

КРАТКИЙ ОБЗОР ТИПОЛОГИИ НЕЙРОНОВ И АНАЛИЗ ИСПОЛЬЗОВАНИЯ МЕМРИСТОРНЫХ КРОССБАРОВ

© 2024 г. А. А. Токарев^{1,*}, И. А. Хорин^{2,**}

¹Российский технологический университет (РТУ МИРЭА), Москва, Россия

²Физико-технологический институт им. К.А. Валиева РАН, Москва, Россия

*E-mail: santokar5@gmail.com

**E-mail: khorin@ftian.ru

Поступила в редакцию 24.06.2024 г.

После доработки 20.09.2024 г.

Принята к публикации 20.09.2024 г.

Нейроморфные технологии, использующие искусственные нейроны и синапсы, могут предложить более эффективное решение для исполнения алгоритмов искусственного интеллекта, чем традиционные вычислительные системы. Недавно были разработаны искусственные нейроны, использующие мемристоры, однако они имеют ограниченную биологическую динамику и не могут взаимодействовать напрямую с искусственными синапсами в интегрированной системе. Целью работы является обзор уровней сложности и функций нейронов и синапсов, а также анализ схемотехнического воплощения отдельных типов нейронов и нейронных сетей.

Ключевые слова: мемристор, архитектура нейронных сетей, мемристивный кроссбар, искусственный нейрон, искусственный синапс

DOI: 10.31857/S0544126924060044

1. ВВЕДЕНИЕ

Выдающиеся возможности нейронных сетей были недавно продемонстрированы производительностью AlphaGo и его вариаций [1], показавших, что ИИ, реализованный на стандартных вычислительных платформах, способен обучаться самостоятельно и превышать человеческие навыки в определенных ограниченных классах задач.

Импульсные нейронные сети (ИНС) — сети третьего поколения принципиально отличаются от предшественников в способе передачи информации, при котором импульсные сигналы заменяют реальные значения [2]. В ИНС импульсные сигналы также как в традиционных нейронных сетях необходимо умножать на синаптические веса и суммировать перед передачей нейрону, но результат напрямую не передается в функцию активации. Вместо этого нейрон накапливает воздействие этих сигналов с течением времени, пока не достигнет порогового значения, после чего посылает импульс на следующий уровень. Это означает, что, в отличие от традиционных нейронных сетей, ИНС не нужно активировать все синаптические веса и нейроны для вычисления на каждом временном шаге. Вместо этого они полагаются на совокупный эффект нескольких импульсных сигналов, чтобы

определить, следует ли продолжать передачу сигнала, что значительно снижает потребление энергии и повышает эффективность.

Следует сразу отметить, что традиционные КМОП-устройства и схемы неэффективно используют подходы к нейроморфным вычислениям, что неудивительно, поскольку транзисторы не создавались и не оптимизировались для этой цели.

С переходом от единичных нейронных сетей к нейросетевым кластерам, таким как GPT-4 и другие сверхмасштабные ИНС, стало очевидно, что несмотря на способность превосходить человеческие возможности в отдельных аспектах, энергопотребление таких ИНС значительно превышает энергопотребление человеческого мозга, имея при этом гораздо меньшую нейронную сеть. Например, обучение AlphaGo Zero выполнялось с использованием 64 графических процессоров и 19 центральных процессоров, а выполнение логического вывода осуществлялось четырьмя тензорными процессорами.

В современных компьютерах, основанных на архитектуре фон Неймана, вычисления и хранение данных физически разделены, что приводит к необходимости передачи данных между вычислительным блоком и блоком хранения и обратно.

Это неизбежно приводит к проблемам задержки и энергопотребления [3]. Особенно для нейронных сетей, требующих интенсивных вычислений с использованием памяти, основными проблемами являются ограничения полосы пропускания и энергопотребление [4]. Напротив, в человеческом мозге вычисления и хранение данных могут происходить в одном и том же физическом пространстве. Эта новая вычислительная архитектура называется “вычислениями в оперативной памяти” [5].

Мемристоры, в основе работы которых лежат процессы ионной динамики, ведут себя аналогично биологическим синапсам и нейронам, и могут точно имитировать их соответствующие функции [6], а значит должны обеспечить более эффективную реализацию нейронной сети. Был достигнут прогресс в создании аппаратных ИНС на базе окислительно-восстановительных мемристоров [7–11], мемристоров с фазовым переходом [12], органических транзисторах [13] и обычных КМОП-схемах [14–16] для эмуляции синапсов путем использования их настраиваемой проводимости в качестве синаптических весов. Однако в большинстве этих ИНС функции обработки сигналов были реализованы либо в КМОП-схемах (примерно с 10 транзисторами и более), либо в программном обеспечении, работающем на процессорах для моделирования нейронов [17], что ограничивает дальнейшее улучшение масштабируемости, стекируемости и энергоэффективности сетей.

Следующий рубеж в повышении производительности вычислений должен включать в себя динамические и адаптивные возможности природных и биологических систем. Биологические системы на всех уровнях реагируют на окружающую среду и историю. Простые молекулярные системы, такие как нуклеиновые кислоты, могут проявлять адаптивное поведение, включая репликацию и самовосстановление, вызванное местной средой. Нейроны, основные элементы обработки информации в биологических системах, выражают более 20 различных динамических поведений, обусловленных электрохимической стимуляцией из их истории и окружающей среды. Точно так же системы с более сложной организацией, такие как глазная и иммунная системы, вплоть до организмов, выражают пропорционально более высокую функциональную сложность и адаптацию. В отличие от них, современные вычислительные системы строятся на основе статических элементов со сложностью нулевого порядка [18].

2. ПЕРСПЕКТИВЫ УВЕЛИЧЕНИЯ СЛОЖНОСТИ И АДАПТИВНОСТИ СИСТЕМ

Термин “нейроморфный” был впервые использован Кейвером Мидом в 1990 году и изначально относился к аналоговым схемам, предназначенным

для имитации поведения нейронных систем [19, 20]. В настоящее время этот термин охватывает различные вычислительные системы, основанные на принципах организации и функционирования, аналогичных биологической активности мозга. Особенность нейроморфных систем заключается в необходимости обучения для их работы и дальнейшего использования. Обучение связано с синаптической пластичностью, то есть способностью системы изменять силу связей между нейронами. В искусственных системах синаптический вес является ключевым параметром, отвечающим за связь между нейронами. Главная функция синапсов заключается в хранении значений синаптических весов, изменении этих значений в соответствии с сигналами от нейронов или согласно правилам обучения, а также в обработке проходящих сигналов.

Суть обучения нейронных сетей заключается в корректировке синаптических весов с помощью определенных методов. Поскольку в ИНС импульсный сигнал дискретен, а функция активации недифференцируема, алгоритм обратного распространения нельзя использовать напрямую, как в искусственных нейронных сетях, для точного обновления синаптических весов. Методы обучения ИНС можно условно разделить на два типа: основанные на правилах и основанные на оптимизации. Методы первого типа основаны на пластичности, зависящей от времени импульса (STDP), и STDP, модулируемой вознаграждением (R-STDP).

Основной принцип правила STDP заключается в том, что прочность синаптической связи между пренейроном и постнейроном усиливается или ослабевает в зависимости от порядка их активации. Таким образом, STDP позволяет реализовать самоорганизацию и функциональную дифференциацию нейронов, то есть неконтролируемые функции обучения [21]. При R-STDP обучение происходит под управлением внешних сигналов вознаграждения, что позволяет реализовать слабо контролируемые функции обучения [22]. Эти два алгоритма обучения, основанные на правилах, обладают высокой эффективностью и адаптивностью, однако они игнорируют сложный динамический процесс и информацию о задачах внутри сети, в результате чего точность прогнозирования быстро снижается с увеличением числа уровней [23].

Методы, основанные на оптимизации, сводят к минимуму погрешность между выходными данными и ожидаемым значением с помощью математических средств, внедряя обратное распространение для глобальной оптимизации сети в процессе обучения.

С 2006 года стало ясно, что необходимо перейти к новой парадигме вычислений, называемой “пост-КМОП”, которая позволит избежать проблем, связанных с уменьшением размеров устройств, преодолеть разрыв между хранением

и обработкой данных (проблема фон Неймана) и ограничения традиционных цифровых методов обработки информации [24–26]. Это привело к появлению множества новых идей в области пост-КМОП вычислений, включая идеи, разработанные ранее [26–33], а также неэлектронных подходов, таких как оптические, квантовые и биомолекулярные вычисления [34–43].

Устройства и схемы в основном преобразуют входы в новые выходы, следуя предписанным логическим таблицам, математическим функциям и правилам условного ветвления, которые формулируются биологической системой наверху: человеком-программистом. Тем не менее, эта система является инструментом, зависящим от человека, который использует ее надлежащим образом, динамически адаптируя инструкции на основе прошлых (возможно, ошибочных) результатов и новых требований окружающей среды.

Первый прорыв в сторону от вышеупомянутой парадигмы произошел в области нейронных сетей, в частности, глубокого обучения [44, 45]. Обучение ИНС для выполнения задачи, такой как распознавание изображений, больше не использует традиционное программирование, а вместо этого основано на предоставлении размеченных примеров в сеть. Благодаря динамическому обновлению весовых коэффициентов сети, руководствуясь алгоритмами обучения (например, обратным распространением), происходит своего рода адаптация и обучение. Важно отметить, что такой подход вывел адаптацию и сложную динамику на уровень ниже пользователя, на алгоритмическом уровне, исключив явное программирование инструкций. Однако базовое оборудование остается статичным и малосложным как на этапе обучения, так и на этапе обучения, а также при последующем развертывании ИНС при выполнении вывода на основе входных данных. Следующий прорыв произойдет за счет внедрения способности к адаптации и сложной динамики на самих аппаратных уровнях.

В этом обсуждении иллюстрируется понятие сложности. Вообще говоря, сложность — это совокупность взаимодействующих процессов, выражающих нетривиальное поведение, которое было определено различными способами в разных дисциплинах [46]. Другими словами, сложность является мерой размерности динамической системы. Для наших целей мерой (“порядком”) сложности является количество дифференциальных уравнений первого порядка (или эквивалентов), необходимых для описания поведения системы, каждое из которых связано с “переменной состояния” и ее динамической эволюцией.

3. ПРЕДСТАВЛЕНИЕ СЛОЖНОЙ МОДЕЛИ РАЗНЫХ УРОВНЕЙ

Сложность проявляется на разных уровнях, от микроскопического до макроскопического: в биологических системах [47], химических процессах, экономических системах [48, 49], социальных структурах и экологических процессах.

Системы нулевого порядка — это те, которые на самом деле не имеют сложности. Они четко следуют всем изменениям входных данных, без учета временного компонента (например, временного запаздывания). Ни одна система не может быть абсолютно нулевой. Объект без массы, имеющий конечное трение, является примером такой системы.

Системы первого порядка описываются одним дифференциальным уравнением первого порядка. Объект с конечной массой и трением (где состояние характеризуется скоростью) реагирует на приложенную силу с временной задержкой из-за трения. В электрической цепи сопротивление и емкость вместе образуют цепь с характерным временем отклика (постоянной времени RC). Это требует наличия сложности первого порядка, то есть учета временной задержки.

Для систем второго порядка требуются две различные и взаимосвязанные переменные состояния, представленные двумя дифференциальными уравнениями первого порядка (или одним уравнением второго порядка), чтобы описать их соответствующую динамику. В электрической схеме сопротивление, индуктивность и емкость объединяются в цепь, которая реагирует на импульс входного напряжения затухающими колебаниями.

4. ФУНКЦИИ ИСКУССТВЕННЫХ НЕЙРОНОВ И СВИНОПАСОВ

4.1. Синаптические функции

Как реальный, так и искусственные синапсы имеют разные типы функций. На рис. 1 приведены примеры синаптического поведения, требующие разных порядков сложности. Они представляют собой идеализированные модели реакции синапсов на входные сигналы. При создании искусственного синапса на основе мемристоров добиться идеального повторения данных функций будет невозможно, а значит качество синапса можно оценить по приближенности выходного сигнала к идеализированной модели. Ниже приведены описания каждой отдельной функции синапса:

1) Направленная проводимость (нулевого порядка): необходимо, чтобы проводимость была асимметрична полярности приложенного напряжения.

2) Синаптическая задержка (первого порядка): создает определенную временную задержку между

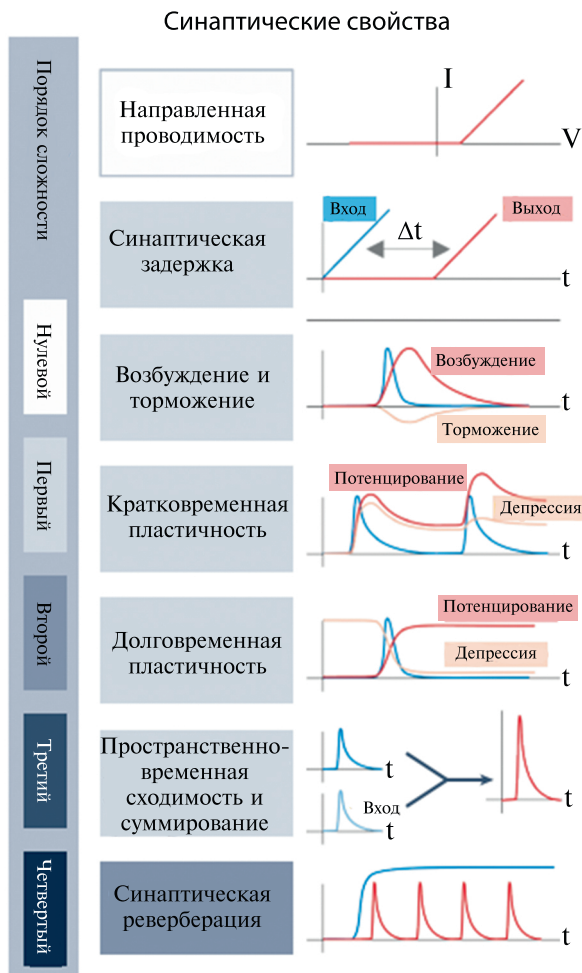


Рис. 1. Примеры синаптического поведения, требующие различных порядков сложности.

входом и выходом. Вспомним, что временная задержка требует сложности первого порядка.

3) Возбуждение и торможение (сложность первого порядка): реакция на входной сигнал, при которой выходной сигнал либо увеличивается (возбуждение), либо уменьшается (торможение). Обычно это включает временной компонент, такой как характерное время отклика, что добавляет сложность первого порядка.

4) Краткосрочная пластичность (сложность первого порядка): это реакция на временное событие (например, импульс), которая сохраняется в течение определенного времени, а затем возвращается к своему состоянию до события [50].

5) Долговременная пластичность (первый порядок): реакция на временное воздействие (например, всплеск) сохраняется практически бесконечное время, либо через потенцирование, либо через депрессию.

6) Пространственно-временная сходимость и суммирование (первый порядок): входящие

сигналы из разных каналов суммируются, включая характерную временную шкалу, что приводит к сложности первого порядка.

7) Синаптическая реверберация (второй порядок): в ответ на стимул могут возникать колебания от синаптических структур с обратной связью. Обратная связь и временные задержки создают два различных динамических процесса, которые вместе приводят к сложности второго порядка.

4.2. Нейронные функции

Как и в случае с искусственными синапсами искусственные нейроны также имеют идеализированные модели выходных сигналов (рис. 2). В среднем в сравнении с синаптическими функциями функции нейронов имеют больший порядок сложности. Для оценки точности нейронных функций также подходит сравнение результирующего сигнала с идеализированной моделью. Ниже приведены описания нейронных функций:

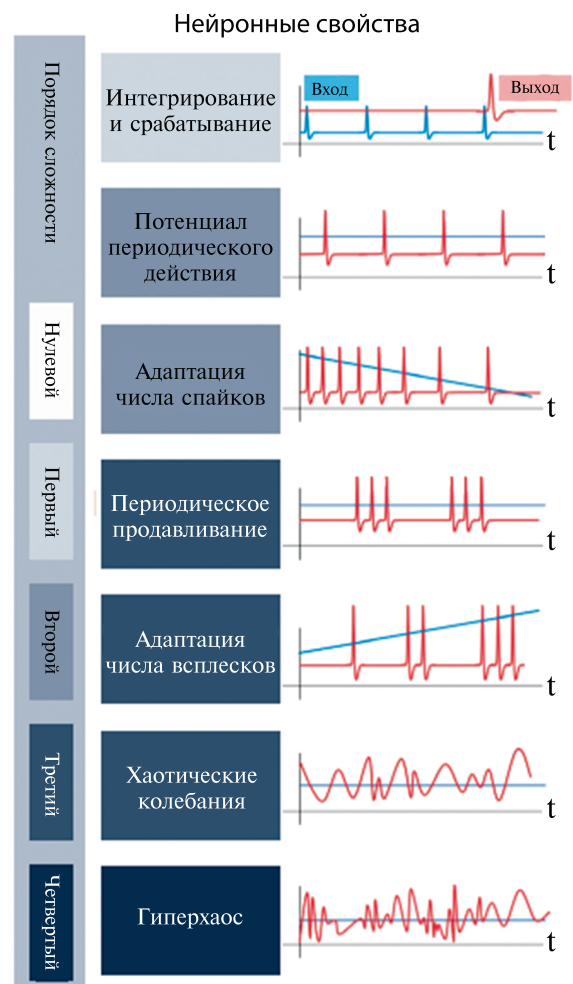


Рис. 2. Примеры нейронного поведения, требующие различных порядков сложности.

1) Интегрирование и срабатывание (сложность первого порядка): Нейроны накапливают потенциал на основе множества входных сигналов в течение определенного периода времени. Когда накопленный потенциал достигает порога, они выдают временный ответ. Подобно краткосрочной потенциации в синапсах, данный процесс вносит сложность первого порядка, аналогичную задержке [51].

2) Потенциал периодического действия (второго порядка): в ответ на постоянный входной сигнал нейроны могут производить периодические всплески, также известные как потенциалы действия. Любое колебательное поведение требует сложности второго порядка.

3) Адаптация числа спайков (второй порядок): этот процесс аналогичен процессу, производящему периодические потенциалы действия (сложность второго порядка), но с дополнительным параметром, который модулирует частоту в зависимости от входного уровня.

4) Периодическое продавливание (третий порядок): в ответ на постоянный входной сигнал нейроны могут производить периодические всплески спайков. Многопериодные колебания требуют сложности третьего порядка (например, для ведомого простого маятника).

5) Адаптация числа всплесков (третий порядок): этот процесс аналогичен процессу, производящему периодическое продавливание (сложность третьего порядка), но с дополнительным параметром, который модулирует количество всплесков в каждом пакете в зависимости от входного уровня.

6) Хаотические колебания (третий порядок): нейроны могут производить хаотическую динамику в ответ на постоянный вход, что, проще говоря, является детерминированным (неслучайным) поведением, которое трудно отследить. Напомним, что хаотическая динамика требует сложности третьего порядка.

7) Гиперхаос (четвертый порядок): технически гиперхаос — это поведение, содержащее, по крайней мере, два показателя Ляпунова, что, говоря простым языком, означает, что его гораздо труднее отследить, чем хаос. Иногда нейронные системы порождают гиперхаос, который требует сложности четвертого порядка [30].

5. ПРИМЕРЫ РЕАЛИЗАЦИИ НЕЙРОНОВ

Были предложены различные модели нейронов для описания процесса генерации ими потенциалов действия. Наверное, самая распространенная и изученная из биофизических моделей — это модель Ходжкина–Хаксли, целью которой является имитация электрофизиологического состояния мембраны нейрона [52]. На рис. 3(а) показана схема модели Ходжкина–Хаксли. Здесь два переменных резистора с разными напряжениями включения (R_{Na} и R_K) представляют каналы Na^+ и K^+ биологических нейронов соответственно. В то же время конденсатор представляет мембрану, а фиксированное сопротивление R_L представляет путь утечки из мембраны. После получения входных сигналов мембранный потенциал повышается и последовательно переводит два ионных канала, управляемых напряжением, в открытое состояние, когда мембранный потенциал превышает пороговое значение, в результате чего возникает потенциал действия. Несмотря на хорошее соответствие результатам электрофизиологических экспериментов с биологическими нейронами, высокая вычислительная стоимость модели Ходжкина–Хаксли делает ее непригодной для крупномасштабных сетей [53].

Упрощением модели Ходжкина–Хаксли стала модель Ижикевича, которая учитывает мембранный потенциал нейрона и переменную восстановления. Последняя представляет собой эффект активации тока ионов калия и эффект инактивации

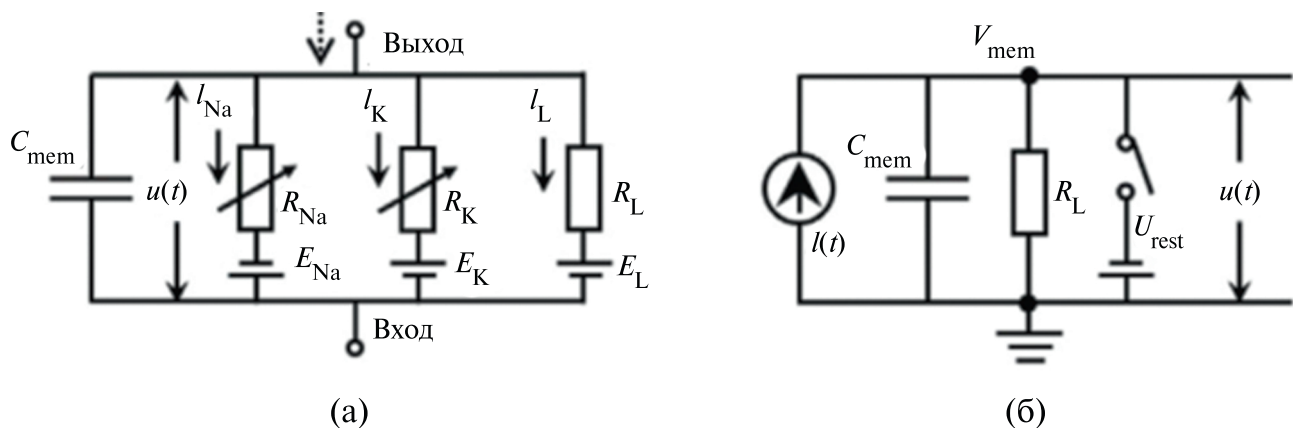


Рис. 3. Принципиальные схемы (а) модели нейронной цепи Ходжкина–Хаксли и (б) модели нейронной цепи LIF [59].

ионов натрия, обеспечивая, таким образом, отрицательную обратную связь с мембранным потенциалом [54]. В результате, модель Ижикевича работает при низкой вычислительной сложности, но при этом сохраняет различные режимы возбуждения биологических нейронов. Хотя эти модели бионических нейронов работают хорошо, не все приложения требуют такого высокого уровня биомимикрии, потребляющего слишком много энергии. Поэтому необходимо сбалансировать степень биомимикрии и вычислительные затраты в соответствии с потребностями конкретных приложений.

Другая, феноменологическая модель — LIF (leaky integrate-and-fire), цель которой состоит в том, чтобы использовать простые математические абстракции [55] для описания поведения нейронов на входе и выходе. По сравнению с моделью Ходжкина—Хаксли модель нейрона LIF относительно проста, как показано на рис. 3(б). В этой модели конденсатор используется в качестве мембраны для интеграции входного сигнала, RL используется в качестве сопротивления контура утечки, а пороговый переключатель (или переменное сопротивление) служит ионным каналом.

Модель LIF успешно зарекомендовала себя в исследованиях нейронных сетей [56—58]. Превосходный баланс, достигнутый между идеальной имитацией и энергоэффективностью, делает ее более подходящей для крупномасштабных применений в нейронных сетях.

В 2018 году был разработан ноцицептор в соответствии с его биологическим аналогом [61]. Ноцицепторы можно найти повсюду в организме человека, их расположение — на концах аксонов сенсорных нейронов. При поступлении вредного раздражителя нейроном, расположенным на свободном нервном окончании, генерируется электрический сигнал и посылается на ноцицептор, который сравнивает амплитуду сигнала с его пороговым значением и решает, будет ли сгенерирован потенциал действия и отправлен в головной мозг через спинной мозг (центральную нервную систему) или нет. Точно так же диффузионный мемристор переключается быстрее (с более высокой частотой) при более высокой интенсивности стимула. Кроме того, амплитуда выходного тока мемристора, которая могла бы быть эквивалентна частоте возбуждения ноцицептора в биосистеме, пропорциональна амплитуде входного электрического импульса. С другой стороны, если входной электрический импульс недостаточно силен, мемристор остается в исходном состоянии с высоким сопротивлением, и ток на выход не поступает, что говорит о том, что внешний раздражитель не вреден.

6. ПРИМЕРЫ РЕАЛИЗАЦИИ АРХИТЕКТУР НЕЙРОМОРФНЫХ СИСТЕМ

Аппаратная реализация нейроморфных вычислений в основном включает в себя физические нейронные сети, которые можно условно разделить на два типа: на основе КМОП-транзисторов и без использования КМОП-транзисторов. В обычном КМОП-процессе нейронные и синаптические цепи являются основой для аппаратной реализации. Для достижения сложных взаимосвязей, вдохновляемых мозгом, нейроморфные чипы обычно реализуются с помощью многоуровневой схемы горизонтальных и вертикальных перекрестных матриц (X-bar), сетей на кристалле (NoC) и многоядерных межсоединений [62, 63]. На основе этой архитектуры было успешно разработано множество чипов, демонстрирующих большой потенциал применения в таких областях, как “умный город”, обработка информации в режиме реального времени для автономного вождения и распознавание лиц: IBM TrueNorth [64, 65], типичный цифровой чип с глобальной асинхронностью и локальной синхронностью; Intel Loihi [63], реализующий онлайн-обучение на основе технологии полевых FinFET транзисторов; разработанной Стэнфордским университетом Neurogrid [66], представляющей собой цифро-аналоговый гибридный программируемый нейроморфный чип. Проблема всех этих устройств в том, что они реализуют нейронные схемы с пиковыми нагрузками [67]. Однако из-за отсутствия динамических характеристик, аналогичных нейронам, нейронным схемам на КМОП-транзисторах для реализации функций нейронов требуются десятки транзисторов, что приводит к большим затратам, проблемам энергоэффективности и масштабируемости.

Нейронные сети, выполненные по модели LIF, с относительно простой структурой также должны состоять из конденсатора, цепей восстановления и компараторов. Например, когда импульсные нейронные сети (ИНС) должны взаимодействовать с внешней средой в режиме реального времени, временные константы сети должны соответствовать постоянным в реальном мире. Обычно это приводит к использованию огромных конденсаторов, которые могут занимать до 60% площади чипа [68].

Новаторская работа в реализации ИНС на основе мемристоров с пороговым переключением, заложившая основу для масштабируемых и КМОП-совместимых нейроморфных схем появилась в 2013 году [69]. На рис. 4 показана схема нейристора, построенного с использованием двух наноразмерных мемристоров в соответствии с моделью Ходжкина—Хаксли. В этой схеме два мемристора действуют как каналы для ионов Na^+ и K^+ , соответственно. Каждый канал состоит из мемристора и емкости, соединенных параллельно.

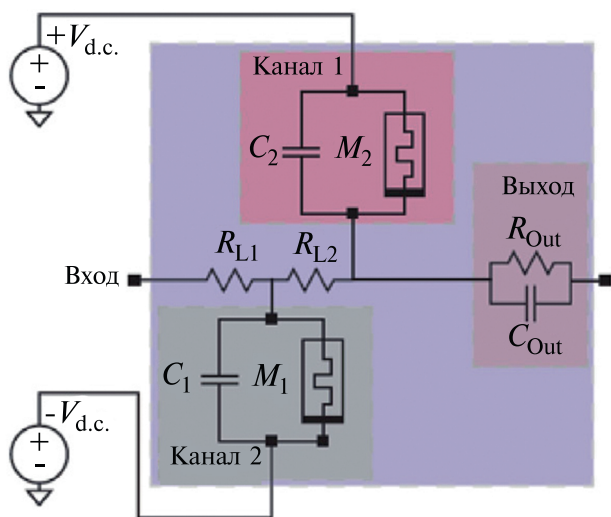


Рис. 4. Схема нейристора, построенного с использованием двух наноразмерных мемристоров в соответствии с моделью Ходжкина—Хаксли [69].

Каналы соединены нагрузочным резистором. Такой нейристор может реализовать такие поведения нейронов, как пороговое возбуждение, потенциал действия “все или ничего”, распространение импульсов без потерь, период невосприимчивости, тонизирующее возбуждение и быстрое импульсное возбуждение.

Позже эта схема была оптимизирована для реализации большего количества режимов запуска нейронов [71]. В результате получилось реализовать на мемристорах 23 режима работы биологических нейронов, которые реализуются путем настройки пассивных элементов R и C без необходимости изменения параметров мемристорного устройства. Это значительно упрощает проектирование и изготовление интегральных схем.

Следует отметить, что схемы нейронов, основанные на модели Ходжкина—Хаксли, предъявляют высокие требования к единообразию устройств и согласованию параметров между схемами, что усложняет создание крупномасштабных устройств и применение. В то же время модель LIF привлекает к себе внимание именно благодаря простой структуре и низкой вычислительной сложности. В нейронной цепи, построенной по модели LIF, необходим только один конденсатор, соединенный с мемристором. Конденсатор отвечает за интеграцию, в то время как мемристор выполняет пороговую оценку и генерирует импульсный сигнал. Был продемонстрирован нейрон LIF, основанный на мемристорах, соединенных с нагрузочным резистором, играющим роль синапса [71]. В 2020 году появилась работа, предложившая приемлемое решение для реализации нейронных цепей высокой плотности для создания эффективной системы, подобной мозгу [72]. В ней получены нескольких

режимов работы нейронов в одном устройстве и впервые проверена интеграция на кристалле.

Еще одна схема нейрона показана на рис. 5 [73]. Здесь в качестве нейрона используются соединенные вместе мемристор и выходное сопротивление, параллельно подключенные к конденсатору. Эта схема обеспечивает выполнение четырех основных функций нейрона: повышение потенциала действия по принципу «все или ничего», пороговое повышение, рефрактерный период (период времени, в течение которого клетка неспособна повторить определенное действие) и частотная характеристика с модуляцией силы. Позже на основе данной схемы был сконструирован гибридный мемристорно-КМОП-нейрон, который выполняет основную функцию нейрона LIF и позволяет настраивать подключаемые синапсы на месте [74]. Таким образом, был предложен новый способ реализации обучения на месте для будущих нейроморфных вычислительных систем.

Для того, чтобы снизить напряжения переключения до уровня биологических нейронов, (ниже 100 мВ), был разработан новый тип мемристора с протеиновыми нанопроволоками в качестве катализатора [75].

Первая ИНС размером 8×8 , полностью сделанная на мемристорных нейронах и синапсах была продемонстрирована в 2018 году [60]. На основе этой системы проверяются операции сверточного логического вывода и реализуется неконтролируемое обучение входных шаблонов.

Большинство современных ИНС упрощают нейроны до простых точечных моделей и упрощают их вычислительную функцию до функции интеграции и возбуждения, игнорируя функцию обработки информации дендритами. В результате ИНС все еще сильно отстают от биологических нейронных сетей в гибкости, надежности и энергопотреблении при выполнении сложных задач. Чтобы решить эту задачу была продемонстрирована нейронная сеть на основе двухполюсных мемристоров, состоящая из энергонезависимого мемристора на основе HfO_x в качестве искусственного синапса,

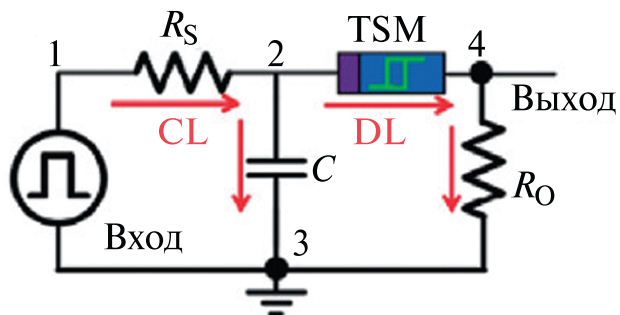


Рис. 5. Схематическая иллюстрация схемы нейрона [76].

динамического мемристора на основе $\text{TaO}_x/\text{AlO}_x$ в качестве искусственного дендрита и мемристора Мотта на основе NbOx в качестве искусственной сомы [76]. Эта работа стала важным шагом на пути к созданию более полных, энергоэффективных и точных нейронных сетей.

Генерация нейроморфных потенциалов действия в элементе схемы теоретически требует, по крайней мере, третьего порядка сложности, в то время как большинство нейронов построены из нейроморфных элементов первого или второго порядка. Был построен мемристор NbOx с динамикой третьего порядка и продемонстрированы нейронные сети из элементов третьего порядка без транзисторов [77].

6.1. Пример полностью аппаратно-реализованной мемристорной сверточной нейронной сети

Реализация практической нейроморфной вычислительной системы на основе мемристоров обычно требует интеграции нескольких мемристорных решеток. В целом, разделение весов на разные массивы полезно для параллельных вычислений, которые становятся все более необходимыми с увеличением масштабов сети. Тем не менее,

предыдущие демонстрации на основе мемристоров основывались на одном массиве [27], в основном из-за проблемы создания массивов с высокой повторяемостью.

Была предложена универсальная вычислительная архитектура нейронных сетей на основе мемристоров, показанная на рис. 6. Мемристорная ячейка использует стек материалов $\text{TiN}/\text{TaO}_x/\text{HfO}_x/\text{TiN}$ [51] и демонстрирует способность к непрерывной настройке проводимости как при потенцировании (SET), так и при депрессии (RESET) путем модуляции электрического поля и тепла [77]. Была создана аппаратная система с несколькими мемристорными матрицами с использованием специализированной печатной платы (PCB) и оценочной платы программируемой вентиляционной матрицы (ZC706, Xilinx).

Как видно из схемы, система в основном состоит из восьми мемристорных обрабатывающих элементов (ОЭ). Каждый ОЭ имеет свою собственную интегрированную матрицу мемристоров из 2048 ячеек. Каждый мемристор соединен с выходящей клеммой транзистора, а именно в конфигурации 1T1R. Каждая матрица мемристоров имеет сборку из 128×16 ячеек 1T1R. Имеется 128 параллельных

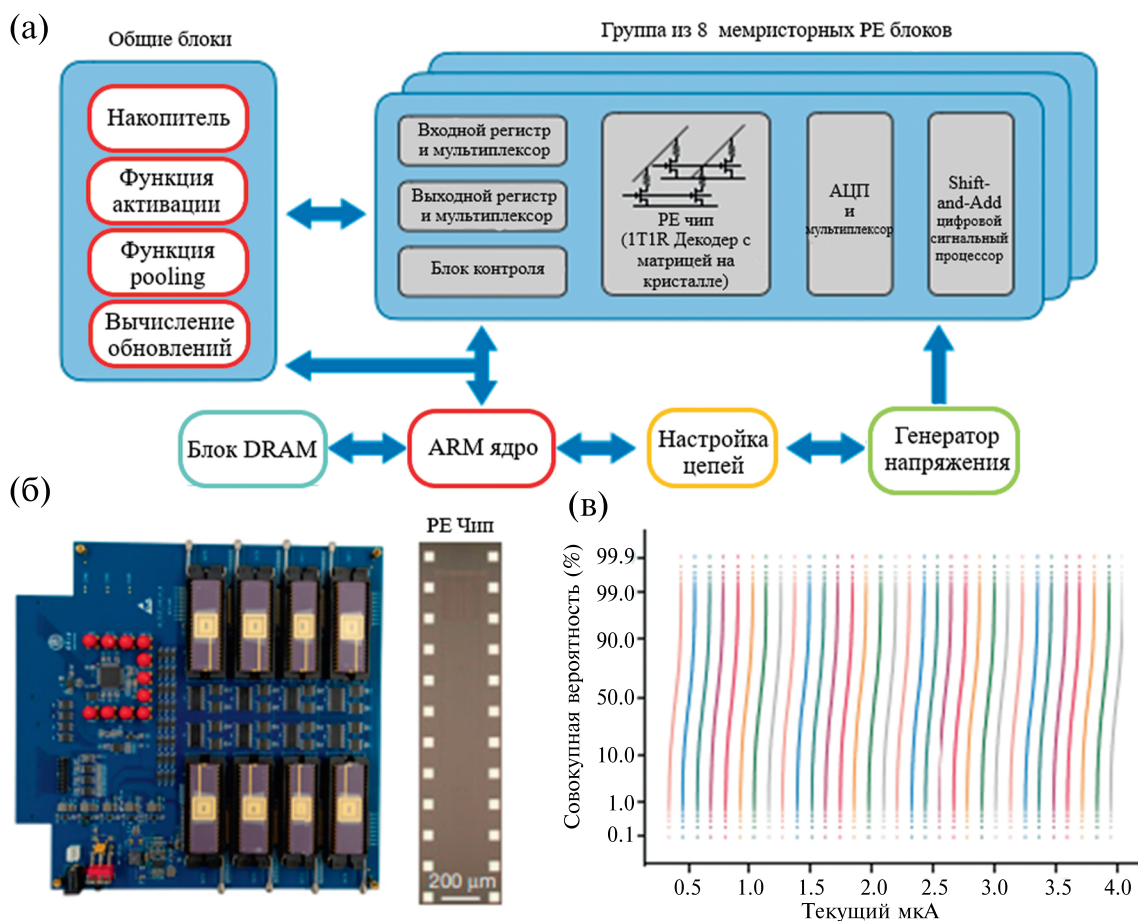


Рис. 6. Универсальная вычислительная архитектура нейронных сетей на основе мемристоров [51].

строк слов и 128 исходных строк по горизонтали и 16 битовых строк по вертикали. Эта решетка демонстрирует удивительно воспроизводимые многоуровневые состояния проводимости, как показано результатами испытаний и данными измерений на остальных 2048 ячеек.

6.2. Алгоритмическая оптимизация устройств на основе мемристоров

На сегодняшнем этапе развития технологии невозможно изготовить оптимальные устройства на основе мемристоров из-за высокой вариативности характеристик последних. Поэтому приходится разрабатывать специальные структуры алгоритмов для нивелирования влияния их неидеальных характеристик.

Было показано, что ИНС обладают высокой устойчивостью к нелинейности устройства и могут напрямую использовать асимметричные и нелинейные мемристормы в качестве синапсов [78]. С другой стороны, синапсы в сети ИНС часто имеют большое количество весов с различными значениями, а существующие мемристормы имеют лишь ограниченное число состояний проводимости. Даже самый выдающийся на сегодняшний день мемристор достиг 2^{11} состояний проводимости благодаря высокоточной настройке [79], но нельзя игнорировать сложность регулирования и надежность устройства после программирования. Для энергонезависимых мемристоров состояния с высоким и низким сопротивлением (двоичные) считаются наиболее стабильными и наименее случайными состояниями. Поэтому именно бинарные нейронные сети (БНС) демонстрируют большой потенциал применения [80, 81]. В них значительно снижаются требования к величине проводимости и повышается совместимость с энергонезависимыми мемристорами. Кроме того, вместо арифметических операций используются битовые операции (такие как XNOR), что позволяет значительно повысить эффективность алгоритма. Концепция БНС также используется для разработки удобных для аппаратного обеспечения бинарных ИНС (БИНС). В процессе их обучения используют методы суррогатного градиента, чтобы заменить производные от импульсного обучения и бинаризовать веса. Однако бинаризация весов приводит к снижению точности работы отдельного синапса, что неизбежно приводит к увеличению масштаба сети и энергопотребления.

Существуют и другие способы решить проблему неидеальных характеристик мемристоров. Например, была предложена архитектура алгоритма STELLAR, которая обладает примерно в 75 раз большей энергоэффективностью, чем цифровые ускорители, и при этом обеспечивает возможности обучения [82]. Этот алгоритм разработан как

универсальный метод, подходящий для различных структур нейронных сетей. В нем не используется схема проверки записи, а настраивается мемристор без верификации, заставляя аппаратное обеспечение адаптироваться к нелинейности и асимметрии устройства. Вычислительные модули на чипе не выполняют точных вычислений веса, а используются для определения направления обновления перекрестной матрицы весовых мемристоров. Для двухуровневой сети направление обновления зависит только от знаков выходных данных первого и второго уровней и ошибки. После определения направления на мемристорный блок подается тот же SET или RESET импульс напряжения. Для адаптации к различным задачам обучения вводится настраиваемый порог, который позволяет отфильтровать небольшие значения ошибок, избежать слишком частых обновлений весов и колебаний состояний сети и улучшить сходимость алгоритма.

Другой способ решения проблемы неидеальных характеристик мемристоров — использование гибридных алгоритмов обучения, когда модель сначала обучается в программном обеспечении “ex situ” для получения высокоточных весовых коэффициентов. Затем эти веса квантуются и сопоставляются с перекрестной матрицей мемристоров. На этом шаге происходит некоторая потеря точности и возникают ошибки в отображении. Для их компенсации последний полностью подключенный уровень (ППУ) обучается “in situ”, используя часть данных. Для такого обучения применяется нейронный процессор (NPU) на основе мемристоров, который вычисляет выходные данные, сравнивает их с меткой, чтобы получить сигнал ошибки, и вводит его в модуль обратного распространения ошибок (ЕВМ) для генерации соответствующей последовательности импульсов. Импульсы вводятся с частотой обучения, которая используется для регулировки значения проводимости мемристора на уровне ППУ до тех пор, пока вся система не достигнет заданной точности или времени [51].

При сопоставлении весов сети с большими массивами мемристоров необходимо учитывать задержки и высокое энергопотребление, что связано с использованием общих шин. Минимизировать пиковую задержку и энергопотребление позволяет Алгоритм проектирования SpiNeMap [83], который сопоставляет ИНС с нейроморфным оборудованием на основе кроссбаров. По сравнению с наиболее эффективной технологией ИНС-картографирования, SpiNeMap снижает среднее энергопотребление на 45%, а пиковую задержку на 21%. Также были предложены некоторые энергосберегающие реконфигурируемые архитектуры, такие как RESPARC для мемристорных массивов нейронных сетей с сильными импульсами [23]. Эти вычислительные архитектуры позволяют лучше

сопоставлять веса с массивами мемристоров с неидеальными характеристиками для повышения эффективности и практичности.

Однако алгоритмы обучения “in situ” могут избежать сопоставления весов и повысить устойчивость сети к неидеальностям. В качестве примера можно привести обучающуюся “in situ” многослойную нейронную сеть на основе массивов мемристоров из оксида гафния (1T1R) [18]. Сначала подается положительный импульс на нижний электрод мемристора для инициализации матрицы, затем подются синхронные положительные импульсы напряжения на верхний электрод мемристора и затвор последовательно подключенного транзистора, чтобы изменить состояние проводимости мемристора. Поскольку напряжение на затворе транзистора управляет предельным током мемристора, эта схема позволяет добиться линейности и симметричности значения проводимости, уменьшая при этом изменения проводимости между циклами и устройствами. Эта обучающая ИНС использует для обучения стохастический градиентный спуск и позволяет достичь точности классификации в 91,71% в наборе данных MNIST, что близко к результатам моделирования без дефектов. При этом алгоритм обучения предполагает, что 11% устройств в матрице не реагируют на импульсы (типичное соотношение значений дефектов, наблюдаемое в эксперименте). Таким образом, продемонстрировано, что обучение “in situ” может адаптироваться к аппаратным дефектам. Также показано, что многоуровневые сети в большей степени способствуют повышению устойчивости к дефектам. Даже если 50% устройств находятся в неисправном состоянии, сеть все равно может достичь уровня точности в 60%.

Можно еще отметить метод, имитирующий случайные колебания, генерируемые мемристорами во время переключения проводимости, путем добавления шума к весам во время тренировки [23].

7. МЕМРИСТОРНЫЕ КРОССБАРЫ

Мемристоры обычно изготавливаются в виде компактной кроссбар-структуры, состоящей из вертикальных и горизонтальных нанопроводов, расположенных в двух перпендикулярных друг другу плоскостях. Сами мемристоры располагаются на стыках каждой пары горизонтальных и вертикальных нанопроводов. Поскольку мемристоры ячейки действуют как запоминающие устройства, то кроссбар можно использовать для создания системы резистивной памяти. Кроме того, непосредственно на кроссбаре путем подачи соответствующих напряжений на линии строк и столбцов могут выполняться некоторые вычисления. Такая вычислительная парадигма, при которой хранение

информации и вычисления выполняются в одном и том же аппаратном блоке, называется вычислениями в оперативной памяти. При этом пропадает необходимость в передаче большого объема данных, а значит повышается производительность вычислений.

Для нейронной сети, реализованной с использованием мемристорных кроссбаров, качество вывода зависит от того, насколько точно в мемристорах могут быть сохранены веса. В таких системах обучение может осуществляться двумя способами: либо прямой загрузкой, либо комплексным методом. В обоих случаях внешняя система (хост) выполняет все необходимые для расчета веса вычисления. Как только тренировка завершается, на кроссбар загружаются окончательно сгенерированные веса. При использовании метода прямой загрузки обучение выполняется в полностью автономном режиме, в то время как при использовании комплексного метода кроссбар взаимодействует с хост-системой для вычисления ошибки сети. Однако при наличии в кроссбаре неисправных мемристоров веса могут быть запрограммированы не так, как того требует хост-система. Эта проблема решается как программными, так и аппаратными методами [84–87]. В общем, можно выделить три подхода.

а) Обучение с учетом ошибок: в процессе обучения модель проверяется с помощью обучающих выборок. Процесс повторяется до тех пор, пока не будет достигнут оптимальный результат. Однако, поскольку этот подход не учитывает реальные ошибки, имеющиеся в микросхеме, фактическая производительность может отличаться от ожидаемого результата.

б) Переназначение: перестановка строк с целью минимизировать общую ошибку. При таком подходе сначала определяется местоположение неисправных мемристоров, а затем применяется метод матричного согласования, который позволяет свести к минимуму общую ошибку путем приведения весов неисправных мемристоров к весу сети, рассчитанному с помощью алгоритма. Этот процесс повторяется для минимизации ошибки. В худшем случае процесс должен быть отменен, если решение, полученное на предыдущих этапах, является нежелательным.

в) Избыточные нейроны: дополнительные мемристоры (полноценные ряды матриц) позволяют преодолеть эффект неисправных мемристоров. Однако такой подход требует сложных соединений между мемристорами и инициализации веса дополнительных мемристоров. Кроме того, использование дополнительных мемристоров приводит к увеличению площади и энергопотребления.

Следует отметить, что на практике часто выбираются комбинации этих трех подходов.

В 2024 на базе мемристоров разработана новая вычислительная парадигма MobileNetV3 и выполнен ее аппаратный дизайн [88], которые обеспечивают точность > 90% в наборе данных CIFAR-10 и дают значительные преимущества в отношении вычислительных ресурсов, времени вычисления и энергопотребления. В данном решении приоритет отдается простоте, особенно в сверточных и полностью подключенных слоях, за счет сокращения количества операционных усилителей на 50% по сравнению с традиционными методами реализации. Также представлена платформа автоматического отображения с открытым исходным кодом для облегчения быстрого построения и моделирования нейронных сетей на основе мемристоров.

ЗАКЛЮЧЕНИЕ

Для нейронных цепей на основе мемристоров функции, которые могут выполнять нейроны, относительно просты. Однако необходимы более глубокие исследования, начиная с проверки функционирования устройств и заканчивая оптимизацией конструкции периферийных вспомогательных цепей с использованием мемристорных нейронов. Кроме того, невысокая надежность, проблема создания больших массивов и вариативность мемристоров ограничивают их широкое применение [89, 90]. Однородность нейронов на уровне устройства из-за случайности формирования нитей в мемристорах и интеграция кроссбар-матриц в нейронные цепи остаются основными проблемами для современных нейроморфных вычислительных чипов на основе мемристоров.

Эффективные алгоритмы ИНС являются основой нейроморфных чипов для реализации искусственного интеллекта, но существующие алгоритмы недостаточно совершенны. Причем алгоритмы для аппаратной платформы на базе мемристоров только начали разрабатываться.

Сенсорную систему на основе мемристоров необходимо объединить с системой глубокой нейроморфной обработки информации, чтобы создать систему искусственного интеллекта, объединяющую сознание, память и вычисления, которая действительно напоминает человеческий мозг.

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках Государственного задания ФТИАН им. К.А. Валиева РАН Минобрнауки РФ по теме № FFNN-2022-0019.

СПИСОК ЛИТЕРАТУРЫ:

1. *Strukov D.B., Snider G.S., Stewart D.R., Williams R.S.* The missing memristor found // *Nature*, 2008. V. 453. P. 80–83.
2. *Roy K., Laiswal A., Panda P.* Towards spike-based machine intelligence with neuromorphic computing // *Nature*, 2019. V. 575. P. 607–617.
3. *Jeon W.K.G., Lee J., Lee H. et al.* Deep learning with GPUs. In S. Kim, & G. C. Deka (Eds.), *Hardware Accelerator Systems for Artificial Intelligence and Machine Learning* (2021). (pp. 167–215). (Advances in Computers; Vol. 122). Academic Press Inc.
4. *Capra M., Bussolino B., Marchisio A. et al.* Hardware and Software Optimizations for Accelerating Deep Neural Networks: Survey of Current Trends, Challenges, and the Road Ahead // *IEEE Access*, 2020. V. 8. P. 225134–225180.
5. *Sebastian A., Le Gallo M., Khaddam-Aljameh R. et al.* Memory devices and applications for in-memory computing // *Nat. Nanotechnol.*, 2020. V. 15. P. 529–544.
6. *Xia Q.-F., Yang J.-J.* Memristive crossbar arrays for brain-inspired computing // *Nat. Mater.*, 2019. V. 18(4). P. 309–323.
7. *Jo S.H., Chang T., Ebong I. et al.* Nanoscale memristor device as synapse in neuromorphic systems // *Nano Lett.*, 2010. V. 10. P. 1297–1301.
8. *Yu S., Wu Y., Jeyasingh R. et al.* An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation // *IEEE Trans. Electron Devices*, 2011, V. 58. P. 2729–2737.
9. *Ohno T., Hasegawa T., Tsuruoka T. et al.* Short-term plasticity and long-term potentiation mimicked in single inorganic synapses // *Nat. Mater.*, 2011. V. 10. P. 591–595.
10. *Pershin Y.V., Di Ventra M.* Neuromorphic, digital, and quantum computation with memory circuit elements // *Proc. IEEE*, 2012. V. 100. P. 2071–2080.
11. *Hu S.G., Liu Y., Liu Z. et al.* Associative memory realized by a reconfigurable memristive Hopfield neural network // *Nat. Commun.*, 2015. V. 6. 7522 (2015).
12. *Burr G.W., Shelby R.M., Sidler S. et al.* Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element // *2014 IEEE International Electron Devices Meeting*, pp. 29.5.1–29.5.4.
13. *Wong H.-S.P., Lee H.-Y., Yu S. et al.* Metal-oxide RRAM // *Proceedings of the IEEE*, 2012. V. 100. № 6. P. 1951–1970.
14. *Yang J.J., Strukov D.B., Stewart D.R.* Memristive devices for computing // *Nature Nanotechnology*, 2013. V. 8. № 1. P. 13–24.
15. *van de Burgt Y., Lubberman E., Fuller E.J. et al.* A non-volatile organic electrochemical device as a

- low-voltage artificial synapse for neuromorphic computing // *Nat. Mater.*, 2017. V. 16(4). P. 414–418.
16. Merolla P.A., Arthur J.V., Alvarez-Icaza R. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface // *Science*, 2014. V. 345. P. 668–673.
 17. Sourikopoulos I., Hedayat S., Loyez C. et al. A 4-fJ/spike artificial neuron in 65 nm CMOS technology // *Front. Neurosci.*, 2017. V. 11. 123.
 18. Li C., Belkin D., Li Y. et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks // *Nat. Commun.*, 2018. V. 9. 2385.
 19. Trepel M. *Neuranatomie: Struktur und Funktion* 7. Auflage, Elsevier. Munchen. 2012. 439 P.
 20. Mead C. Neuromorphic electronic systems // *Proc. IEEE*, 1990. V. 78. P. 1629–1636.
 21. Kheradpisheh S.R., Ganjtabesh M., Thorpe S.J. et al. STDP-based spiking deep convolutional neural networks for object recognition // *Neural Networks*, 2018. V. 99. P. 56–67.
 22. Mozafari M., Ganjtabesh M., Nowzari-Daliniet A. al. Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks // *Pattern Recognition*, 2019. V. 94. P. 87–95.
 23. Peng H., Gan L., Guo X. Memristor based Spiking Neural Networks: Cooperative Development of Neural Network Architecture/Algorithms and Memristors // *Chip*, 2024. V. 3(2). 100093.
 24. Ianov D., Chezhegov A., Kiselev M. et al. Neuromorphic artificial intelligence systems // *Frontiers in Neuroscience*, 2022. DOI: 10.3389/fnins.2022.959626
 25. Chen A., Datta S., Hu X.S. et al. A survey on architecture advances enabled by emerging beyond-CMOS technologies // *IEEE Design & Test*, 2019. V. 36. P. 46–68.
 26. Dragoman M., Dragoman D. *Atomic-Scale Electronics Beyond CMOS*. Springer Nature Switzerland AG. 2021.
 27. Hebb D.O. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, Inc. 1949.
 28. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // *Psychol. Rev.*, 1958, V. 65. 386.
 29. Ивахненко А.Г., Лапа В.Г. *Кибернетические предсказывающие устройства*. Киев — 1965. 219 с.
 30. Werbos P.J. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Harvard Univ. Cambridge, Massachusetts. 1974. 453 p.
 31. Kelley H.J. Gradient theory of optimal flight paths // *ARS J.*, 1960. V. 30. P. 947–954.
 32. Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by back-propagating errors // *Nature*, 1986. V. 323. P. 533–536.
 33. Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities // *Proc. Natl Acad. Sci. USA*, 1982. V. 79. P. 2554–2558.
 34. Kohonen T. Self-organized formation of topologically correct feature maps // *Biol. Cybern.*, 1982. V. 43. P. 59–69.
 35. Manipatrun, S., Nikonov D.E., Young I.A. Beyond CMOS computing with spin and polarization // *Nat. Phys.*, 2018. V. 14. P. 338–343.
 36. Preskill J. Quantum computing in the NISQ era and beyond // *Quantum*, 2018. V. 2. P. 79.
 37. Solli D.R., Jalali B. Analog optical computing // *Nat. Photonics*, 2015. V. 9. P. 704–706.
 38. Паун Г., Розенберг Г., Саломая А. *ДНК-компьютер. Новая парадигма вычислений: пер. с англ.* — М.: Мир, 2003. — 528 с.
 39. Mead C. Neuromorphic electronic systems // *Proc. IEEE*. 1990. V. 78. P. 1629–1636.
 40. Kephart J.O., Chess D.M. The vision of autonomic computing // *Computer*, 2003. V. 36. P. 41–50.
 41. Kish L.B. Thermal noise driven computing // *Appl. Phys. Lett.*, 2006. V. 89. 144104.
 42. Unger R., Moullet J. Towards computing with proteins // *Proteins*, 2006. V. 63. P. 53–64.
 43. Upadhyay N.K., Jiang H., Wang Z. et al. Emerging memory devices for neuromorphic computing // *Adv. Mater. Technol.*, 2019. V. 4. 1800589.
 44. Bengio Y. Deep Learning of Representations for Unsupervised and Transfer Learning // *JMLR: Workshop and Conference Proceedings*, 2012, V. 27. P. 17–37.
 45. Schmidhuber J. Deep learning in neural networks: An overview // *Neural Netw.*, 2015. V. 61 P. 85–117.
 46. Chua L.O. Local activity is the origin of complexity // *Int. J. Bifurc. Chaos*, 2005. V. 15. P. 3435–3456.
 47. Csete M.E., Doyle J.C. Reverse engineering of biological complexity // *Science*, 2002. V. 295. P. 1664–1669.
 48. Stanley H.E., Amaral L.A., Buldyrev S.V. et al. Self-organized complexity in economics and finance // *Proc. Natl Acad. Sci. USA*, 2002. V. 99. P. 2561–2565.
 49. Arthur W.B. Foundations of complexity economics // *Nat. Rev. Phys.*, 2021. V. 3. P. 136–145.
 50. Wilson H.R. Voluntary generation of hyperchaotic visuo-motor patterns // *Sci. Rep.*, 2019. V. 9. 13819.
 51. Yao P., Wu H., Gao B., et al. Fully hardware-implemented memristor convolutional neural network // *Nature*, 2020. V. 577 P. 641–646.
 52. Hodgkin A.L., Huxley A.F. A quantitative description of membrane current and its application to conduction and excitation in nerve // *J. Physiol.*, 1952. V. 117(4). P. 500–544.
 53. Kistler W.G.W.M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press. 2002. 498 p.

54. *Izhikevich E.M.* Simple model of spiking neurons // IEEE Trans. Neural Netw., 2003. V. 14. P. 1569–1572.
55. *Segee B.* Methods in Neuronal Modeling: from Ions to Networks, 2nd Edition // Comput. Sci. Eng., 1999. V. 1(1). P. 81–81.
56. *Taherkhani A., Belatreche A., Liet Y. et al.* A review of learning in biologically plausible spiking neural networks // Neural Netw., 2020. V. 122. P. 253–272.
57. *Fang W., Yu Z., Chen Y. et al.* Incorporating Learnable Membrane Time Constant to Enhance Learning of Spiking Neural Networks // 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, 2021. P. 2641–2651.
58. *Duan Q., Jing Z., Zou X. et al.* Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks // Nat. Commun., 2020. V. 11. 3399.
59. *Wang Y.-H., Gong T.-C., Ding Y.-X. et al.* Redox memristors with volatile threshold switching behavior for neuromorphic computing // Journal of Electronic Science and Technology, 2022. V. 20. 100177.
60. *Wang Z.-R., Joshi S., Savell'ev S. et al.* Fully memristive neural networks for pattern classification with unsupervised learning // Nat. Electron., 2018. V. 1(2). P. 137–145.
61. *Kim Y., Kwon Y.J., Kwon D.E., Yoon K.J.* Nociceptive Memristor // Adv. Mater., 2018. V. 30, 1704320.
62. *Davies M., Srinivasa N., Lin T.-H., et al.* Loihi: a neuromorphic manycore processor with on-chip learning // IEEE Micro, 2018. V. 38. №1. P. 82–99.
63. *Imam N., Cleland T.A.* Rapid online learning and robust recall in a neuromorphic olfactory circuit // Nat. Mach. Intell., 2020. V. 2(3). P. 181–191.
64. *Merolla P.A., Arthur J.V., Alvarez-Icaza R., et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface // Science, 2014. V. 345(6197). P. 668–673.
65. *DeBole M.V., Taba B., Amir A., et al.* TrueNorth: Accelerating from zero to 64 million neurons in 10 years // Computer, 2019. V. 52(5). P. 20–29.
66. *Benjamin B.V., Gao P.-R., Mcquinn E. et al.* Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations // Proc. IEEE, 2014. V. 102(5). P. 699–716.
67. *Beck M.E., Shylendra A., Sangwan V.K., et al.* Spiking neurons from tunable Gaussian heterojunction transistors // Nat. Commun., 2020. V. 11(1). P. 1565, 1–8.
68. *Covi E., Donati E., Liang X.-P., et al.* Adaptive extreme edge computing for wearable devices // Front. Neurosci., 2021. V. 15. P. 611300, 1–27.
69. *Pickett M.D., Medeiros-Ribeiro G., Williams R.S.* A scalable neuristor built with Mott memristors // Nat. Mater., 2013. V. 12(2). P. 114–117.
70. *Yi W., Tsang K.K., Lam S.K., et al.* Biological plausibility and stochasticity in scalable VO₂ active memristor neurons // Nat. Commun., 2018. V. 9 P. 4661, 1–10.
71. *Gao L.-G., Chen P.-Y., Yu S.-M.* NbO_x based oscillation neuron for neuromorphic computing // Appl. Phys. Lett., 2017. V. 111(10). P. 103503, 1–4.
72. *Kumar S., Williams R.S., Wang Z.-W.* Third-order nanocircuit elements for neuromorphic engineering // Nature, 2020. V. 585. 7826. P. 518–523.
73. *Zhang X.-M., Wang W., Liu Q., et al.* An artificial neuron based on a threshold switching memristor // IEEE Electron. Device Lett., 2018. V. 39(2). P. 308–311.
74. *Zhang X.-M., Lu J., Wang Z.-R., et al.* Hybrid memristor-CMOS neurons for in-situ learning in fully hardware memristive spiking neural networks // Sci. Bull., 2021. V. 66(16). P. 1624–1633.
75. *Fu T.-D., Liu X.-M., Gao H.-Y., et al.* Bioinspired bio-voltage memristors // Nat. Commun., 2020. V. 11(1). P. 1861, 1–10.
76. *Li X.-Y., Tang J.-S., Zhang Q.-T., et al.* Power-efficient neural network with artificial dendrites // Nat. Nanotechnol., 2020. V. 15(9) P. 776–782.
77. *Wu W., Wu H., Gao B. et al.* A methodology to improve linearity of analog RRAM for neuromorphic computing // 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2018. P. 103–104.
78. *Kim T., Hu S., Kim J. et al.* Spiking Neural Network (SNN) With Memristor Synapses Having Non-linear Weight Update // Front. Comput. Neurosci., 2021. V. 15. 646125.
79. *Rao M., Tang H., Wu J. et al.* Thousands of conductance levels in memristors integrated on CMOS // Nature, 2023. V. 615. P. 823–829.
80. *Liang S., Yin S., Liu L. et al.* FP-BNN: Binarized neural network on FPGA // Neurocomputing, 2018. V. 275. P. 1072–1086.
81. *Simons T., Lee D.-J.* A Review of Binarized Neural Networks // Electronics, 2019. V. 8. P. 661.
82. *Zhang W., Yao P., Gao B. et al.* Edge learning using a fully integrated neuro-inspired memristor chip // Science, 2023. V. 381. P. 1205–1211.
83. *Balaji A., Das A., Wu Y. et al.* Mapping Spiking Neural Networks to Neuromorphic Hardware // IEEE Trans. Very Large Scale Integr. (VLSI) Syst., 2020. V. 28. P. 76–86.
84. *Yadav D.N., Thangkhiew P.L., Chakraborty S., et al.* Efficient grouping approach for fault tolerant weight mapping in memristive crossbar array // Memories — Materials, Devices, Circuits and Systems, 2023. V. 4. 100045.
85. *Zhang B., Uysal N., Fan D. et al.* Handling Stuck-at-Fault Defects Using Matrix Transformation for Robust Inference of DNNs // IEEE Transactions on

- Computer-Aided Design of Integrated Circuits and Systems*, 2020. V. 39. №10. P. 2448—2460.
86. Xu Y., Jin S., Wang Y., Qi Y. Aggressive Fault Tolerance for Memristor Crossbar-Based Neural Network Accelerators by Operational Unit Level Weight Mapping // *IEEE Access*, 2021. V. 9. P. 102828—102834.
 87. Xia L., Huangfu W., Tang T., et al. Stuck-at Fault Tolerance in RRAM Computing Systems // *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2018. V. 8 №1. P. 102—115.
 88. Li J., Ma L., Sham C.-W., Fu C. A Novel Computing Paradigm for MobileNetV3 using Memristor // arXiv:2402.10512v1 [cs.AR]. 16 Feb 2024. 1—15.
 89. Yan B.-N., Chen Y.-R., Li H. Challenges of memristor based neuromorphic computing system // *Sci. China Inf. Sci.*, 2018. V. 61(6). P. 060425, 1—3.
 90. Sung C., Hwang H., Yoo I.K. Perspective: a review on memristive hardware for neuromorphic computation // *J. Appl. Phys.*, 2018. V. 124(15). P. 151903, 1—13.

A Brief Overview of the Typology of Neurons and Analysis of Using Memristor Crossbars

A. Tokarev^{a,*}, I. A. Khorin^{b,**}

^aMIREA — Russian Technological University (RTU MIREA), Moscow, Russia

^bK.A. Valiev Institute of Physics and Technology of the Russian Academy of Sciences, Moscow, Russia

*E-mail: santokar5@gmail.com

**E-mail: khorin@ftian.ru

Neuromorphic technologies using artificial neurons and synapses can offer a more efficient solution for the execution of artificial intelligence algorithms than traditional computing systems. Artificial neurons using memristors have recently been developed, but they have limited biological dynamics and cannot interact directly with artificial synapses in an integrated system. The purpose of the work is to review the levels of complexity and functions of neurons and synapses, as well as to analyze the circuitry of certain types of neurons and neural networks.

Keywords: memristor, neural network architecture, memristive crossbar, artificial neuron, artificial synapse

REFERENCES

1. Strukov D.B., Snider G.S., Stewart D.R., Williams R.S. The missing memristor found // *Nature*, 2008. V. 453. P. 80—83.
2. Roy K., Laiswal A., Panda P. Towards spike-based machine intelligence with neuromorphic computing // *Nature*, 2019. V. 575. P. 607—617.
3. Jeon W.K.G., Lee J., Lee H. et al. Deep learning with GPUs. In S. Kim, & G. C. Deka (Eds.), *Hardware Accelerator Systems for Artificial Intelligence and Machine Learning* (2021). (pp. 167—215). (Advances in Computers; Vol. 122). Academic Press Inc.
4. Capra M., Bussolino B., Marchisio A. et al. Hardware and Software Optimizations for Accelerating Deep Neural Networks: Survey of Current Trends, Challenges, and the Road Ahead // *IEEE Access*, 2020. V. 8. P. 225134—225180.
5. Sebastian A., Le Gallo M., Khaddam-Aljameh R. et al. Memory devices and applications for in-memory computing // *Nat. Nanotechnol.*, 2020. V. 15. P. 529—544.
6. Xia Q.-F., Yang J.-J. Memristive crossbar arrays for brain-inspired computing // *Nat. Mater.*, 2019. V. 18(4). P. 309—323.
7. Jo S.H., Chang T., Ebong I. et al. Nanoscale memristor device as synapse in neuromorphic systems // *Nano Lett.*, 2010. V. 10. P. 1297—1301.
8. Yu S., Wu Y., Jeyasingh R. et al. An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation // *IEEE Trans. Electron Devices*, 2011, V. 58. P. 2729—2737.
9. Ohno T., Hasegawa T., Tsuruoka T. et al. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses // *Nat. Mater.*, 2011. V. 10. P. 591—595.
10. Pershin Y.V., Di Ventra M. Neuromorphic, digital, and quantum computation with memory circuit elements // *Proc. IEEE*, 2012. V. 100. P. 2071—2080.

11. *Hu S.G., Liu Y., Liu Z. et al.* Associative memory realized by a reconfigurable memristive Hopfield neural network // *Nat. Commun.*, 2015. V. 6. 7522 (2015).
12. *Burr G.W., Shelby R.M., Sidler S. et al.* Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element // 2014 IEEE International Electron Devices Meeting, pp. 29.5.1–29.5.4.
13. *Wong H.-S.P., Lee H.-Y., Yu S. et al.* Metal-oxide RRAM // *Proceedings of the IEEE*, 2012. V. 100. № 6. P. 1951–1970.
14. *Yang J.J., Strukov D.B., Stewart D.R.* Memristive devices for computing // *Nature Nanotechnology*, 2013. V. 8, № 1. P. 13–24.
15. *van de Burgt Y., Lubberman E., Fuller E.J. et al.* A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing // *Nat. Mater.*, 2017. V. 16(4). P. 414–418.
16. *Merolla P.A., Arthur J.V., Alvarez-Icaza R. et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface // *Science*, 2014. V. 345. P. 668–673.
17. *Sourikopoulos I., Hedayat S., Loyez C. et al.* A 4-fJ/spike artificial neuron in 65 nm CMOS technology // *Front. Neurosci.*, 2017. V. 11. 123.
18. *Li C., Belkin D., Li Y. et al.* Efficient and self-adaptive in-situ learning in multilayer memristor neural networks // *Nat. Commun.*, 2018. V. 9. 2385.
19. *Trepel M.* *Neuranatomie: Struktur und Funktion* 7. Auflage, Elsevier. Munchen. 2012. 439 P.
20. *Mead C.* Neuromorphic electronic systems // *Proc. IEEE*, 1990. V. 78. P. 1629–1636.
21. *Kheradpisheh S.R., Ganjtabesh M., Thorpe S.J. et al.* STDP-based spiking deep convolutional neural networks for object recognition // *Neural Networks*, 2018. V. 99. P. 56–67.
22. *Mozafari M., Ganjtabesh M., Nowzari-Daliniet A. al.* Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks // *Pattern Recognition*, 2019. V. 94. P. 87–95.
23. *Peng H., Gan L., Guo X.* Memristor based Spiking Neural Networks: Cooperative Development of Neural Network Architecture/Algorithms and Memristors // *Chip*, 2024. V. 3(2). 100093.
24. *Ianov D., Chezhegov A., Kiselev M. et al.* Neuromorphic artificial intelligence systems // *Frontiers in Neuroscience*, 2022. DOI: 10.3389/fnins.2022.959626
25. *Chen A., Datta S., Hu X.S. et al.* A survey on architecture advances enabled by emerging beyond-CMOS technologies // *IEEE Design & Test*, 2019. V. 36. P. 46–68.
26. *Dragoman M., Dragoman D.* *Atomic-Scale Electronics Beyond CMOS*. Springer Nature Switzerland AG. 2021.
27. *Hebb D.O.* *The Organization of Behavior: A Neuro-psychological Theory*. John Wiley & Sons, Inc. 1949.
28. *Rosenblatt F.* The perceptron: a probabilistic model for information storage and organization in the brain // *Psychol. Rev.*, 1958, V. 65. 386.
29. *Ивахненко А.Г., Лана В.Г.* Кибернетические предсказывающие устройства. Киев – 1965. 219 с.
30. *Werbos P.J.* *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Harvard Univ. Cambridge, Massachusetts. 1974. 453 p.
31. *Kelley H.J.* Gradient theory of optimal flight paths // *ARS J.*, 1960. V. 30. P. 947–954.
32. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning representations by back-propagating errors // *Nature*, 1986. V. 323. P. 533–536.
33. *Hopfield J.J.* Neural networks and physical systems with emergent collective computational abilities // *Proc. Natl Acad. Sci. USA*, 1982. V. 79. P. 2554–2558.
34. *Kohonen T.* Self-organized formation of topologically correct feature maps // *Biol. Cybern.*, 1982. V. 43. P. 59–69.
35. *Manipatrun, S., Nikonov D.E., Young I.A.* Beyond CMOS computing with spin and polarization // *Nat. Phys.*, 2018. V. 14. P. 338–343.
36. *Preskill J.* Quantum computing in the NISQ era and beyond // *Quantum*, 2018. V. 2. P. 79.
37. *Solli D.R., Jalali B.* Analog optical computing // *Nat. Photonics*, 2015. V. 9. P. 704–706.
38. *Паун Г., Розенберг Г., Саломаа А.* ДНК-компьютер. Новая парадигма вычислений: пер. с англ. – М.: Мир, 2003. – 528 с.
39. *Mead C.* Neuromorphic electronic systems // *Proc. IEEE*. 1990. V. 78. P. 1629–1636.
40. *Kephart J.O., Chess D.M.* The vision of autonomic computing // *Computer*, 2003. V. 36. P. 41–50.
41. *Kish L.B.* Thermal noise driven computing // *Appl. Phys. Lett.*, 2006. V. 89. 144104.
42. *Unger R., Moulton J.* Towards computing with proteins // *Proteins*, 2006. V. 63. P. 53–64.
43. *Upadhyay N.K., Jiang H., Wang Z. et al.* Emerging memory devices for neuromorphic computing // *Adv. Mater. Technol.*, 2019. V. 4. 1800589.
44. *Bengio Y.* Deep Learning of Representations for Unsupervised and Transfer Learning // *JMLR: Workshop and Conference Proceedings*, 2012, V. 27. P. 17–37.
45. *Schmidhuber J.* Deep learning in neural networks: An overview // *Neural Netw.*, 2015. V. 61 P. 85–117.
46. *Chua L.O.* Local activity is the origin of complexity // *Int. J. Bifurc. Chaos*, 2005. V. 15. P. 3435–3456.
47. *Csete M.E., Doyle J.C.* Reverse engineering of biological complexity // *Science*, 2002. V. 295. P. 1664–1669.

48. Stanley H.E., Amaral L.A., Buldyrev S.V. et al. Self-organized complexity in economics and finance // Proc. Natl Acad. Sci. USA, 2002. V. 99. P. 2561–2565.
49. Arthur W.B. Foundations of complexity economics // Nat. Rev. Phys., 2021. V. 3. P. 136–145.
50. Wilson H.R. Voluntary generation of hyperchaotic visuo-motor patterns // Sci. Rep., 2019. V. 9. 13819.
51. Yao P., Wu H., Gao B., et al. Fully hardware-implemented memristor convolutional neural network // Nature, 2020. V. 577 P. 641–646.
52. Hodgkin A.L., Huxley A.F. A quantitative description of membrane current and its application to conduction and excitation in nerve // J. Physiol., 1952. V. 117(4). P. 500–544.
53. Kistler W.G.W.M. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press. 2002. 498 p.
54. Izhikevich E.M. Simple model of spiking neurons // IEEE Trans. Neural Netw., 2003. V. 14. P. 1569–1572.
55. Segee B. Methods in Neuronal Modeling: from Ions to Networks, 2nd Edition // Comput. Sci. Eng., 1999. V. 1(1). P. 81–81.
56. Taherkhani A., Belatreche A., Liet Y. al. A review of learning in biologically plausible spiking neural networks // Neural Netw., 2020. V. 122. P. 253–272.
57. Fang W., Yu Z., Chen Y. et al. Incorporating Learnable Membrane Time Constant to Enhance Learning of Spiking Neural Networks // 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, 2021. P. 2641–2651.
58. Duan Q., Jing Z., Zou X. et al. Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks // Nat. Commun., 2020. V. 11. 3399.
59. Wang Y.-H., Gong T.-C., Ding Y.-X. et al. Redox memristors with volatile threshold switching behavior for neuromorphic computing // Journal of Electronic Science and Technology, 2022. V. 20. 100177.
60. Wang Z.-R., Joshi S., Savell'ev S. et al. Fully memristive neural networks for pattern classification with unsupervised learning // Nat. Electron., 2018. V. 1(2). P. 137–145.
61. Kim Y., Kwon Y.J., Kwon D.E., Yoon K.J. Nociceptive Memristor // Adv. Mater., 2018. V. 30, 1704320.
62. Davies M., Srinivasa N., Lin T.-H., et al. Loihi: a neuromorphic manycore processor with on-chip learning // IEEE Micro, 2018. V. 38. №1. P. 82–99.
63. Imam N., Cleland T.A. Rapid online learning and robust recall in a neuromorphic olfactory circuit // Nat. Mach. Intell., 2020. V. 2(3). P. 181–191.
64. Merolla P.A., Arthur J.V., Alvarez-Icaza R., et al. A million spiking-neuron integrated circuit with a scalable communication network and interface // Science, 2014. V. 345(6197). P. 668–673.
65. DeBole M.V., Taba B., Amir A., et al. TrueNorth: Accelerating from zero to 64 million neurons in 10 years // Computer, 2019. V. 52(5). P. 20–29.
66. Benjamin B.V., Gao P.-R., Mcquinn E. et al. Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations // Proc. IEEE, 2014. V. 102(5). P. 699–716.
67. Beck M.E., Shylendra A., Sangwan V.K., et al. Spiking neurons from tunable Gaussian heterojunction transistors // Nat. Commun., 2020. V. 11(1). P. 1565, 1–8.
68. Covi E., Donati E., Liang X.-P., et al. Adaptive extreme edge computing for wearable devices // Front. Neurosci., 2021. V. 15. P. 611300, 1–27.
69. Pickett M.D., Medeiros-Ribeiro G., Williams R.S. A scalable neuristor built with Mott memristors // Nat. Mater., 2013. V. 12(2). P. 114–117.
70. Yi W., Tsang K.K., Lam S.K., et al. Biological plausibility and stochasticity in scalable VO₂ active memristor neurons // Nat. Commun., 2018. V. 9 P. 4661, 1–10.
71. Gao L.-G., Chen P.-Y., Yu S.-M. NbOx based oscillation neuron for neuromorphic computing // Appl. Phys. Lett., 2017. V. 111(10). P. 103503, 1–4.
72. Kumar S., Williams R.S., Wang Z.-W. Third-order nanocircuit elements for neuromorphic engineering // Nature, 2020. V. 585. 7826. P. 518–523.
73. Zhang X.-M., Wang W., Liu Q., et al. An artificial neuron based on a threshold switching memristor // IEEE Electron. Device Lett., 2018. V. 39(2). P. 308–311.
74. Zhang X.-M., Lu J., Wang Z.-R., et al. Hybrid memristor-CMOS neurons for in-situ learning in fully hardware memristive spiking neural networks // Sci. Bull., 2021. V. 66(16). P. 1624–1633.
75. Fu T.-D., Liu X.-M., Gao H.-Y., et al. Bioinspired bio-voltage memristors // Nat. Commun., 2020. V. 11(1). P. 1861, 1–10.
76. Li X.-Y., Tang J.-S., Zhang Q.-T., et al. Power-efficient neural network with artificial dendrites // Nat. Nanotechnol., 2020. V. 15(9) P. 776–782.
77. Wu W., Wu H., Gao B. et al. A methodology to improve linearity of analog RRAM for neuromorphic computing // 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2018. P. 103–104.
78. Kim T., Hu S., Kim J. et al. Spiking Neural Network (SNN) With Memristor Synapses Having Non-linear Weight Update // Front. Comput. Neurosci., 2021. V. 15. 646125.
79. Rao M., Tang H., Wu J. et al. Thousands of conductance levels in memristors integrated on CMOS // Nature, 2023. V. 615. P. 823–829.
80. Liang S., Yin S., Liu L. et al. FP-BNN: Binarized neural network on FPGA // Neurocomputing, 2018. V. 275. P. 1072–1086.
81. Simons T., Lee D.-J. A Review of Binarized Neural Networks // Electronics, 2019. V. 8. P. 661.

82. *Zhang W., Yao P., Gao B. et al.* Edge learning using a fully integrated neuro-inspired memristor chip // *Science*, 2023. V. 381. P. 1205—1211.
83. *Balaji A., Das A., Wu Y. et al.* Mapping Spiking Neural Networks to Neuromorphic Hardware // *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, 2020. V. 28. P. 76—86.
84. *Yadav D.N., Thangkhiew P.L., Chakraborty S., et al.* Efficient grouping approach for fault tolerant weight mapping in memristive crossbar array // *Memories — Materials, Devices, Circuits and Systems*, 2023. V. 4. 100045.
85. *Zhang B., Uysal N., Fan D. et al.* Handling Stuck-at-Fault Defects Using Matrix Transformation for Robust Inference of DNNs // *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2020. V. 39. №10. P. 2448—2460.
86. *Xu Y., Jin S., Wang Y., Qi Y.* Aggressive Fault Tolerance for Memristor Crossbar-Based Neural Network Accelerators by Operational Unit Level Weight Mapping // *IEEE Access*, 2021. V. 9. P. 102828—102834.
87. *Xia L., Huangfu W., Tang T., et al.* Stuck-at Fault Tolerance in RRAM Computing Systems // *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2018. V. 8 №1. P. 102—115.
88. *Li J., Ma L., Sham C.-W., Fu C.* A Novel Computing Paradigm for MobileNetV3 using Memristor // *arXiv:2402.10512v1 [cs.AR]*. 16 Feb 2024. 1—15.
89. *Yan B.-N., Chen Y.-R., Li H.* Challenges of memristor based neuromorphic computing system // *Sci. China Inf. Sci.*, 2018. V. 61(6). P. 060425, 1—3.
90. *Sung C., Hwang H., Yoo I.K.* Perspective: a review on memristive hardware for neuromorphic computation // *J. Appl. Phys.*, 2018. V. 124(15). P. 151903, 1—13.