandyopadhyay S, Hamerly R, Englund D *Optica* **8** 1247 (2021)
21. Envise. Lightmatter, https://lightmatter.co/products/envise/
22. "High-efficiency multi-slot waveguide nano-opto-electromechanical phase modulator", Grant US-11281068-B2, https://app.dimensions.ai/details/patent/US-10884313-B2
23. Feng Y et al. *Opt. Express* **28** 38206 (2020)
24. Baghdadi R et al. *Opt. Express* **29** 19113 (2021)
25. Jacob B et al., in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, June 18–22, 2018* (Piscataway, NJ: IEEE, 2018) p. 2704
26. Machupalli R, Hossain M, Mandal M *Microprocess. Microsyst.* **89** 104441 (2022)
27. Lightelligence, https://www.lightelligence.ai
28. Lightelligence. PACE: Photonic Arithmetic Computing Engine, https://www.lightelligence.ai/index.php/product/index/2.html
29. Qu Y et al. *Sci. Bull.* **65** 1177 (2020)
30. Chao C-Y, Fung W, Guo L J *IEEE J. Sel. Top. Quantum Electron.* **12** 134 (2006)
31. Feldmann J et al. *Nature* **589** 52 (2021)
32. Burr G W et al. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **6** 146 (2016)
33. Wu C et al. *Nat. Commun.* **12** 96 (2021)
34. Prabhu M et al. *Optica* **7** 551 (2020)
35. Lucas A *Front. Phys.* **2** 5 (2014) https://doi.org/10.3389/fphy.2014.00005
36. Ахманов С А, Никитин С Ю *Физическая оптика* 2-е изд. (М.: Изд-во Московского ун-та, 2004)
37. Goodman J W *Introduction to Fourier Optics* (Englewood, CO: Roberts and Co., 2005)
38. Lin X et al. *Science* **361** 1004 (2018)
39. Yann LeCun's Home Page, http://yann.lecun.com/exdb/mnist/
40. Qian C et al. *Light Sci. Appl.* **9** 59 (2020)
41. Luo Y et al. *Light Sci. Appl.* **8** 112 (2019)
42. Mengu D et al. *IEEE J. Sel. Top. Quantum Electron.* **26** 3700114 (2020) https://doi.org/10.1109/JSTQE.2019.2921376
43. Papers with Code. Fashion-MNIST, https://paperswithcode.com/dataset/fashion-mnist
44. Papers with Code. Residual Network, https://paperswithcode.com/method/resnet
45. He K et al. "Deep residual learning for image recognition", in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, 27–30 June 2016* (Piscataway, NJ: IEEE, 2016) p. 770, https://doi.org/10.1109/CVPR.2016.90
46. Kulce O et al. *Light Sci. Appl.* **10** 25 (2021)
47. Luo X et al. *Light Sci. Appl.* **11** 158 (2022)
48. Miscuglio M et al. *Optica* **7** 1812 (2020)
49. The CIFAR-10 dataset, https://www.cs.toronto.edu/~kriz/cifar.html
50. Colburn S et al. *Appl. Opt.* **58** 3179 (2019)
51. Krizhevsky A, Sutskever I, Hinton G E, in *Advances in Neural Information Processing Systems* Vol. 25 (Eds F Pereira et al.) (Red Hook, NY: Curran Associates Inc., 2012) p. 1097
52. Kaggle. Datasets. Cats-vs-Dogs: image dataset for binary classification, https://www.kaggle.com/datasets/shaunthesheep/microsoft-catsvsdogs-dataset
53. Yan T et al. *Phys. Rev. Lett.* **123** 023901 (2019)

Photonics approaches to the implementation of neuromorphic computing

A.I. Musorin, A.S. Shorokhov, A.A. Chezhegov, T.G. Baluyan, K.R. Safronov, A.V. Chetvertukhin, A.A. Grunin, A.A. Fedyanin [(a)]
Lomonosov Moscow State University, Faculty of Physics, Leninskie gory 1, str. 2, 119991 Moscow, Russian Federation
E-mail: [(a)] *fedyanin@nanolab.phys.msu.ru*

Physical limitations on the operation speed of electronic devices has motivated the search for alternative ways to process information. The past few years have seen the development of neuromorphic photonics — a branch of photonics where the physics of optical and optoelectronic devices is combined with mathematical algorithms of artificial neural networks. Such a symbiosis allows certain classes of computation problems, including some involving artificial intelligence, to be solved with greater speed and higher energy efficiency than can be reached with electronic devices based on the von Neumann architecture. We review optical analog computing, photonic neural networks, and methods of matrix multiplication by optical means, and discuss the advantages and disadvantages of existing approaches.

Keywords: neuromorphic photonics, artificial intelligence, machine learning, reservoir computing, matrix–vector multiplication, photonic computing, neural networks, optical coprocessor, photonic tensor computing, optical Fourier transform, integrated photonics, Mach–Zehnder interferometer, ring resonators, waveguides

PACS numbers: 07.05.Mh, 42.79.Hp

Bibliography — 53 references
Uspekhi Fizicheskikh Nauk **193** (12) 1284 – 1297 (2023)
DOI: https://doi.org/10.3367/UFNr.2023.07.039505

Received 8 November 2022, revised 4 July 2023
Physics – Uspekhi **66** (12) (2023)
DOI: https://doi.org/10.3367/UFNe.2023.07.039505