DOI: 10.18137/RNU.V9187.24.01.P.137 УДК 004.021:004.932

А.В. Смирнов, Е.Е. Власов, Е.Н. Пистун

АНАЛИЗ ПЕРСПЕКТИВЫ ПРИМЕНЕНИЯ ПЛАТФОРМЫ ПЛИС ДЛЯ РЕАЛИЗАЦИИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ С ЦЕЛЬЮ ВЫСОКОСКОРОСТНОЙ ОБРАБОТКИ ИНФОРМАЦИИ ИЗ ВИДЕОПОТОКА

Аннотация. Статья посвящена исследованию перспективы применения платформы программируемых логических интегральных схем (ПЛИС) для реализации алгоритмов машинного обучения с целью высокоскоростной обработки информации из видеопотока. Обосновывается актуальность и значимость темы исследования. Дается краткое обоснование роли и значимости развития применения платформы ПЛИС для реализации алгоритмов машинного обучения. Постулируется, что в настоящее время ПЛИС в среднем более производительна в общих задачах, чем процессор специального назначения. Когда производительность процессора недостаточна и нет готовых схем для решения проблемы, используются ПЛИС, особенно когда речь идет о 100-процентной гарантированной обработке и гарантированной задержке. В отличие от мощного процессора, мощная ПЛИС не требует дополнительной памяти, потребляет меньше энергии и может быть реализована в компактном устройстве. ПЛИС более производительна, но менее гибка с точки зрения реализации некоторых задач. В то же время сфера применения технологии довольно широка. Анализ научной литературы позволил сделать заключение о ее эффективности в различных областях.

Ключевые слова: платформы программируемых логических интегральных схем, ПЛИС, машинное обучение, видеокамеры, цифровые видеокамеры, программируемые интегральные схемы, платформы, обработка, видеопоток.

A.V. Smirnov, E.E. Vlasov, E.N. Pistun

ANALYSIS OF THE PROSPECTS OF USING THE FPGA PLATFORM FOR THE IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS FOR THE PURPOSE OF HIGH-SPEED PROCESSING OF INFORMATION FROM THE VIDEO STREAM

Abstract. The article addresses the prospects of using the FPGA platform for the implementation of machine learning algorithms for the purpose of high-speed processing of information from a video stream. The authors substantiate the relevance and significance of the research topic. A brief justification of the role and significance of the development of the FPGA platform for the implementation of machine learning algorithms is given. It is stated that at present the FPGA is on average more productive in general tasks than a special-purpose processor. When the processor performance is insufficient and there are no ready-made circuits to solve the problem, FPGAs are used. Especially when it comes to 100-percent guaranteed processing and guaranteed latency. Unlike a powerful processor, a powerful FPGA does not require additional memory, consumes less energy and can be implemented in a compact device. FPGA is more productive, but less flexible in terms of implementing some tasks. At the same time, the scope of the technology is quite wide. The analysis of the scientific literature allowed us to conclude about its effectiveness in various fields.

Keywords: FPGA, machine learning, video cameras, digital video cameras, programmable integrated circuits, platforms, processing, video stream.

Смирнов Артем Владимирович
аспирант, МИРЭА – Российский технологический университет, Москва. Сфера научных интересов: математическое моделирование алгоритмов обнаружения движущихся объектов. Автор восьми опубликованных научных работ.
Электронный адрес: 5725757@mail.ru

Власов Евгений Евгеньевич
кандидат технических наук, доцент кафедры программного обеспечения систем радиоэлектронной аппаратуры, МИРЭА – Российский технологический университет, Москва. Сфера научных интересов: математическое моделирование алгоритмов обнаружения движущихся объектов. Автор более 20 опубликованных научных работ.
Электронный адрес: vlasov@mirea.ru

Пистун Евгений Николаевич
ассистент кафедры программного обеспечения систем радиоэлектронной аппаратуры, МИРЭА – Российский технологический университет, Москва. Сфера научных интересов: мониторинг и диагностика вычислительных комплексов. Автор четырех опубликованных научных работ.
Электронный адрес: pistun@awooo.ru

Введение

Как известно, передовая сетевая инфраструктура развертывается не только для решения проблемы резкого увеличения объема передаваемых данных, но и для обеспечения возможности обработки данных в различных частях сети, например, на периферии, в ядре и в облаке. Неудивительно, что бóльшая часть данных – это либо видео, либо изображения, объем которых растет в геометрической прогрессии и будет продолжать расти в ближайшие годы. Область машинного обучения, которая подпадает под сферу применения искусственного интеллекта (далее – ИИ), позволяет машинам извлекать уроки из данных без необходимости явного программирования. Он использует алгоритмы для обнаружения закономерностей и составления прогнозов или решений на основе входных данных. И наоборот, программируемые в полевых условиях вентильные матрицы представляют собой тип интегральной схемы, которая может быть реконфигурирована после производства для выполнения конкретных задач. *Программируемые логические интегральные схемы* (далее – ПЛИС) специально созданы для одновременного выполнения широкого спектра задач, что делает их привлекательным решением для приложений, требующих больших вычислительных затрат, включая машинное обучение. Требуется больше вычислительных ресурсов, чтобы приспособиться к массовому росту объема данных [1]. Поскольку типы приложений разнообразны, в центрах обработки данных существует большое количество рабочих нагрузок по обработке видео или изображений.

Перспективы применения платформы программируемых логических интегральных схем для реализации алгоритмов машинного обучения

ПЛИС – это электронный компонент, используемый для создания цифровых интегральных схем. ПЛИС являются привлекательным вариантом для машинного обучения, поскольку их можно адаптировать для ускорения конкретных вычислений и обработки огромных объемов данных в режиме реального времени.

Автоматический целостный подход к проектированию может значительно улучшить качество реализации алгоритмов машинного обучения с целью высокоскоростной обработки информации из видеопотока на ПЛИС. Рассмотрим методику, которая включает в себя оптимизацию алгоритмов машинного обучения с целью высокоскоростной обработки информации из видеопотока с учетом аппаратного обеспечения и окончательной оптимизированной реализации для платформы ПЛИС. Реализованная в системе комплексного автоматического машинного обучения на ПЛИС (Holistic Auto Machine Learning for FPGAs – HALF), она объединяет алгоритм эволюционного поиска, различные этапы оптимизации и библиотеку параметризуемых аппаратных модулей обработки информации из видеопотока. HALF автоматизирует как процесс поиска, так и реализацию оптимизированных решений на целевой платформе ПЛИС для различных приложений. Благодаря программированию ПЛИС для выполнения определенного алгоритма машинного обучения производительность значительно повышается по сравнению с обычным графическим процессором.

Кроме того, ПЛИС обладают высокой энергоэффективностью и низкой задержкой, благодаря чему ПЛИС являются более востребованной технологией для многочисленных приложений машинного обучения. По сути, ПЛИС представляют собой многообещающую платформу для улучшения операций машинного обучения благодаря их высокому параллелизму, низкой задержке и возможности настраивать аппаратную реализацию для конкретных задач [2; 3].

Конструкции на основе ПЛИС обеспечивают баланс, гибкость, простоту конфигурирования, а также быструю и энергоэффективную обработку.

Применение ПЛИС для реализации алгоритмов машинного обучения с целью высокоскоростной обработки информации из видеопотока

ПЛИС являются привлекательным вариантом для обработки и сжатия видео, поскольку они обеспечивают ресурсами, необходимыми для реализации инновационных алгоритмов обработки видео. Кроме того, ПЛИС предоставляют гибкое решение, которое сокращает время вывода на рынок и обеспечивает непрерывное обновление и внедрение новых функций в течение всего срока службы решения.

Спрос на транскодирование видео резко возрос, чтобы сделать потоковую передачу быстрой и эффективной. Большинство существующих предложений используют программный подход, который не может соответствовать требованиям обработки потокового видео с высокой пропускной способностью и качеством вещания. Поставщики потокового видео и/или облачных услуг сталкиваются с проблемами низкой пропускной способности, высокого энергопотребления, длительной задержки и большого объема занимаемых ими программных решений [4].

Объем трафика видеоданных будет неуклонно расти из года в год во всех приложениях, включая видео по запросу, прямую трансляцию и видеонаблюдение. Рост числа приложений для потоковой передачи видео, таких как Netflix и YouTube, стимулирует спрос на перекодирование видео. Наиболее заметным различием между традиционным вещанием и потоковым видео является объем контента и количество каналов. Для поддержки широкого спектра принимающих устройств, начиная от ПК и заканчивая смартфонами, контент должен быть перекодирован в различные разрешения и форматы сжатия «на лету». В результате потоковая передача видео потребляет огромное количество вычислительных ресурсов.

Обработка видеопотока обычно включает в себя операции либо с синхронизацией видеосигнала, либо с необработанными растровыми данными отдельных кадров или полей.

Архитектура ПЛИС хорошо подходит для обработки видео по следующим причинам.

1. Генерация синхронизации видео относительно проста с помощью ПЛИС. Даже логическая структура недорогих семейств ПЛИС обычно способна поддерживать компоненты интеллектуальной собственности с частотой более 150 МГц, что позволяет генерировать высокую четкость разрешения (далее – HD-разрешения).

2. При необработанных данных кадра можно использовать преимущества защищенных блоков цифрового процессорного блока, чтобы снизить требования к синхронизации для самой логической структуры. Вместе с конвейерной обработкой отдельных операций алгоритма это позволяет проектировать сложные пути обработки видео даже с разрешением HD [5].

3. Благодаря «близости к металлу» алгоритмы на ПЛИС могут быть более эффективными с точки зрения энергопотребления, чем системы, использующие ядро центрального процессора для выполнения функций обработки.

4. Благодаря гибкости ПЛИС схема обработки видео может быть адаптирована к конкретным требованиям проекта.

5. Гибкость архитектуры ПЛИС может оказаться полезной для небольших производственных серий, где затраты на разработку интегральной схемы специального назначения могут быть непомерно высокими.

В силу своей структуры ПЛИС используются для обработки изображений в реальном времени, поскольку несколько операций могут выполняться одновременно. В интеллектуальной камере со встроенными ПЛИС будет обрабатываться большая часть мультимедиа, в то время как датчик будет передавать пакеты данных. В этом случае камера выдает поток обработанных выходных данных, а не серию изображений.

ПЛИС позволяют перекодировать видео 8K в режиме реального времени без задержек.

В таких областях, как обработка видео и машинное обучение, ПЛИС могут обеспечить экстремальное аппаратное ускорение. Например, известно, что Blackmagic Design использует Xilinx ПЛИС при разработке. На Рисунке ниже можно видеть ПЛИС Xilinx Spartan-6, питающую экран управления камерой.

FPGA[1](Field Programmable Gate Array) – это безопасный способ работы с данными. Хотя существует вероятность того, что злоумышленник сможет получить контроль над FPGA, он не сможет контролировать систему. Поэтому в последние годы ПЛИС часто используются в области безопасности, обработки сигналов и машинного обучения.

Более того, в современной научно-исследовательской литературе в области нейронных сетей обнаруживаются доказательства, что нейронная сеть хорошо справляется с множеством задач, связанных с классификацией и обработкой видеоданных, причем даже лучше человека. Большинство современных архитектур имеют в своем составе сверточные (convolution) блоки. Такие сети называются Convolutional Neural Nets (далее – CNN). CNN – это тип машинного обучения, который анализирует изображение таким образом, чтобы изучать функции, которые могут помочь компьютеру идентифицировать шаблоны в изображении [6].

Этот процесс использования машинного обучения для глубокого обучения чрезвычайно интенсивен. Так, добавив процессоры и протестировав FPGA для обработки алго-

[1] ПЛИС и FPGA – это аббревиатуры, обозначающие один и тот же класс электронных компонентов, микросхем.

ритмов, можно улучшить результат производительности почти в два раза. Ускоритель на основе ПЛИС FPGA был разработан для эффективного расчета прямого распространения сверточных слоев. Таким образом, ускоритель машинного обучения, а именно CNN, должен иметь возможность принимать входное изображение и обрабатывать несколько сверточных слоев подряд [7]. При разработке архитектура должна включать несколько факторов. Поскольку система должна обрабатывать несколько уровней, вычислительный движок системы должен быть настроен для поддержки этих слоев. Управление памятью имеет решающее значение, поэтому проект должен включать эффективную схему буферизации данных и сеть повторного распространения на кристалле. Наконец, архитектура проекта должна иметь возможность содержать пространственно распределенный массив элементов обработки, которые можно легко масштабировать до тысяч единиц. Это позволяет ускорителю CNN принять входное изображение, а затем выполнить анализ многочисленных сверточных слоев подряд. Метод, в котором система обрабатывает сверточные слои, очень зависит от используемого оборудования. ПЛИС FPGA стали явным выбором для достижения большей эффективности обработки [8].

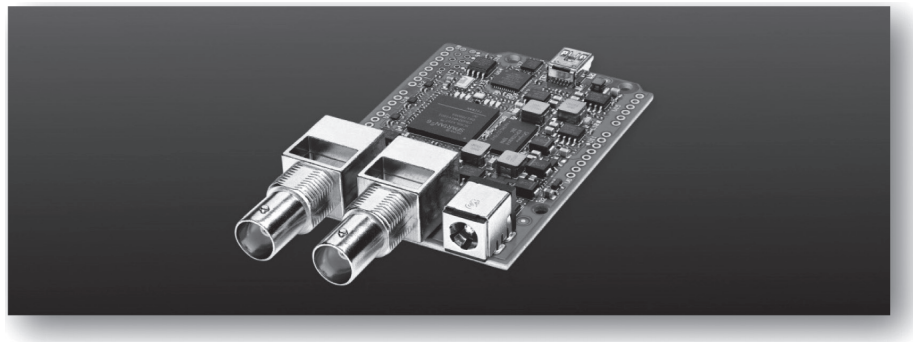

Рисунок. Blackmagic 3G-SDI Shield для Arduino
Источник: Blackmagic Design. URL: https://www.blackmagicdesign.com/developer/product/Arduino (дата обращения: 18.01.2024).

Так, выполнение более быстрых и эффективных вычислений позволяет разрабатывать передовые приложения. Глубокое обучение и другие приложения, интенсивно использующие вычислительные ресурсы, находятся на переднем крае исследований. Поскольку графические процессоры смогли удовлетворить спрос, они зачастую используются в центрах обработки данных. Но с улучшением производительности ПЛИС FPGA центры обработки данных могут использовать их для удовлетворения потребностей в вычислительных возможностях и энергопотреблении. Это позволяет эмулировать глубокое обучение путем последовательной обработки нескольких слоев сверточных нейронных сетей.

Следует согласиться с мнением Д.М. Чумакова и С.В. Козлова, что чип FPGA может обеспечивать программируемое программное обеспечение и производительность обработки профессионального оборудования [9]. В видеоприложениях микросхемы FPGA могут обеспечивать в 10…100 раз бóльшую вычислительную мощность на платформе персонального компьютера, и для изменения функции системы нужно лишь повторно загрузить файл, чтобы изменить конфигурацию оборудования пользователя.

Хотя технология FPGA уже давно используется в индустрии вещательного видео, до сих пор большинство этих приложений ограничены конкретными функциями на профес-

сиональных платформах, и лишь некоторые из них применяются для определенных функций на платах PCI в персональных компьютерах.

Тем не менее вышесказанное позволяет заключить, что вычислительная мощность ПЛИС в последние годы стремительно развивалась, эволюционировав до такой степени, что микросхема может выполнять функции платы или функции системы (SoC). Поскольку на плате PCI можно установить несколько микросхем FPGA, можно разработать полную систему, эквивалентную профессиональному устройству на плате.

Более того, в современной научно-исследовательской литературе обнаруживаются доказательства того, в рамках существующей технологии стандартные функциональные модули обработки видео могут быть загружены в новый тип высокопроизводительной микросхемы FPGA. Эти модули похожи на несколько функциональных модулей в профессиональном устройстве. Другими словами, теперь одна микросхема может содержать всю систему без необходимости использования профессионального устройства, как раньше.

Анализ современного педагогического опыта, публикуемого на различных профильных форумах, сайтах образовательных организаций, а также в научных журналах, позволяет сделать вывод, что по сравнению с разработкой системы профессионального оборудования этот метод разработки имеет следующие преимущества:

- значительно меньшая стоимость разработки;
- значительно более низкая стоимость системы;
- чрезвычайно удобный пользовательский дизайн и обновление;
- можно напрямую использовать прикладное программное обеспечение платформы персонального компьютера.

Заключение

Вышесказанное позволяет сделать объективное заключение, что реализация и применение платформы ПЛИС для реализации алгоритмов машинного обучения с целью высокоскоростной обработки информации из видеопотока превосходит другие оптимизированные модели. ПЛИС можно эффективно применять для самого широкого спектра прикладных задач вне зависимости от предъявляемых требований к надежности и производительности. ПЛИС-платформа предоставляет пользователю большую гибкость для быстрой разработки и тестирования сложных алгоритмов. В качестве примера кодеры JPEG или MPEG-4 могут быть реализованы аппаратными ускорителями с использованием предлагаемой архитектуры для извлечения кадров в пользовательскую логику. Более того, динамическая частичная самонастройка может быть использована для очень сложных алгоритмов, требующих больших ресурсов. Это позволит осуществлять временное отображение алгоритмов в отличие от пространственного отображения, которое может оказаться неосуществимым для одной ПЛИС. Требуемая логика может быть значительно сокращена, и на одной платформе может быть реализовано несколько модулей обработки.

Литература

1. *Xu Liu, Hibat Allah Ounifi, Abdelouahed Gherbi, Yves Lemieux, Wubin Li.* A hybrid GPU-FPGA-based computing platform for Machine Learning // Procedia Computer Science. 2018. Vol. 141. Pp. 104–111. DOI: https://dx.doi.org/10.1016/j.procs.2018.10.155

2. *Зотов В.Б., Терехова К.О., Царапов М.Н.* Анализ программ цифровизации в городе Москве // Муниципальная академия. 2020. № 4. С. 8–17. EDN PIFPBW.

3. *Зотов В.Б., Терехова К.О., Царапов М.Н.* Структура цифровых технологий в Москве // Муниципальная академия. 2022. № 2. С. 10–15. EDN WUJRNG. DOI: 10.52176/2304831X_2022_02_10
4. *Sun Y., Li J., Xu X., Shi Y.* Adaptive Multi-Lane Detection Based on Robust Instance Segmentation for Intelligent Vehicles // IEEE Transactions on Intelligent Vehicles. 2023. Vol. 8. No. 1. Pp. 888–899. DOI: 10.1109/TIV.2022.3158750
5. *Harada K., Kanazawa K., Yasunaga M.* FPGA-Based Object Detection for Autonomous Driving System // 2019 International Conference on Field-Programmable Technology (ICFPT), Tianjin, China, 09–13 December 2019. Pp. 465–468. DOI: 10.1109/ICFPT47387.2019.00094
6. *Угрюмов Е.П.* Цифровая схемотехника : Учебное пособие для вузов. Гл. 7. Программируемые логические матрицы, программируемая матричная логика, базовые матричные кристаллы. Изд. 2. СПб. : БХВ-Петербург, 2004. С. 357–389.
7. *Слюсарь В.И.* Разработка схемотехники ЦАР: некоторые результаты. Часть 2 // Первая миля. 2018. № 2 (71). С. 74–78. EDN YWHYJC. DOI: 10.22184/2070-8963.2018.71.2.74.78
8. *Scott C.* GOWIN Semiconductor Releases the First FPGA with Integrated Bluetooth Radio // GOWIN: Programming for the Future. 2019. November 12. URL: https://www.gowinsemi.com/en/about/detail/latest_news/47/ (дата обращения: 18.01.2024).
9. *Чумаков Д.М., Козлов С.В.* Использование ПЛИС в специализированных цифровых видеокамерах // Известия Томского политехнического университета. 2008. Т. 312. № 2. С. 333–335. URL: http://earchive.tpu.ru/handle/11683/2102 (дата обращения: 18.01.2024).

References

1. Xu Liu, Hibat Allah Ounifi, Abdelouahed Gherbi, Yves Lemieux, Wubin Li (2018) A hybrid GPU-FPGA-based computing platform for Machine Learning. *Procedia Computer Science*. Vol. 141. Pp. 104–111. DOI: https://dx.doi.org/10.1016/j.procs.2018.10.155
2. Zotov V.B., Terekhova K.O., Tsarapov M.N. (2020) Analysis of digitalization programs in the city of Moscow. *Municipal Academy*. No. 4. Pp. 8–17. (In Russian).
3. Zotov V.B., Terekhova K.O., Tsarapov M.N. (2022) Management organization structure of digital technologies in Moscow. *Municipal Academy.* No. 2. Pp. 10–15. DOI: 10.52176/2304831X_2022_02_10 (In Russian).
4. Sun Y., Li J., Xu X., Shi Y. (2023) Adaptive Multi-Lane Detection Based on Robust Instance Segmentation for Intelligent Vehicles. In: *IEEE Transactions on Intelligent Vehicles.* Vol. 8. No. 1. Pp. 888–899. DOI: 10.1109/TIV.2022.3158750
5. Harada K., Kanazawa K., Yasunaga M. (2019) FPGA-Based Object Detection for Autonomous Driving System. In: *2019 International Conference on Field-Programmable Technology (ICFPT)*, Tianjin, China, 09–13 December 2019. Pp. 465–468. DOI: 10.1109/ICFPT47387.2019.00094
6. Ugryumov E.P. (2004) *Tsifrovaya skhemotekhnika* [Digital circuitry] : Textbook for universities. Chapter 7. Programmable logic matrices, programmable matrix logic, basic matrix crystals. 2nd edition. St. Petersburg : BVHK-Petersburg Publ. Pp. 357–389. (In Russian).
7. Slyusar V.I. (2018) Development of circuitry of digital antenna arrays: some results. Part 2. *Last mile*. No. 2. Pp. 74–78. DOI: 10.22184/2070-8963.2018.71.2.74.78 (In Russian).
8. Scott C. (2019) GOWIN Semiconductor Releases the First FPGA with Integrated Bluetooth Radio. *GOWIN: Programming for the Future*. November 12. URL: https://www.gowinsemi.com/en/about/detail/latest_news/47/ (accessed 18.01.2024).
9. *Chumakov D.M., Kozlov S.V.* (2008) The use of FPGAs in specialized digital video cameras. *Bulletin of the Tomsk Polytechnic University*. Vol. 312. No. 2. Pp. 333–335. URL: http://earchive.tpu.ru/handle/11683/2102 (accessed 01.2024).