DOI: 10.22184/1992-4178.2018.178.7.116.123

Аппаратное обеспечение СнК с поддержкой искусственного интеллекта

И. Попов[1]

УДК 004.8:004.31 |ВАК 05.13.05

Понятие «искусственный интеллект» (ИИ) появилось в академических кругах в середине 1950-х годов для обозначения способности систем выполнять «творческие» функции, а также науки и технологии создания таких систем. Задачи, для решения которых могут применяться интеллектуальные системы, охватывают широкий круг областей, включая классификацию больших массивов данных, распознавание голоса и графических объектов, робототехнику и экспертные системы. С философских позиций ИИ делят на «слабый» и «сильный». Под слабым ИИ понимают специализированные интеллектуальные средства, направленные на решение задач определенного класса, например распознавание текста или речи. Сильный ИИ обладает возможностями для решения задач, с которыми он еще не сталкивался, или, по другому определению, способен к пониманию смысла задач, то есть в некотором роде обладает разумом. Все решения, которые на сегодняшний день имеются на рынке, относятся к слабому ИИ, однако в научных кругах активно идет разработка алгоритмов и программно-аппаратного обеспечения, способного справляться с решениями широкого круга задач, которые ранее системе не встречались.

Само словосочетание «искусственный интеллект» наводит на мысль о реализации человеческого интеллекта в программно-аппаратном комплексе. Поэтому неудивительно, что в основе разработок в большой степени лежит изучение и попытка смоделировать процессы, происходящие в мозге человека. У человека за восприятие и последующий анализ поступившей сенсорной информации отвечает сеть из миллиардов нейронов, каждый из которых имеет несколько тысяч возможных каналов связи (передачи нервных импульсов) – аксонов. Когда человек воспринимает изображение или звук, импульсы, поступающие от того или иного органа чувств, проходят через сеть нейронов посредством связывающих их синапсов. Обучение, происходящее в течение жизни человека, создает устойчивые цепи – связи нейронов. Сравнение вновь поступивших импульсов с уже усвоенным образцом позволяет человеку делать выводы более высокого уровня сложности, например узнавать лицо знакомого человека в толпе из нескольких сотен других людей. Каждый новый шаблон, классифицированный как подходящий, усиливает связи в цепи. Этот эффект мы наблюдаем при многократном повторении одной и той же информации: чем большее количество раз мы, например, прочитаем слово, тем более надежная нейронная цепь, отвечающая за распознавание этого слова, будет создана. Данная модель является лишь небольшой частью механизмов интеллекта человека и не включает такую важную функцию, как творческое мышление [1, 2, 3].

Класс методов ИИ, в которых задача анализа и распознавания шаблонов решается на основе предварительного анализа большого количества входных данных, а не с помощью жестко заданных правил (программ), получил название *машинное обучение (МО)*. К данному классу относятся алгоритмы дерева принятия решений, метод ближайших соседей, статистический метод логистической регрессии, эвристические методы (генетический алгоритм), искусственные нейронные сети. На базе последних был создан *метод глубинного обучения (Deep Learning)* (рис. 1), отличительной особенностью которого является способность не только выделять отдельные сложные объекты, но и принимать решения на основе иерархических взаимосвязей между ними. К примеру, распознавание дорожного знака «СТОП» происходит на основе совокупности объектов: цвета (красный), формы (восьмигранник), буквы (S, T, O и P).

Данный метод появился в 1980-х годах, но получил распространение только в 2000-х, так как требовал больших вычислительных ресурсов, связанных с параллельной обработкой данных. В основе метода глубинного обучения используется модель сети нейронов человека, которая получила название *глубинная нейронная сеть (DNN – Deep Neural Network)*. В мозге человека каждый нейрон отвечает за обработку определенного небольшого атрибута входного сигнала. Глубинная

[1] Synopsys, ведущий специалист по IP-блокам, ilya.popov@synopsys.com.

Рис. 1. Соотношение методов глубинного обучения, машинного обучения и искусственного интеллекта

нейронная сеть состоит из множества слоев, каждый из которых отвечает за свою функцию обработки отдельного элемента входных данных. Каждый отдельный элемент сети имеет множественные входные связи от элементов предыдущего слоя. Таким образом, первый слой выявляет наличие элементов шаблона во входящем потоке, следующий уровень выявляет наличие заданных признаков в выходном потоке предыдущего слоя и т.д. Современные нейронные сети используют от пяти до полутора сотен слоев [4]. Существует большое количество вариантов организации искусственных нейронных сетей, которые делятся на однонаправленные – сети без обратных связей, в которых информация от слоя к слою передается только в одном направлении, и сети с обратной связью – рекуррентные нейронные сети. Благодаря наличию слоя хранения и обратной связи в сетях рекуррентной архитектуры возникает возможность обработки последовательности событий, распределенных во времени. Одним из вариантов однонаправленной искусственной нейронной сети для глубинного обучения является архитектура *сверточной нейронной сети (CNN – convolutional neural network)*, которая приобрела большую популярность благодаря эффективности при решении задач в таких областях, как распознавание образов, виртуальная или дополненная реальность и стремительно набирающее обороты направление компьютерного зрения.

Системы на базе алгоритмов глубинного обучения способны воспринимать необработанную информацию в виде входных сигналов, подаваемых в систему без какой-либо формальной организации или шаблона, и строить иерархические представления, которые позволяют классифицировать данные. Нейронные сети, являющиеся основным инструментом систем глубинного обучения, используются для обработки и классификации данных, собираемых разнообразными датчиками. Данные анализируются, и полезная информация затем отправляется обратно пользователям в режиме реального времени либо используется на последующих этапах работы ИИ.

Сверточная нейронная сеть является частным случаем нейронной сети. Она состоит из нескольких слоев различного типа: *сверточных, подвыборочных* и *полносвязанных*. Задача сети заключается в классификации объекта путем многократного перехода от конкретного изображения к более обобщенному представлению посредством выделения определенных свойств (сверточный слой) с их последующим объединением (полносвязанный слой).

Каждый слой сети представляет собой двумерную матрицу, над которой выполняются определенные операции и алгоритмы. Следует отметить, что для различных задач используются различные алгоритмы и структуры сетей. Также каждая сеть имеет свой уникальный набор весовых коэффициентов для выделения элементов, необходимых для эффективной классификации.

Вне зависимости от архитектуры нейронной сети, для выполнения требуемой функции сеть должна быть предварительно обучена. Цель обучения – определение набора коэффициентов каждого нейрона в сети, в результате которого она становится способной выдавать результат с приемлемой для решаемой задачи погрешностью (рис. 2).

Для обучения DNN/CNN требуются значительные вычислительные ресурсы, которые были недоступны в начале работ по созданию ИИ. Появление ресурсов для реализации этих алгоритмов в настоящее время обуславливает рост популярности DNN/CNN и ИИ в целом.

Обучение сети производится на большом подмножестве заранее охарактеризованных объектов. Так например, система по распознаванию образов использует библиотеки изображений с описанием того, что на этих снимках содержится (рис. 3). Чем больше объектов в библиотеке, тем лучше результат обучения, однако создание крупных библиотек требует времени и дорогостоящих решений для хранения.

Процесс обучения характеризуется очень большим количеством итераций. Каждый слой сети может обладать обратной связью, что не только существенно увеличивает количество итераций и необходимых вычислительных ресурсов, но и определяет жесткие требования по скорости обмена данными и производительности подсистем памяти. Вычислительная сложность задач обучения такова, что она неизбежно требует применения многоядерных и распределенных систем вычисления, при этом должна быть обеспечена синхронная работа множества устройств – как специализированных ускорителей для работы с матрицами, так и процессоров общего назначения (CPU). Во многих случаях

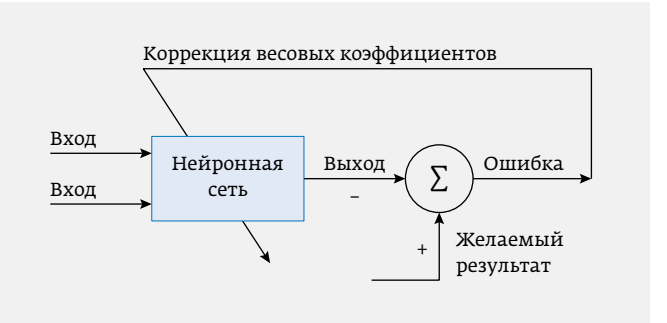

Рис. 2. Схема обучения нейронной сети

обучение происходит на облачных вычислительных кластерах. Полученные в результате графы со специфическими весовыми коэффициентами используются в конечных устройствах для непосредственной обработки входных данных в режиме реального времени.

На этапе обучения во главу угла ставится производительность, которая включает в себя вычислительную мощность, эффективность подсистемы памяти, а также скорость передачи больших массивов данных, то есть пропускную способность протоколов межсоединений, которые реализуются на различных уровнях иерархии вычислительного кластера.

Традиционные процессоры общего назначения на базе RISC-архитектуры недостаточно эффективны при работе с матрицами. Они обладают высокой производительностью в выполнении сложных операций над небольшим объемом данных, в то время как работа с нейронными сетями требует диаметрально противоположного подхода – эффективного выполнения простых операций над большими массивами. Гораздо лучше с подобными задачами справляются *векторные процессоры SIMD (single instruction, multiple data – одиночный поток команд, множественный поток данных)* и *процессоры со сверхдлинным командным словом (VLIW – very long instruction word)*. Данная архитектура реализована в *графических процессорах (GPU)* и в специализированных ускорителях для обработки нейронных сетей. В обоих случаях специализированные решения обладают ограниченным набором команд и требуют наличия «по соседству» процессора общего назначения. Очевидно, что скорость обмена данными между специализированными вычислителями и процессорами общего назначения является одним из важнейших факторов, определяющих производительность системы в целом.

Традиционно для межсоединения различных устройств, расположенных на одном кристалле либо на одной плате, используется протокол *PCI Express (PCIe)*. Однако для систем с интенсивным обменом данными скорость передачи 16 GT/s может быть недостаточной. Кроме того, в мультипроцессорной системе стоит задача обеспечения когерентности экземпляров данных, хранящихся в локальной памяти каждого устройства. Чтобы решить эти проблемы, группой ведущих компаний – разработчиков аппаратуры и программного обеспечения был создан консорциум для разработки нового протокола, получившего название *CCIX – Cache Coherent Interconnect for Accelerators*. В данном протоколе используются топология и физический уровень протокола PCIe, и его можно рассматривать как расширение PCIe с увеличенной до 25 GT/s скоростью и возможностью обеспечивать когерентность встроенных памятей устройств, подключенных к данной шине с помощью интерфейса *CXS (CCIX Stream Interface)*. Многолетний опыт и лидирующие позиции компании Synopsys в качестве поставщика решений для PCIe в сочетании с ее активным участием в деятельности консорциума CCIX позволяют предложить разработчикам надежное и проверенное решение, отвечающее самым высоким требованиям (рис. 4). Помимо контроллера CCIX и надежного высокоскоростного блока физического уровня (PHY), компания Synopsys также предлагает разработчикам средства верификации и прототипирования [5].

Помимо обеспечения высокой скорости обмена данными между элементами, расположенными на одном кристалле, для облачных кластеров обработки данных также принципиальна пропускная способность и задержка при передаче данных между оборудованием, расположенным в различных центрах обработки информации. Для этой цели компания Synopsys предлагает решения для реализации высокоскоростных протоколов из семейства Ethernet.

В результате процесса обучения создается алгоритм, позволяющий получать результат определенного качества в режиме реального времени, используя в качестве входной информации данные от различных сенсоров.

Очевидно, что требования, предъявляемые к аппаратному обеспечению для обучения и для применения нейронных сетей существенно различны. В отличие от системы обучения, которая рекурсивно работает с огромным объемом данных и, следовательно,

Рис. 3. Обучение сверточной нейронной сети на основе исходного подмножества объектов

определяющими требованиями к которой являются высокая производительность, в том числе достигаемая посредством распараллеливания, и объем памяти, в задачу оборудования по применению нейронных сетей входит лишь однократное (не итерационное) выполнение алгоритма, однако сделать это необходимо наиболее эффективным способом и во многих задачах, таких как распознавание речи и компьютерное зрение – в реальном времени. Для этих систем критическим требованием является скорость получения ответа, при этом распараллеливание выполнения задачи может быть малоэффективным. Например, в системах помощи водителю не имеет особого смысла разбиение видео, полученного с камер, на отдельные кадры с последующей независимой параллельной обработкой каждого из них. Обработка первого кадра в идеале должна завершиться до поступления от камеры следующего.

Компьютерное зрение (рис. 5) – наиболее распространенное применение технологий ИИ, служащее для обработки как статической, так и динамической визуальной информации. В последние годы количество видеоданных растет экспоненциально. Причина тому – огромное количество всевозможных их источников, начиная от смартфонов и заканчивая камерами видеонаблюдения. Большинство этих данных впоследствии передается по коммуникационным сетям и хранится в облачных хранилищах. По приблизительной оценке, доля визуальной информации в общем объеме интернет-трафика составляет порядка 80%. Один лишь популярный онлайн видео-сервис генерирует приблизительно пять часов видео каждую секунду [1]. При столь огромных объемах данных этого типа возникла острая необходимость в применении новых подходов к управлению ими, их анализу и использованию. С другой стороны, видеоданные являются наиболее сложноструктурируемым типом данных. Они представляют собой поток пикселей – точек с заданными атрибутами, который без классификации и атрибуции с большой долей вероятности станет информационным мусором. Вычленение информации из потока пикселей является весьма сложной задачей для вычислительных алгоритмов. Необходимо разделять технологии обработки изображений (в которых, впрочем, также применяются различные методы ИИ) и технологии компьютерного зрения. Первые превращают поток пикселей в изображение, компьютерное зрение превращает изображение в информацию.

Несмотря на постоянный рост вычислительных мощностей, соответствие оборудования для реализации CNN основным метрикам эффективности является нетривиальной задачей. Упрощенно, задачу можно сформулировать следующим образом: необходимо выполнить алгоритм CNN как

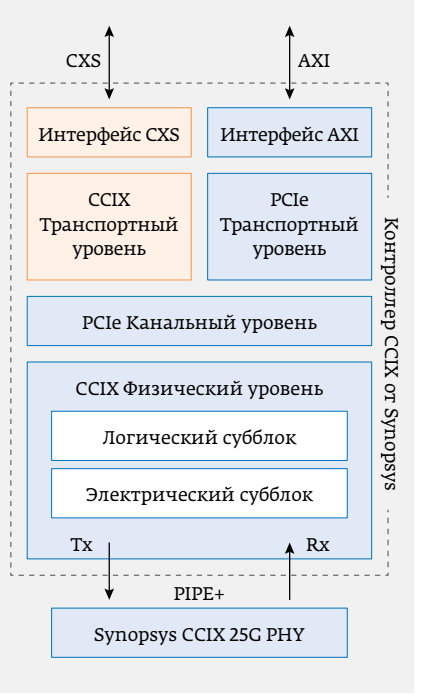

Рис. 4. Контроллер и блок физического уровня протокола CCIX от компании Synopsys

можно более быстро и эффективно. Понятие эффективности для систем ИИ можно рассматривать в двух плоскостях – энергетической и экономической. В качестве метрики энергоэффективности можно определить количество классифицируемых элементов в единицу времени, отнесенное к затрачиваемой энергии. Метрика энергоэффективности в конечном итоге определяет максимальную теоретическую производительность системы. Экономическая эффективность – это стоимость классификации одного объекта в единицу времени.

В настоящее время множество CNN-систем используют оборудование на основе процессоров общего назначения и графических процессоров, которое не подходит для маломощных и недорогих встраиваемых решений. Также существуют реализации CNN на ПЛИС, но они неидеальны для многих встраиваемых систем – отчасти из-за более высокого энергопотребления и низкой эффективности

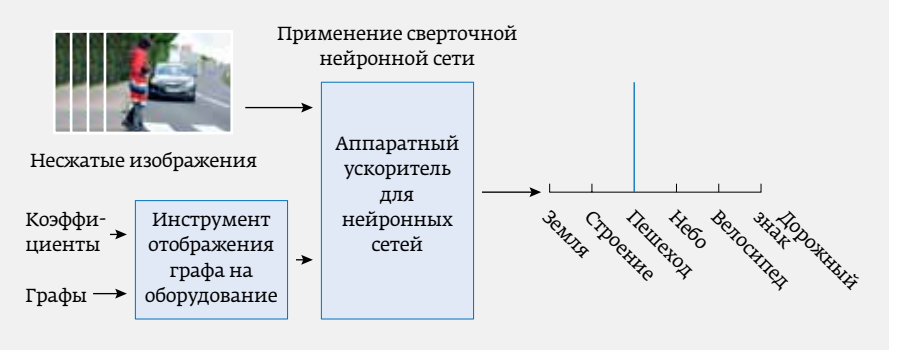

Рис. 5. Применение нейронной сети в компьютерном зрении

использования оборудования. Существуют ПЛИС-решения с жестко запрограммированной логикой, оптимизированной под конкретную CNN [4, 6]. Однако структура CNN изменяется в зависимости от результатов обучения, поэтому непрограммируемая, жестко закодированная логика значительно ограничивает пространство решаемых задач и, следовательно, не подходит для большинства приложений компьютерного зрения. Чтобы удовлетворить потребность в возможности программного изменения структуры сети, можно использовать процессор DSP или VLIW.

В сегменте встраиваемых решений существуют достаточно мощные и энергоэффективные решения, например такие, как ARC HS. Однако даже эти процессоры не обеспечивают необходимую гибкость в случае, когда производительности одного ядра недостаточно. Многоядерные системы, созданные на их основе, в силу специфики вычислительных операций склонны к деградации производительности из-за возникновения узких мест при обращении к общей памяти, а при использовании встроенной в ядра памяти (кэш-памяти и подобной) возникает проблема обеспечения когерентности данных в многоядерной системе. Графические процессоры помогли вступить в эпоху машинного обучения. Повышение производительности благодаря сокращению геометрических размеров кристалла в сочетании с вычислительной мощностью графических процессоров обеспечивает необходимую для выполнения алгоритмов глубинного обучения производительность. Однако графические процессоры были первоначально созданы для обработки графики и лишь позже были применены для задач глубинного обучения. Высокая потребляемая мощность ограничивает их применение в чувствительных к энергопотреблению встраиваемых системах [3, 4].

Исходя из этого, лучшим подходом к реализации оборудования для работы с CNN является подход, при котором определенные задачи будут выполняться на специализированном оптимизированном для эффективного выполнения сверток и связанного с ним перемещения данных оборудовании, имеющем возможность оптимизации и программирования.

С целью создания универсального программируемого процессора для обработки широкого спектра нейронных сетей были разработаны векторные сигнальные (DSP) процессоры, сочетающие в себе элементы архитектуры процессоров VLIW и SIMD. Способность таких процессоров одновременно производить несколько операций умножения-накопления (MAC) позволяет им выполнять двумерные свертки, необходимые для CNN-алгоритмов, более эффективно, чем графическому процессору. Добавление большего количества MAC в векторный DSP-процессор позволит обрабатывать больше CNN за один цикл и повысить частоту кадров при работе с изображениями. Добавив к процессору данного типа специальные ускорители CNN, можно получить более высокую эффективность по потребляемой мощности и площади.

Следующим шагом в повышении эффективности является объединение специализированного, но гибкого модуля обработки CNN с векторным DSP-процессором. Специализированный CNN-модуль должен поддерживать все обычные операции обработки CNN (такие, как свертки, объединение, поэлементные операции). Векторные DSP-процессоры по-прежнему необходимы для пред- и постобработки видеоизображений. Специализированный CNN-модуль призван не только обеспечить заданную производительность, измеряемую в количестве MAC в секунду, но и необходимую пропускную способность подсистемы памяти для обеспечения MAC-устройств необходимыми данными. Различные применения методов ИИ требуют различных комбинаций аппаратного обеспечения – VLIW, DSP, SIMD и CNN-модулей.

Компанией Synopsys в линейке процессоров ARC EV (Embedded Vision – встраиваемое зрение) предлагается гибкое и энергоэффективное решение для систем компьютерного зрения на базе сверточных нейронных сетей. Процессоры ARC EV 6-й серии (рис. 6) наряду со скалярным 32-битным ядром ARC HS также располагают 512-битным SIMD/VLIW DSP-процессором и специализированным CNN-модулем. DSP-процессор исполняет функции обработки изображений, такие как фильтрация, геометрические преобразования, преобразование цветовых пространств, обнаружение объектов и другие функции из библиотек обработки изображений OpenCV и OpenVX. Специализированные CNN-модули отвечают за свертку, сегментацию и классификацию объектов. Количество CNN-модулей и количество MAC-устройств в каждом CNN-модуле выбирается исходя из сложности нейронной сети и требований по производительности и потребляемой мощности. Модули объединены специализированной шиной ARConnect, которая, помимо низкой задержки и высокой пропускной способности, обеспечивает когерентность кэш-памятей, встроенных в каждый модуль. Обмен данными подсистемы с остальными компонентами системы на кристалле (СнК) ARC EV осуществляется по стандартному протоколу AXI4.

Процессоры ARC EV6x способны развивать производительность до 4,5 TMAC/с и обеспечивать обработку нескольких видеопотоков с разрешением 4K. ARC EV поддерживают любые типы сверточных сетей, включая такие распространенные сети как AlexNet, VGG-16, GoogLeNet, YOLO, Faster R-CNN, SqueezeNet и ResNet.

Особенности реализации специализированного CNN-модуля заключаются в том, что он способен обрабатывать 32-битые CNN-графы, используя 12-битные CNN-модули при сохранении качества распознавания и при этом существенно снижая энергозатраты. ARC EV, реализованный на технологии FinFET 16 нм, обладает энергоэффективностью 2000 ГMAC/(Вт·с).

Линейка процессоров ARC Embedded Vision представлена тремя продуктами: EV61 – одноядерный процессор с одним 32-битным скалярным модулем и 512-битным DSP,

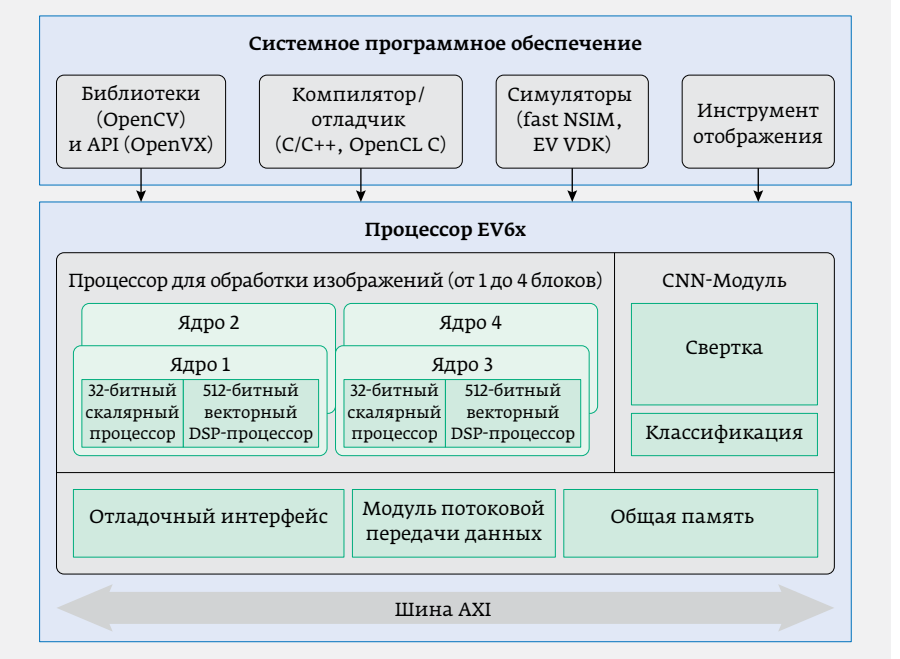

Рис. 6. Структура процессора ARC EV 6-й серии

EV62 – двухъядерный и EV64 – четырехъядерный процессор. Все три продукта могут быть расширены конфигурируемым CNN-модулем, содержащим в себе 880, 1760 или 3520 устройств умножения-накопления (далее: 880 MAC CNN, 1760 MAC CNN и 3520 MAC CNN соответственно) [7].

Одной из областей применения технологии компьютерного зрения, получившей активное развитие в последние несколько лет, стала область транспорта и транспортной инфраструктуры. Особое место здесь занимают современные системы помощи водителю (Advanced Driver-Assistance Systems – ADAS) и электроника для беспилотных автомобилей. К этому классу продуктов промышленность предъявляет особые требования в плане надежности и безопасности, регламентируемые международным стандартом ISO 26262.

Стандарт ISO 26262, выпущенный в 2011 году и обновленный в 2018-м, регламентирует функциональную безопасность электрических и/или электронных систем, используемых на транспортных средствах, в соответствии с уровнями безопасности автомобильной электроники (ASIL). Цель состоит в том, чтобы путем определения функциональных требований к компонентам системы свести к минимуму вероятность случайного аппаратного сбоя. Стандарт регламентирует как требования к процессу разработки, так и необходимые конструктивные меры, призванные предотвратить системные отказы. Линейка продуктов ARC EV6x отвечает самому высокому уровню функциональной безопасности – ASIL D.

Очевидно, что любое программируемое решение малоэффективно без соответствующих средств разработки. Для семейства процессоров EV6x таким продуктом является DesignWare ARC MetaWare EV Development Toolkit. Это комплексная среда программирования, основанная на стандартах компьютерного зрения и нейронных сетей, включая OpenCV, OpenCL C, OpenVX, которая также включает инструменты, позволяющие портировать на ARC EV6x нейронные сети, полученные с помощью таких популярных средств, как Caffe и TensorFlow. Помимо поддержки уже известных алгоритмов нейронных сетей, DesignWare ARC MetaWare EV Development Toolkit предоставляет возможность создания новых алгоритмов, оптимизированных под конкретные задачи системы ИИ [1, 2, 6].

Вернемся к обсуждению вопроса энергоэффективности. Потребляемая мощность напрямую связана с количеством выделяемого тепла, которое в свою очередь лимитировано возможностями системы охлаждения кристалла. Например, автономному транспортному средству может потребоваться значительная производительность системы ИИ и построенной на ее основе системы компьютерного зрения. Одна или несколько 8-мегапиксельных камер, работающих с частотой 60 кадров в секунду, могут требовать производительности 20–30 TMAC/с в рамках минимально возможного бюджета мощности. В линейке EV6x от Synopsys эта задача решается двумя способами: путем масштабирования количества MAC-устройств в каждом CNN-модуле и путем масштабирования количества экземпляров CNN-модулей. В качестве примера рассмотрим систему, построенную на базе ARC EV61. В верхней части рис. 7 показан процессор EV61 с 880 MAC CNN для небольших приложений, таких как маломощные интеллектуальные устройства для Интернета вещей. EV61 может интегрировать в себе 880 MAC, 1760 MAC или 3520 MAC CNN для удовлетворения требований конкретных задач. Для приложений, требующих более высокой производительности, несколько процессоров EV могут быть объединены с помощью шины AXI или специально подобранной высокопроизводительной матрицы сети на кристалле (NoC) [4, 7, 8].

Для энергоэффективных решений компьютерного зрения на базе ИИ выбор решений с конфигурируемыми CNN-модулями кажется интуитивным. Однако в этом случае возникает проблема определения необходимой и достаточной конфигурации и оценки производительности и энергоэффективности на ранних этапах проекта.

Например, функция распознавания лица, часто используемая в мобильных устройствах с питанием от батареи, в зависимости от требуемого размера кадра, частоты кадров и других параметров может потребовать несколько

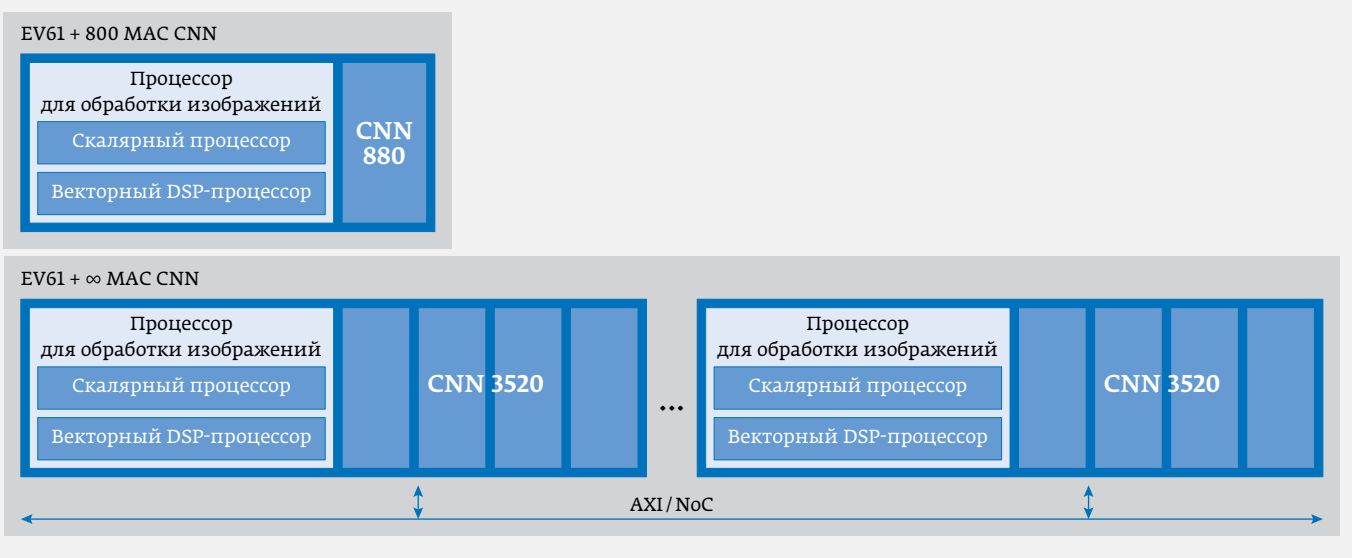

Рис. 7. Масштабирование в линейке процессоров ARC EV6x

сотен ГМАС/с. СнК должна, с одной стороны, обладать достаточной производительностью, чтобы обработать эту сеть, с другой – оставаться в рамках бюджета мощности проекта, который может составлять лишь несколько сотен милливатт.

К сожалению, сравнение различных решений с поддержкой ИИ – задача непростая. Аппаратные решения обладают большим пространством конфигураций и часто еще не доступны в виде реальных кристаллов. Все реализации отличаются друг от друга, что затрудняет сравнение энергоэффективности или производительности различных решений и конфигураций. Кроме того, в отличие от классических задач, для CNN на сегодняшний день нет стандартов для сравнения производительности. Платформы для прототипирования на базе ПЛИС могут обеспечить оценку производительности системы, но не способны дать точную оценку потребляемой мощности [2, 3, 8].

Одним из наиболее распространенных способов оценки энергопотребления является запуск целевого алгоритма на уровне межрегистровых передач (register transfer level – RTL) или списка цепей. Моделирование позволяет зафиксировать переключательную активность всей логики. Для небольших проектов симуляция может быть завершена за несколько часов (например, путем запуска тестов CoreMark или Dhrystone на встроенном ядре RISC). Для больших проектов симуляция выполняется медленно. В случае больших графов CNN, обрабатывающих потоки с высокой частотой кадров, чтобы достичь устойчивого состояния для измерения мощности, симуляция может занять недели.

Для сокращения временных затрат используются ускорители моделирования – аппаратные эмуляторы, предназначенные для имитации поведения проектируемых микросхем. При этом время выполнения симуляции по сравнению с программным моделированием может снижаться в десятки тысяч раз.

Примером такого решения является эмулятор ZeBu фирмы Synopsys (рис. 8). Такой аппаратно-программный комплекс позволяет сохранять прозрачность моделируемого RTL-описания путем вывода большого количества диагностической информации.

Synopsys ZeBu дает огромное преимущество для анализа и измерения потребляемой мощности для разработчиков как IP-блоков, так и систем на кристалле. Сервер ZeBu – это самая быстрая система эмуляции в отрасли для моделирования СнК. Система поддерживает расширенные режимы использования, включая тестирование функций управления

Рис. 8. Эмулятор Synopsys ZeBu

питанием на проектируемом чипе, всестороннюю отладку и интеграцию со средствами анализа и отладки Verdi. На базе ZeBu возможно построение гибридных систем эмуляции с виртуальными прототипами, а также систем моделирования для исследования архитектурных решений. ZeBu обладает дополнительными возможностями для точного вычисления потребляемой мощности на протяжении миллиона машинных циклов моделируемого проекта. Использование эмулятора ZeBu позволяет разработчикам аппаратного обеспечения СнК с поддержкой ИИ подобрать оптимальную конфигурацию и настроить энергопотребление всех элементов в системе [8].

Необходимость в применении специализированных решений для аппаратного обеспечения СнК с поддержкой ИИ продиктована требованиями промышленности к гибкости и эффективности выполнения специфических задач. Системы ИИ представляют собой наиболее сложные изделия электроники, когда-либо созданные человеком. В ближайшие несколько лет аналитики прогнозируют трехкратное увеличение размера рынка оборудования для глубинного обучения [3, 4, 6]. Synopsys работает с ведущими разработчиками и поставщиками оборудования по всему миру. Этот опыт реализован в эффективных и надежных решениях, которые снижают риск, ускоряют время выхода на рынок и обеспечивают привлекательность продукта для заказчиков.

ЛИТЕРАТУРА

1. Introduction to Convolutional Neural Networks for Visual Recognition // Stanford University School of Engineering. – https://www.youtube.com/watch?v=vT1JzLTH4G4.
2. **Nandra N.** IP With Near Zero Energy Budget Targets Machine Learning Applications // SNUG Silicon Valley 2018 – IP Track. – Synopsys. March 22, 2018.
3. **Lowman R.** The DNA of an Artificial Intelligence SoC. – Synopsys. – https://www.synopsys.com/designware-ip/technical-bulletin/the-dna-of-an-ai-soc-dwtb_q318.html.
4. **Gwennap L., Demler M., Case L.** A Guide to Processors for Deep Learning. First Edition. – The Linley Group. September 2017.
5. **Solomon R.** CCIX – What And Why? – Semiconductor Engineering. September 14, 2017. – https://semiengineering.com/ccix-what-and-why/.
6. Artificial Intelligence: Powering the Next Generation of Processors. – SEMICO Research Corporation. April 2018.
7. DesignWare EV61, EV62 and EV64 Processors Datasheet. – https://www.synopsys.com.
8. **Cooper G.** Implementing High-Performance Deep Learning without Breaking Your Power Budget. – Synopsys. – https://www.synopsys.com/designware-ip/technical-bulletin/implementing-high-performance-deep-learning-2018q1.html.
9. **Thompson M.** Embedding Artificial Intelligence into Our Lives. – Synopsys. – https://www.synopsys.com/designware-ip/technical-bulletin/embedding-artificial-intelligence-into-our-lives-2018q1.html.