Повышение способности искусственного интеллекта к самостоятельному обучению при помощи ReRAM

Ключевые слова: искусственный интеллект, машинное обучение, нейронные сети, резистивные ОЗУ.

Улучшение параметров систем искусственного интеллекта неразрывно связано со все более глубоким пониманием принципов работы мозга человека. Реализация этих принципов в долгосрочной перспективе способна привести к созданию систем с высокой производительностью и малой потребляемой мощностью. Одним из направлений работ микроэлектронных фирм в этом плане можно считать создание нейронных сетей с использованием перспективных схем памяти и новых архитектур.

Недавние исследования фирмы Weebit Nano с использованием технологии резистивного ОЗУ на основе оксида кремния (SiO_x) описывают систему искусственного интеллекта (ИИ), построенную по принципу человеческого мозга и способную выполнять задачи неконтролируемого обучения с высокой точностью результатов. Работа проводилась совместно с исследователями Миланского технического университета (Италия), результаты представлены в недавнем совместном докладе, где подробно описывается демонстрация самообучения ИИ на основе ReRAM фирмы Weebit Nano. Технология ReRAM на основе SiO_x считается одной из основных альтернатив флеш-памяти NAND-типа благодаря потенциально в 1000 раз большему быстродействию при в 1000 раз меньшем энергопотреблении и в 100 раз большем сроке службы. Привлекательность нового ReRAM от Weebit Nano обусловлена возможностью использования существующих производственных процессов (рис. 1).

В рамках совместных работ специалисты Миланского технического университета разработали конструкцию аппаратного обеспечения, использующего ReRAM фирмы Weebit Nano, что позволило объединить эффективность сверточных нейронных сетей (CNN) с пластичностью импульсных нейронных сетей (SNN)[11], действующих аналогично человеческому мозгу. Созданное аппаратное обеспечение формирования логических выводов способно изучать новые объекты, не забывая ранее полученную в ходе обучения информацию. Кроме того, система способна адаптировать рабочую частоту для снижения энергопотребления до 50% за счет неклассифицируемых объектов, что делает ее пригодной для использования в автономных системах ИИ. Такой подход позволяет оптими-

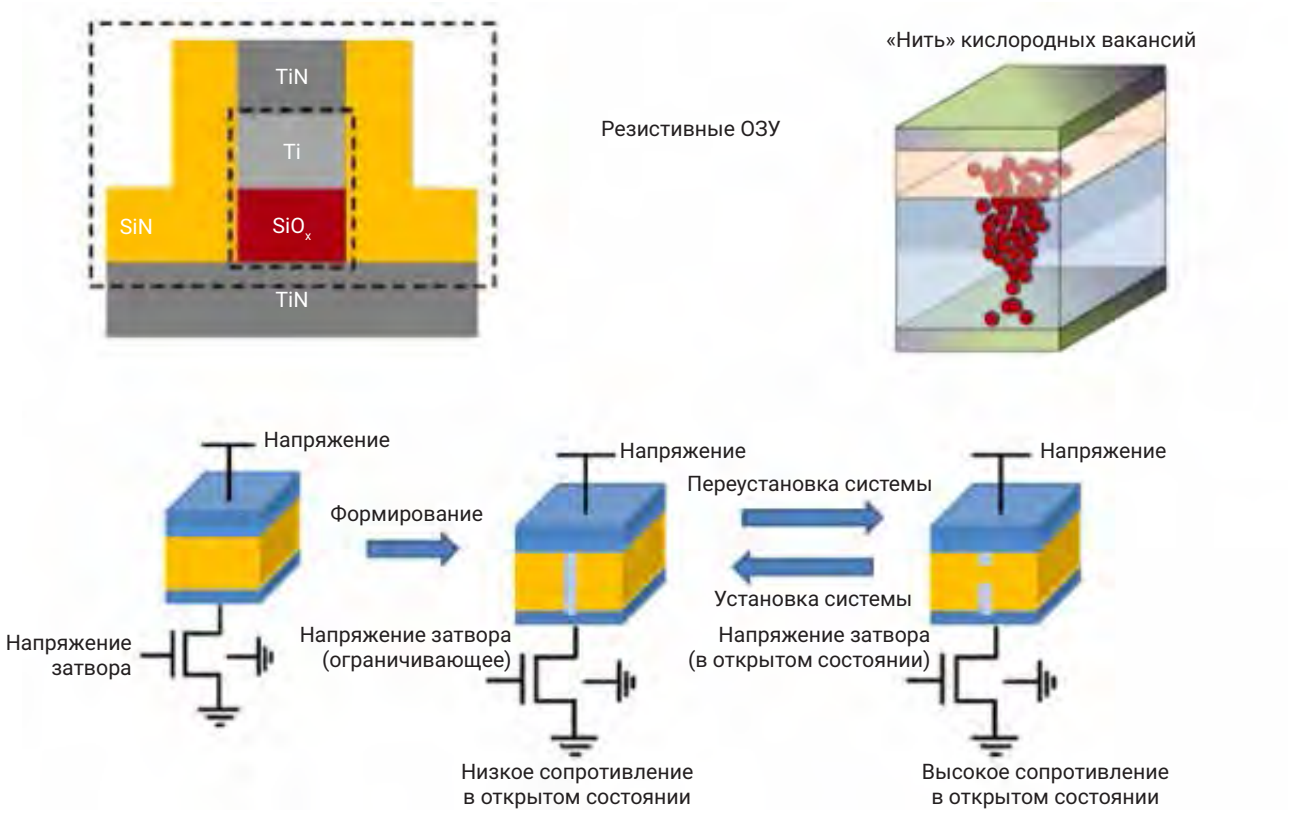

Рисунок 1. Ячейка ReRAM фирмы Weebit Nano, состоящая из двух слоев металлизации и слоя оксида кремния (SiO_x) между ними, пригодна для изготовления на существующих производственных линиях

зировать процесс классификации и повторно обучать фильтры, т. е. преодолевать основной недостаток стандартных искусственных нейронных сетей, связанный с сохранностью полученных данных.

Самая большая проблема аппаратного обеспечения ИИ на сегодня – ограничения на предмет обучения. Например, если система обучена распознавать определенные цифры, то она будет распознавать только их, игнорируя дополнительные цифры. Точно так же и потому же она не сможет самостоятельно распознавать буквы. Цель нового исследования заключалась в разработке аппаратного обеспечения на основе ReRAM, способного к непрерывному обучению. Результаты показали, что благодаря использованию уже накопленной при решении предшествующих задач обучения информации система формирования выводов способна изучить на 50% больше данных. Например аппаратное обеспечение, обученное на 100 фигурах, способно распознавать дополнительные 100 фигур без обучения. Это именно то, что происходит в мозге, когда человек чему-то учится.

В мозге при распознании какого-либо объекта нейрон формирует импульс, на что затрачивается определенная энергия. Так происходит каждый раз, но за счет существования внутренней обратной связи величина импульса при опознании уже известного объекта снижается. Исследователи смогли имитировать этот эффект на своем аппаратном обеспечении, используя ReRAM. Утверждается, что при этом был достигнут высокий уровень энергоэффективности системы, являющийся в настоящее время пределом для аппаратного обеспечения ИИ.

Специалисты Weebit Nano отмечают, что им было важно показать возможность ис-

пользования ReRAM на основе SiOx не только в качестве перспективной технологии памяти, но и в других сложных приложениях. В частности, разработчики компании всегда утверждали, что ReRAM обладает большим потенциалом для применения в нейроморфных и других сложных приложениях.

В целях выявления потенциальных областей использования ReRAM фирма Weebit Nano взаимодействует со многими исследовательскими организациями. Совместная работа с Миланским техническим университетом добавляет пластичности существующим системам ИИ за счет использования мультимодальной глубокой аналитики (рис. 2). Сегодняшний общепринятый подход к ИИ основан на контролируемом обучении, при котором необходимо прилагать значительные усилия для обучения системы. При этом в дальнейшем эта система сможет выполнять только ту задачу, для решения которой она была обучена. Однако человеческий мозг может классифицировать объекты, не подвергаясь массовому обучению, поскольку он обладает пластичностью и способен проецировать лишь несколько изображений. Weebit – не единственная компания, изучающая потенциал ReRAM для применения в области ИИ. В начале 2019 г. фирма Crossbar, один из ведущих разработчиков и производителей ReRAM, сформировала консорциум SCAiLE (SCalable AI for Learning at the Edge – «Масштабируемый ИИ для обучения в краевых вычислениях»). Задача консорциума, в который входит и Weebit Nano, – создание платформ ИИ с использованием ReRAM.

Несмотря на то что Weebit Nano по-прежнему оптимистично оценивает потенциал нейроморфных приложений, она, как коммерческая компания и стартап, работающий с целью получения дохода, продолжает уделять основное внимание выводу на рынок встраиваемых решений и разработке дискретных ReRAM. Тем не менее специалисты компании готовятся к завтрашнему дню, предполагая, что внедрение нейроморфных систем может привести к существенным изменениям. Такие гиганты, как Google, Facebook, Microsoft и Intel, прилагают значительные усилия в этой области, так как верят в огромный потенциал искусственных нейронных сетей и необходимость улучшения существующих систем. Во многом эти улучшения связаны с тем, что облачные сети требуют массового развертывания аппаратного обеспечения, а это связано с очень высоким

Источник: Crossbar

Рисунок 2. Мультимодальная глубокая аналитика

МНЕНИЕ ЭКСПЕРТА

ИИ в подавляющем большинстве случаев реализуется с помощью ускорителей нейронных сетей на основе архитектуры фон Неймана с «узким горлышком» – графических и тензорных сопроцессоров, эффективно реализующих матричные операции. Определенный сектор занимают реализации на заказных проблемно-ориентированных СБИС и ПЛИС. Эти решения уже освоены в производстве, в них используются динамическая DRAM и дорогая статическая SRAM-память с относительно большими занимаемой площадью на кристалле и энергопотреблением, тактированным доступом.

Обход ограничений архитектуры и памяти ведется в трех направлениях: поиск наиболее биологически правдоподобных («нейроморфных») принципов построения искусственной нейронной сети, поиск быстродействующей энергоэффективной памяти на новых принципах, переход к нейроморфным архитектурам. Последнее направление реализуется как на процессорных ядрах, так и в соответствии с концепцией «вычислений в памяти» – например, с помощью матриц ReRAM, как и описано в статье. Количество групп, работающих в этом направлении в России и за рубежом, весьма велико. Однако ни одна из них пока не смогла обеспечить одновременное достижение удовлетворительных значений сразу по всем основным параметрам: числу циклов переключения; времени удержания; разбросу технологических параметров от устройства к устройству; разбросу параметров при множественном считывании; температурной зависимости параметров. Описанные в статье отдельные характеристики носят скорее маркетинговый характер, чем подтверждают готовность к серийному выпуску.

Спрос на такие решения есть: отечественные разработчики и изготовители серийных беспилотных летательных аппаратов и сложных робототехнических комплексов, принявшие участие в Форуме «Микроэлектроника 2020», отмечали,

что испытывают необходимость в высокоплотной быстродействующей памяти и нейроморфных устройствах для реализации на борту задач ИИ в реальном времени в части аудио-, видео- и текстовой обработки сигнала, задач управления и помощи при принятии решений.

В НИИМЭ ведется комплекс работ по реализации нейроморфных вычислений, в частности, с помощью мемристивных элементов. Разработана комбинированная модель для физического и схемотехнического моделирования, позволяющая синтезировать уравнения работы и описывать эволюцию состояния мемристора. Также при активном взаимодействии с МФТИ, ИПТМ РАН, ИФП СО РАН, НИЦ «Курчатовский институт», ННГУ им. Н. И. Лобачевского и другими организациями непрерывно ведутся работы в области новых видов памяти – уже имеются готовые образцы.

Олег Тельминов, кандидат технических наук, начальник лаборатории исследования нейроморфных систем АО «НИИМЭ»

энергопотреблением. Соответственно, требуются решения, позволяющие снизить энергопотребление без потери эффективности.

Фирма Weebit Nano уже подала заявки на несколько патентов, касающихся технологий производства, оптимизации и программирования ReRAM на основе SiO_x для приложений памяти и инновационных разработок. Ожидается, что ReRAM также станут основой будущих нейроморфных систем.

Последняя патентная заявка подана совместно с CEA-Leti, давним партнером Weebit Nano. В ней описывается эффективный метод реализации надежного многоуровневого ЗУ на основе ReRAM, позволяющего хранить в ячейке более одного бита данных (многоуровневые ячейки, MLC). Благодаря этому возрастает емкость памяти без увеличения числа ячеек памяти или размера матрицы памяти, что повышает рентабельность ЗУ. Хотя этот метод основан на ReRAM (на основе SiO_x) фирмы Weebit Nano, его также можно распространить на любую технологию ReRAM. Эксперты компании считают, что их патент потребуется многим фирмам, специализирующимся на ReRAM и желающим освоить технологию MLC. Даже тем, кто сможет изготавливать память на MLC без использования этого патента, скорее всего, потребуется реализовать для достижения цели схожий подход.

Hilson Gary. ReRAM Research Improves Independent AI Learning. EE Times, August 31, 2020: https://www.eetimes.com/reram-research-improves-independent-ai-learning/