Платформы для искусственного интеллекта: в поисках оптимальной архитектуры

Реализация систем искусственного интеллекта требует решения целого набора задач. Это и выбор эффективной вычислительной платформы, и поиск новых моделей и методов обучения нейронных сетей, и исследование структур для физической реализации нейроморфных систем. Поиском путей построения эффективных систем искусственного интеллекта сегодня занимаются ведущие исследовательские центры и компании как в России, так и за рубежом.

Мы попросили представителей компаний и научных организаций, работающих в области создания систем ИИ, оценить современный этап развития отечественных аппаратных платформ для ИИ и обозначить ключевые тренды, существующие сегодня в области разработки специализированных архитектур и способов реализации нейроморфных вычислений.

Вячеслав Дёмин, директор-координатор по направлению «Природоподобные технологии» НИЦ «Курчатовский институт»

Сегодня аппаратные системы поддержки искусственного интеллекта стремительно развиваются во всем мире. Российские исследователи и разработчики, что весьма отрадно, не отстают от своих зарубежных коллег. Дело в том, что область разработки вычислительных систем ИИ, в том числе с нейроморфной (мозгоподобной) архитектурой, является довольно молодым направлением исследований. Поэтому сейчас в основном наблюдается конкуренция идей в этой области. А российские разработчики всегда отличались оригинальностью и перспективностью своих подходов.

Из существующих отечественных решений в области аппаратных систем поддержки ИИ первого поколения можно отметить серию нейропроцессоров NeuroMatrix компании НТЦ «Модуль», свежую платформу Robodeus разработки АО НПЦ «Элвис», тензорный микропроцессор компании IVA Technologies.

Параллельно разрабатывается второе поколение нейроморфных процессоров, основанных на новой компонентной базе – массивах мемристоров. Мемристор – это наноразмерный энергонезависимый элемент резистивной памяти, который меняет и сохраняет свое резистивное состояние (проводимость для электрического тока) под действием импульсов напряжения. Будучи собранными в массив в виде прослойки специального материала между адресуемыми шинами строк и столбцов (архитектура типа «кроссбар»), они представляют собой реализацию наборов синаптических контактов искусственных нейронов – простейших вычислительных элементов нейроморфной системы. При мемристорной реализации ядра нейропроцессора энергоэффективность и производительность его работы может возрастать на 1–3 порядка величины даже по сравнению со специализированными архитектурами нейропроцессоров 1-го поколения (которые используют стандартную ЭКБ). При этом массогабаритные характеристики таких мемристивных устройств также могут быть снижены в разы при схожей или даже увеличенной производительности. Так, НИЦ «Курчатовский институт» уже несколько лет активно работает над созданием нейроморфного процессора на основе нанокомпозитных мемристоров из ниобата лития ($LiNbO_3$) с наногранулами металла внутри оксидной матрицы. В настоящее время создается дизайн микросхемы ядра такого нейропроцессора, и одновременно идет работа по достижению требуемой воспроизводимости мемристивных матриц.

Современные системы ИИ реализуются, главным образом, на архитектурах типа графических ускорителей, так как основная операция

в нейросетевых вычислениях, являющихся сегодня драйвером бурного развития направления ИИ в целом, – это векторно-матричное умножение. Так получилось, что именно на этой операции специализируются графические процессоры (GPU), разработанные ранее для обработки графики. Их следующим логическим аппаратным усложнением является тензорный процессор, который может умножать уже не двумерные, а трехмерные массивы чисел – тензоры. Дело в том, что в наиболее коммерчески успешной архитектуре нейросетей – нейросетей сверточного типа – используется именно операция тензорного умножения. В результате графические и тензорные ядра, совмещенные с традиционными процессорами общего назначения (CPU), составляют основу всего существующего и возникающего на рынке разнообразия нейропроцессоров 1-го поколения.

Для современных и перспективных задач ИИ требуются еще более специализированные архитектуры. И, в первую очередь, речь идет даже не о мобильности таких устройств, реализующих весьма вычислительно емкие нейросетевые расчеты, а в основном об их энергоэффективности. В наше время использование ИИ становится по-настоящему массовым, причем увеличение объемов использования экспоненциальное, без видимых намеков на насыщение. В этих условиях очень скоро можно ожидать глобальный рост энергопотребления информационно-коммуникационной инфраструктуры вплоть до десятков процентов от всей производимой энергии в развитых странах при использовании традиционных вычислительных устройств. Этот показатель уже сегодня, например, в США составляет около 30%. Поэтому борьба за снижение энергопотребления ускорителей ИИ в настоящее время крайне актуальна.

Кардинальным, естественным и, возможно, единственным решением является разработка нейроморфных вычислительных систем на принципах строения и функционирования центральной нервной системы животных и человека.

Помимо повышения энергоэффективности и производительности нейроморфных вычислительных систем, использование дополнительного типа мемристоров, в которых наблюдается релаксация установленного высокопроводящего состояния в базовое непроводящее состояние (так называемые энергозависимые, или волатильные, мемристоры), позволяет эмулировать на их основе даже сами вычислительные элементы – нейроны. Правда, речь здесь идет об импульсных (спайковых) нейронах и сетях, в которых нейроны обмениваются не статическими уровнями сигналов (напряжения или тока), а импульсами определенной формы (нейропроцессоры 3-го поколения). Методы обучения и принципы работы алгоритмов ИИ на основе импульсных нейронных сетей в настоящее время все еще находятся в стадии разработки, но перспективы огромны: это уже не просто еще на порядок меньшее энергопотребление и дальнейшее уменьшение размеров устройств, но также самообучающиеся в реальном времени и в течение всего срока службы системы, интеллектуальные материалы с распределенной сенсорикой, носимые нейроинтерфейсы с мыслеуправлением, а также нейроимпланты, монтируемые на оболочку нервной ткани для протезирования двигательной активности и даже для расширения когнитивных возможностей человека.

Разработкой таких нейроморфных систем 3-го поколения также занимаются лаборатория технологий искусственного интеллекта Курчатовского института и ряд других отечественных и зарубежных научных организаций.

Николай Ивенев,
председатель совета директоров ГК «ХайТэк»

На сегодняшний день отечественных специализированных платформ для систем искусственного интеллекта (ИИ) в непосредственном доступе нет. Хотя в некоторой степени те или иные из существующих платформ, созданные для других задач, подходят и для ИИ. Однако в каждом случае необходимо проверять их применимость к конкретным задачам, благо большой проблемы в этом нет, поскольку, в отличие от западного, российский ИТ-рынок не столь жестко разделен по вертикали. Как правило, одни и те же поставщики предлагают интегрированные решения, в состав которых входит набор основных компонентов: прикладное ПО, специализированные библиотеки (в том числе библиотеки ИИ), системное ПО и аппаратная составляющая.

Сейчас на рынке представлен очень широкий спектр платформ, на которых уже реализованы системы ИИ. Однако важно понимать, что задачи, решаемые методами ИИ, и сами методы сильно зависят от постановки задачи. Примером может служить способная распознавать голосовые команды «умная» бытовая техника. Узкие специалисты уверенно скажут, что это никакой не ИИ, потому что при идентификации голосовых команд используются традиционные, неспецифичные для обработки речи методы, а аппаратной платформой является даже не микропроцессор, а микроконтроллер. Но для конечного пользователя, который зачастую не специалист в ИТ, подобная бытовая техника несомненно проявляет свойства ИИ.

Если взять другие примеры и рассмотреть распознавание смыслов естественной речи с учетом корпусов языков или машинное зрение, то здесь уже мы столкнемся с необходимостью обработки огромных массивов данных. И это уже безусловно нейросетевые технологии ИИ, для которых создаются крупнейшие по вычислительной мощности компьютеры.

Все зависит от конкретной задачи, поэтому, прежде чем начать применять методы ИИ, важно оценить масштаб проекта, правильно выбрать платформу и уже затем вести разработку. Практика показывает, что основные проблемы и большие потери возникают в тех случаях, когда на старте проекта неудачно выбрана платформа и потом приходится все переделывать.

Следует также отметить, что для очень многих современных задач ИИ вполне подходят универсальные архитектуры, которые используются в серверах, ноутбуках и даже смартфонах. Если отбросить высокоуровневые задачи, будь то распознавание слитной речи, перевод голоса в текст или машинное зрение в произвольных условиях, то остальные успешно решаются вычислительными средствами, созданными задолго до того, как ИИ получил популярность. Ведь методы и задачи ИИ появились еще несколько десятков лет назад. При этом, например, прорыв в области машинного зрения произошел в тот момент, когда хорошо известные методы в сочетании с доступной вычислительной средой стали классифицировать изображения не хуже, а порой даже лучше человека.

Еще на заре становления вычислительной техники и даже до ее появления в привычном для нас виде были доказаны очень важные фундаментальные теоремы, которые показали – все можно вычислить с помощью очень простых устройств. Вопрос исключительно в показателях качества вычислений: как долго производятся вычисления, какие нужны дополнительные предварительные или завершающие действия для того, чтобы данные были подготовлены, а результат стал бы применим. С тех пор постоянно идет поиск новых физических принципов и технологий, которые позволяют выполнять вычисления, в некотором смысле, на чем угодно. Например, на ДНК-компьютерах, в которых вычисления проводятся в пробирке при помощи химических реакций или биологических явлений.

Вычисления искусственных нейронных сетей имеют целый ряд особенностей. Причем эти особенности позволяют использовать специализированные вычислительные средства, создание которых раньше было нецелесообразным именно потому, что они были недостаточно универсальными. Теперь, благодаря современным методам ИИ и их востребованности, вычисления большого объема стали очень однообразными. Для таких вычислений не требуется высокая универсальность, можно сделать гибридный вычислитель, в котором предусмотреть универсальный фрагмент, выполняющий малую часть вычислений, и специализированный фрагмент, который берет на себя основную массу вычислений методами ИИ.

В настоящее время наблюдается ренессанс архитектурных и технологических исследований. Специалисты пытаются разобраться, можно ли весь тот багаж, который был накоплен за десятки лет, переосмыслить, добавить что-то новое и эффективно применить для узкоспециализированных вычислений. В чем-то это даже похоже на подход к квантовыми вычислениям. Ведь, во многом, привычные нам вычисления на квантовом компьютере делать бессмысленно, при этом есть и такие вычисления, в которых он имеет бесспорное и решающее преимущество перед обычным, более универсальным компьютером.

Одним из перспективных направлений развития компьютеров считаются нейроморфные технологии вычислений, более эффективно моделирующие принципы работы естественных нейросетей мозга. Способы обработки

информации в нейроморфных компьютерах плохо подходят для универсальных вычислений, но очень хороши для вычисления искусственных нейросетей. И дальше слово не столько за предлагаемыми технологиями, сколько за спросом на них. Вопрос в том, как скоро возникнут такие задачи, которые исчерпают возможности нынешних промышленных технологий и дадут шанс вновь создаваемым. Возможно, развитие технологий нейроморфных вычислений в какой-то степени повторит на новом витке историю ускорения нейросетевых вычислений более традиционными средствами, когда сначала были разработаны методы, а потом «подтянулись» технологии, то есть начал расти спрос и вложения в технологии, которые затем стали активно усовершенствоваться. Сейчас реализации нейроморфных вычислений существуют в экспериментально-исследовательских и опытно-промышленных изделиях. Отдельные задачи с их помощью решать уже можно. Но пока эффективность решения задач ИИ с помощью более традиционных технологий остается на приемлемом уровне, эти методы останутся востребованными. В том числе такой подход, как применение ускорителей вычислений искусственных нейронных сетей, построенных по классическим кремниевым нанотехнологиям.

Виктор Лучинин,
директор инжинирингового центра «Микротехнологии и диагностика», заведующий кафедрой микро- и наноэлектроники Санкт-Петербургского государственного электротехнического университета «ЛЭТИ»

В настоящее время для организации нейроморфных вычислений на аппаратном уровне используются разнообразные подходы, в том числе разработка специализированных схемотехнических решений для ключевых элементов нейромиметических систем, проектирование сложнофункциональных устройств на базе СБИС с аналоговыми и аналого-цифровыми блоками, применение графических процессоров-ускорителей и процессоров со встроенными программируемыми вентильными матрицами, использование современных достижений нанотехнологии (в области плазмоники, магноники, фотоники и наноэлектроники) для создания принципиально новых физических структур, расширяющих функционал аппаратно-реализуемых нейронных сетей, а также комбинация всех этих методов.

При этом, в качестве основных требований при реализации нейроморфных вычислений выступают обеспечение массивного параллелизма и связности архитектуры на аппаратном уровне, а также эффективная организация трафика данных, доступность большого объема памяти и низкое энергопотребление.

С позиций архитектуры нейроморфных вычислений прослеживается явная тенденция к переходу к асинхронным спайковым нейронным сетям с использованием глобально асинхронных локально синхронных архитектур и событийных протоколов передачи данных. Так, спайковые нейронные сети лежат в основе работы большинства современных нейроморфных процессоров (Neurogrid, TrueNorth, DYNAPs, SpiNNaker, Loihi, Braindrop).

С позиций материаловедения в качестве доминирующего тренда выступает разработка и создание физических структур, последующая интеграция которых лежит в основе аппаратной реализации гиперразмерных вычислений. Такой подход позволяет не только повысить эффективность узкоспециализированных архитектур (например, универсальные ускорители PUMA), но и обеспечивает возможность для разработки и создания нейроморфных модулей с интеграцией восприятия и действия (активно воспринимающие системы).

Из новых физических структур мы особое внимание уделяем мемристивным композициям с аналоговой многоуровневой перестройкой сопротивления между энергонезависимыми состояниями. Такие эффекты достигаются в гетероструктурах за счет комбинирования тонкопленочных металлооксидных мемристивных слоев, а также применения современных технологий атомно-слоевой сборки. Использование таких мемристивных композиций в структуре активных и пассивных кроссбар-массивов обеспечивает возможность их интеграции в СБИС, совместимых с КМОП-технологией, и приводит к увеличению

производительности нейроморфных систем на порядки благодаря аппаратной реализации алгоритмов матрично-векторного умножения, а также позволяет рассчитывать на их применение не только в узкоспециализированных нейроморфных архитектурах (оптимизированных для решения определенного круга задач), но и в универсальных нейроморфных модулях, автономно работающих в сложных окружениях.

Александр Черников,
заместитель начальника отделения разработки СБИС Научно-технического центра «Модуль»

Сегодня НТЦ «Модуль» – практически единственная компания на российском рынке микроэлектроники, предлагающая аппаратные решения собственного производства для систем искусственного интеллекта. У других российских компаний существуют различные прототипы на базе ПЛИС, однако это не серийные продукты. Кроме того, они реализуются или полностью на импортной аппаратной платформе или используют программируемые решения ведущих американских поставщиков ПЛИС. Такая ситуация объясняется тем, что в России пока крайне мало разработчиков процессоров и еще меньше компаний, разрабатывающих собственную архитектуру. Остальные компании пользуются зарубежными решениями. В свою очередь НТЦ «Модуль» предлагает линейку универсальных встраиваемых и автономных нейросетевых ускорителей и вычислителей для самого разного класса задач с поддержкой максимально возможного набора нейросетей, созданных в традиционных средах. Более того, мы создаем механизм портирования такого ПО на нашу архитектуру.

Подавляющее большинство современных систем искусственного интеллекта, нашедших достаточно широкое применение, строятся пока, по крайней мере в России, на вычислителях NVIDIA. Безусловное преимущество этой компании в том, что она предоставляет разработчику очень удобный пакет с полноценным стеком ПО – как для построения и обучения нейронных сетей, так и для исполнения обученных сетей в конечном устройстве. Мы стараемся не отставать от мирового лидера и показываем реальные практические результаты при реализации систем ИИ на базе собственной архитектуры NeuroMatrix. Наши решения уже успешно работают в составе демонстратора нейросетевого программно-аппаратного комплекса НПАК – масштабируемой вычислительно-коммуникационной платформы для внедрения в практику продуктов медицинского ИИ (совместная разработка НТЦ «Модуль» и ФГАУ «Ресурсный центр универсального дизайна и реабилитационных технологий»). На базе наших решений работает и комплекс мобильной медицинской компьютерной и магнитно-резонансной томографии (производитель в РФ – АО «Швабе-Медицинские системы»), а также устройство для обнаружения раковых клеток C2D2 (Cancer Cell Detection Device) компании KeyAsic.

Сегодня отдельные задачи искусственного интеллекта можно решать на универсальных, а не на специализированных процессорах – существуют достаточно простые алгоритмы, которые не очень требовательны к объему вычислений. Такой вариант решения разумен, только если универсальная архитектура уже заложена в проекте и речь не идет о сложных алгоритмах, таких как обработка видео в режиме реального времени с помощью глубоких сверточных нейронных сетей. Однако в подавляющем большинстве случаев для обучения нейронных сетей, облачных сервисов на базе нейросетевых алгоритмов, а также во встраиваемых системах интеллектуальной обработки, для решения задач ИИ необходимы специализированные архитектуры.

Что касается применения новых физических структур для реализации нейроморфных вычислений (например, на базе мемристоров), то такие решения всё еще находятся в стадии исследований. Однако можно отметить ряд наработок в сфере нейроморфных вычислителей на базе кремния, например тестовый чип Loihi от Intel. Среди российских разработок можно выделить нейроморфный процессор «Алтай». Но оценивать коммерческие перспективы таких разработок пока сложно.

Материал подготовлен В. Б. Ежовым

ОБНОВЛЕННАЯ СЕРИЯ УСТАНОВОК ЭЛЕКТРОННО-ЛУЧЕВОГО НАПЫЛЕНИЯ ТОНКИХ ПЛЕНОК В СВЕРХВЫСОКОМ ВАКУУМЕ В ГЕОМЕТРИИ «LIFT-OFF»

SemiTEq
ОТ ЗАМЫСЛА ДО ВОПЛОЩЕНИЯ

- Максимальный размер обрабатываемых подложек – Ø200 мм или 150x150 мм для стеклянных и керамических пластин
- Возможность оптимизации расхода материала за счет изменения расстояния «испаритель-подложка» в пределах 350÷500 мм

STE EB71
Стандартное исполнение

STE EB71M
Исполнение с опцией резистивного испарения в шлюзовой камере

ЗАО «НТО»
пр. Энгельса, 27
Санкт-Петербург, 194156, Россия
Тел.: +7 812 601 06 05,
Факс: +7 812 313 54 29
sales@semiteq.ru

www.semiteq.ru