ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ, ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Գրիգորյան Մուշեղ Տարոնի

ՄԻԿՐՈԷԼԵԿՏՐՈՆԱՅԻՆ ՀԱՇՎՈՂԱԿԱՆ ՀԱՄԱԿԱՐԳԵՐՈՒՄ ՆԵՅՐՈՆԱՅԻՆ ՑԱՆՑԻ ԲԱՇԽՎԱԾ ՄՇԱԿՈՒՄԸ ԵՎ ՀԵՏԱԶՈՏՈՒՄԸ

Ե.27.01 «Էլեկտրոնիկա, միկրո և նանոէլեկտրոնիկա» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2024

МИНИСТЕРСТВО ОБРАЗОВАНИЯ, НАУКИ, КУЛЬТУРЫ И СПОРТА РЕСПУБЛИКИ АРМЕНИЯ

НАЦИОНАЛЬНЫЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ АРМЕНИИ

Григорян Мушег Таронович

РАСПРЕДЕЛЕННАЯ ОБРАБОТКА И ИССЛЕДОВАНИЕ НЕЙРОННОЙ СЕТИ В МИКРОЭЛЕКТРОННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук по специальности 05.27.01- “Электроника, микро- и наноэлектроника”

Ереван 2024

Ատենախոսության թեման հաստատվել է Հայաստանի ազգային պոլիտեխնիկական համալսարանում (ՀԱՊՀ)։

Գիտական ղեկավար՝ տ.գ.դ. Վազգեն Շավարշի Մելիքյան

Պաշտոնական ընդդիմախոսներ՝ ֆ-մ.գ.դ. Ֆերդինանտ Վազգենի Գասպարյան
տ.գ.թ. Արման Ստեփանի Տրդատյան

Առաջատար կազմակերպություն՝ ՀՀ ԳԱԱ Ռադիոֆիզիկայի և էլեկտրոնիկայի ինստիտուտ

Ատենախոսության պաշտպանությունը կայանալու է 2024 հուլիսի 29-ին, ժամը 10^{00}-ին, ՀԱՊՀ-ում գործող «Ռադիոտեխնիկայի և էլեկտրոնիկայի» 046 մասնագիտական խորհրդի նիստում (հասցեն՝ 0009, Երևան, Տերյան փ., 105, 17 մասնաշենք)։

Ատենախոսությանը կարելի է ծանոթանալ ՀԱՊՀ–ի գրադարանում։

Սեղմագիրն առաքված է 2024թ. հունիսի 17-ին

046 Մասնագիտական խորհրդի
գիտական քարտուղար, տ.գ.թ. Բենիամին Ֆելիքսի Բադալյան

Тема диссертации утверждена в Национальном политехническом университете Армении (НПУА)

Научный руководитель: д.т.н. Вазген Шаваршович Меликян

Официальные оппоненты: д.ф.-м.н. Фердинант Вазгенович Гаспарян
к.т.н. Арман Степанович Трдатян

Ведущая организация: НАН РА Институт радиофизики и электроники

Защита диссертации состоится 29-го июля 2024 г. в 10^{00} ч. на заседании Специализированного совета 046 – “Радиотехники и электроники”, действующего при Национальном политехническом университете Армении, по адресу: 0009, г. Ереван, ул. Теряна, 105, корпус 17.

С диссертацией можно ознакомиться в библиотеке НПУА.

Автореферат разослан 17-го июня 2024 г.

Ученый секретарь
специализированного совета 046, к.т.н. Бениамин Феликсович Бадалян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Развитие аппаратного обеспечения, предназначенного для нейронных сетей (НС), стало значительно продвигаться, что позволяет удовлетворить растущий спрос на глубокое обучение. Аппаратное обеспечение включает в себя не только традиционное оборудование, такое как центральные и графические процессоры (ГП), но и специализированное оборудование, такие как массивы программируемых логических вентилей и специализированные интегральные схемы (ИС).

Центральные процессоры (ЦП) являются универсальными и могут запускать сложные операционные системы и программы. Последние модели улучшили возможности параллельной обработки, что помогает произвадить операции с НС. Однако ЦП по-прежнему отстают от другого оборудования в производительности, необходимой для операций с крупными нейронными сетями. Например, процессоры Intel Xeon и AMD Ryzen были улучшены для более высоких параллельных операций, но они все же не сравнимы с эффективностью ГП для целей глубокого обучения.

В настоящее время ГП являются наиболее эффективным оборудованием для глубокого обучения благодаря их высокой параллельности, что идеально подходит для матричных и векторных операций в НС. Модели, такие как серии NVIDIA Tesla и RTX, широко используются в промышленности. Несмотря на свои преимущества, ГП потребляют много энергии и имеют высокую стоимость, что может быть препятствием для небольших организаций.

Массивы программируемой логики предлагают уникальную гибкость, так как они могут быть перепрограммированы после производства, чтобы соответствовать различным требованиям. Это делает их подходящими для реализации на оборудовании и адаптации к новым алгоритмам машинного обучения. Хотя они более гибкие, чем специализированные ИС, однако обычно медленнее и менее энергоэффективнее. Xilinx и подразделение Intel Altera являются примерами в этой области.

Специализированные ИС обеспечивают несравненную эффективность для конкретных задач, поскольку они разработаны для выполнения определенных операций. Тензорные процессоры обработки от Google, разработанные для оптимизации операций НС, являются хорошим примером высокой производительности и эффективности. Однако их фиксированная архитектура и высокие расходы на обработку делают их менее подходящими для общих исследований и приложений, подвергающихся частым изменениям.

Каждый тип оборудования имеет свои сильные и слабые стороны. Развитие оборудования для НС сосредоточено на балансе между эффективностью, энергосбережением и ценовой эффективностью, чтобы соответствовать быстрому прогрессу в технологиях искусственного интеллекта (ИИ) и машинного обучения. Диссертация посвящена фундаментальным вопросам распределенной обработки НС в микроэлектронных вычислительных системах.

Объект исследования. Основные факторы, обусловливающие распределенную обработку НС в микроэлектронных вычислительных системах, и разработка средств повышения их производительности..

Цель работы. Разработка методов, средств и процедур реализации НС в микроэлектронных вычислительных системах с минимальным снижением точности распознавания.

Методы исследования. В ходе исследования были использованы современные подходы к оценке, моделированию и оптимизации распределенной обработки нейронных сетей в микроэлектронных вычислительных системах, а также методы разработки программного обеспечения (ПО).

Научная новизна:

- Предложены подходы к созданию систем распределенной обработки нейронных сетей посредством программируемых вентильных матриц, благодаря которым за счет снижения вычислительной сложности нейронных сетей повышается эффективность распознавания на программируемых вентильных матриц в соответствии с современными требованиями.
- Спроектирован аппаратный ускоритель для распределенной обработки нейронных сетей, который в результате ускорения распознавания обеспечил соответственно 30- и 2,6-кратное ускорение времени распознавания, а также в 3 раза меньшую ошибку локализации по сравнению с центральным и графическим процессорами за счет роста погрешности пространственного смещения на 0,06% для сверточной нейронной сети визуальной локализации.
- Разработана процедура, которая благодаря высокому параллелизму позволила одновременно выполнять разработку четырех и более моделей нейронных сетей при увеличении требуемых ресурсов на 24,375% и необходимой мощности на 24,96%.
- Предложен метод благодаря которому путем снижения вычислительной сложности нейронной сети, квантования весов и активаций обеспечил уменьшение необходимой мощности для работы ускорителя на 21,23% и улучшение скорости распознавания на 32% за счет роста погрешности точности последнего на 0,9%.
- Создана процедура реализации нейронной сети с помощью интегральной схемы, ориентированной на ее применение, которая благодаря характерным для платформы возможностям обеспечила 16,9-кратное уменьшение энергопотребления за счет потери перепрограммируемости.

Практическая ценность работы. В диссертации способы проектирования ускорителей нейронных сетей в разработанных микроэлектронных вычислительных системах были реализованы в программном средстве "Hardware Accelerator Design Tool", которое было внедрено в ООО "ЭНДЖИН" и позволило сократить время проектирования и проверки ускорителей в 3...4 раза. Реализация предложенных методов с помощью программного инструмента "Hardware Accelerator Design Tool" позволила обеспечить 2,6-кратное ускорение времени распознавания за счет роста погрешности пространственного смещения на 0,1%.

На защиту выносятся:

- аппаратный ускоритель для распределенной обработки нейронных сетей;
- процедура распределенной обработки нейронных сетей;

- метод снижения вычислительной сложности нейронной сети;
- процедура реализации нейронной сети с помощью интегральной схемы, ориентированной на ее применение.

Достоверность научных положений. Научные положения были подтверждены экспериментальными результатами моделирования и математическими обоснованиями, представленными в диссертации.

Внедрение. Разработанный программный инструмент "Hardware Accelerator Design Tool" был внедрен в ООО " ЭНДЖИН ". Он используется при проектировании ускорителей нейронных сетей с целью повышения их производительности.

Апробация работы. Основные научные и практические результаты диссертации докладывались на:

- 19-й Международной конференции "East-West Design & Test Symposium (EWDTS)" (Батуми, Грузия, 2023 г.);
- научных семинарах кафедры "Микроэлектронные схемы и системы" НПУА (Ереван, Армения, 2021 - 2024 гг.);
- научных семинарах ЗАО "Синопсис Армения" (Ереван, Армения, 2021 - 2024 гг.).

Публикации. Основные положения диссертации представлены в пяти научных работах, список которых приведен в конце автореферата.

Структура и объём диссертации. Работа состоит из введения, трех глав, основных выводов, списка литературы из 137 наименований и четырех приложений. В первом приложении представлен акт внедрения диссертации, во втором - фрагменты аппаратного описания моделей нейронных сетей, в третьем - фрагменты описания программного средства "Hardware Accelerator Design Tool", а в четвертом - списки использованных рисунков, таблиц и сокращений. Основной объем диссертации составляет 113 страниц, а вместе с приложениями - 142 страницу, включая 89 рисунков и 17 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цель и основные задачи исследования, представлены разработанные методы, научная новизна, практическое значение и основные научные положения, выносимые на защиту.

В первой главе представлены основные типы аппаратного обеспечения для выполнения вычислений, характерных для НС. Обсуждены типы ускорительей НС, их архитектуры и особенности. Изучена эффективность обработки этих средств.

Ускорение для НС включает использование специализированного аппаратного обеспечения для ускорения разработки алгоритмов машинного обучения. Эти ускорители спроектированы для выполнения массивно-параллельных вычислений, характерных для НС, делая их значительно более быстрыми и эффективными по сравнению с процессорами общего назначения. Наиболее распространенные методы аппаратного ускорения следующие:

Графические процессоры представляют собой многоядерные параллельные процессоры, способные одновременно выполнять тысячи операций. Принцип работы ГП делает их эффективными для выполнения матричных и векторных

операций, характерных для НС. Их архитектура (рис. 1) позволяет значительно ускорить этапы как обучения, так и распознавания моделей машинного обучения.

Ускорение свёрточных нейронных сетей (СНС) на ГП связано с эффективным использованием имеющихся ядер. Реструктуризация потока выполнения может решить многие проблемы с памятью, возникающие из-за зависимости от данных. Тем не менее, настоящая сила ГП кроется в группировании. Во многих случаях от СНС требуется выполнить распознавание для многочисленных входных данных. Если в данный момент доступно несколько входных данных, их одновременная группировка и обработка даст лучший результат, поскольку это устраняет необходимость перезагрузки весов для каждого входного набора данных.

Рис.1. Архитектура NVIDIA Turing TU102

Программируемые вентильные матрицы (ПВМ) перенастраиваемые кремниевые ИС, которые могут быть перепрограммированы для выполнения специализированных операций. Они предлагают гибкость и эффективность, позволяя адаптироваться к специфическим задачам НС. Это свойство может обеспечить более высокую производительность для определенных приложений по сравнению с процессорами общего назначения. (рис. 2).

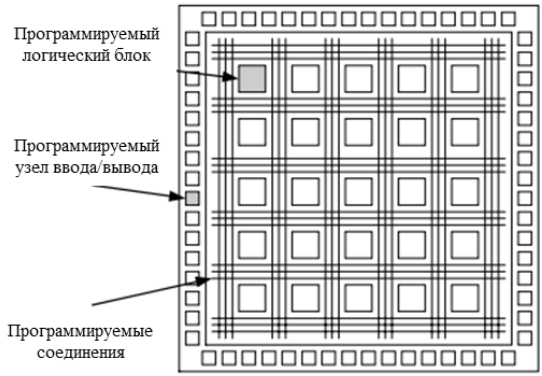


Рис. 2. Архитектура ПВМ

Основным логическим блоком ПВМ является блок таблицы истинности (ТИ) (рис. 3), который содержит: блок памяти (статическое оперативное запоминающее устройство (ОЗУ)) для хранения двоичных данных логической функции, мультиплексор для выбора правильных данных в зависимости от входных значений и ТИ для любой логической функции (И, И-НЕ, И-ИЛИ и т.д.).

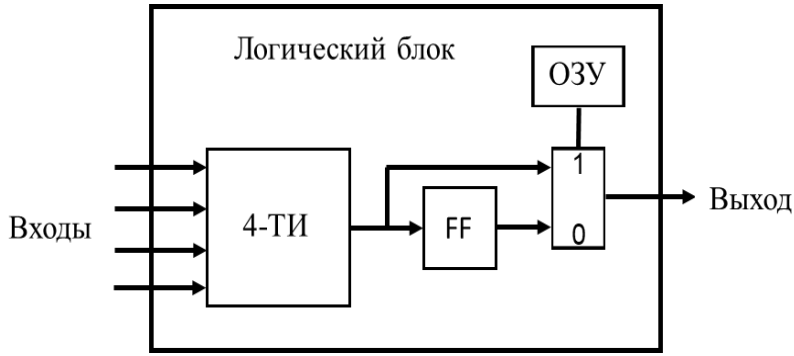
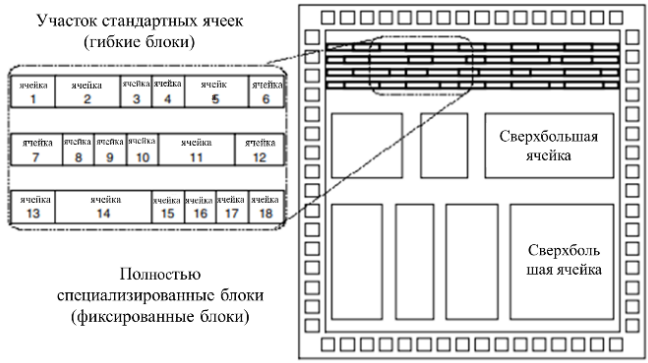


Рис. 3. Архитектура блока таблицы истинности

Интегральные схемы специального назначения (ИССН) специализированные устройства, предназначенные для повышения вычислительной эффективности и производительности алгоритмов ИИ. Эти ИС (рис. 4) адаптированы для специальных задач ИИ, предлагая значительные преимущества с точки зрения скорости обработки и энергопотребления. Они достигают высоких уровней оптимизации, включая архитектуры, явно спроектированные для преобладающих математических операций в вычислениях ИИ, таких как матричное умножение и нелинейные активации.

ИССН на основе массивов транзисторов (рис. 4) состоят из массивов транзисторов p и n типов, причем они могут быть соединены каналом (рис. 4 а) и без канала (рис. 4 а).

а)



б)

Рис. 4. Архитектуры ИССН на основе массивов вентильей (а) и стандартных ячеек (б)

В массивах транзисторов без каналов соединения обеспечиваются с помощью верхних металлических слоев. ИССН на основе стандартных ячеек (рис. 4 б) состоят из логических элементов (вентили, умножители, сумматоры, триггеры), которые проектируются и хранятся в виде библиотек.

Как видно из сравнения эффективности различных аппаратных ускорителей с учетом энергопотребления и производительности (рис. 5), ПВМ и ИССН-ускорители более эффективны, причем ИССН "Synopsys EV6" демонстрирует наивысшую эффективность с результатом 4,5 тера умножение сложение аккумуляцыя (TУСА).

ПВМ могут обеспечить энергоэффективное ускорение благодаря своей настраиваемости для конкретной модели и использованию только необходимых ресурсов. Они также способны динамически конфигурироваться для работы с любой разрядностью данных, в то время как ГП не могут приспосабливаться для обработки данных переменных размеров, будучи в основном пригодными для 32- или 64-разрядных операций с плавающей запятой.

Ускорители на базе ИССН превосходят другие платформы по энергоэффективности, отчасти благодаря способу соединения обрабатывающих элементов. Вероятно, единственным недостатком ИССН является их негибкость. После изготовления повторное использование элементов нежизнеспособно. Однако для нейросетевых ускорителей это не проблема. Проектирование общего нейросетевого ускорителя в ИССН может обрабатывать данные любого вида, просто загружая в память правильную архитектуру сети и веса.

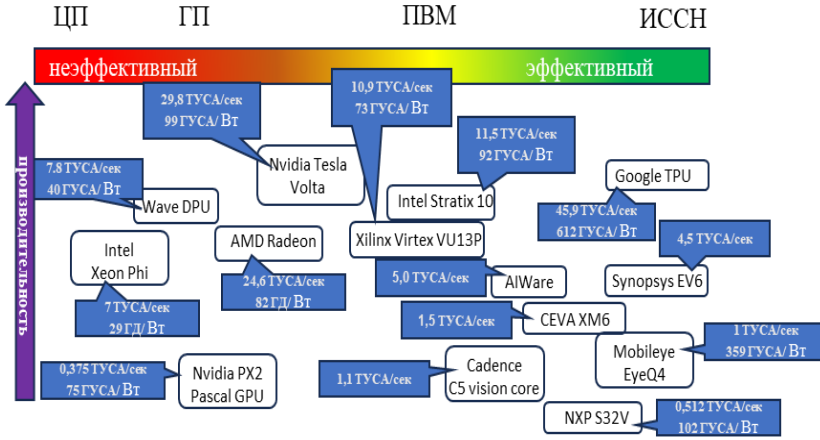


Рис. 5. Сравнение эффективности различных аппаратных ускорутелов

Таким образом, в микроэлектронных вычислительных системах распределенная обработка НС с помощью центральных и графических процессоров сопряжена с рядом ограничений - высоким энергопотреблением, длительным временем отклика, что делает эффективность обработки несоответствующей текущим практическим требованиям. По этой причине обработка подобных сетей с помощью ПВМ и ИССН в настоящее время является чрезвычайно актуальной задачей.

Во второй главе представлены разработанные методы и даются решения проблем, описанных в первой главе.

Метод распределенной обработки нейронной сети в микроэлектронных вычислительных системах с реализацией модели визуальной локализации на спроектированном процессоре глубокого обучения.

Для сравнительного анализа с традиционными процессорами был разработан ускоритель, или, другими словами, процессор глубокого обучения (ПГО) с использованием аппаратного обеспечения ПВМ, с помощью которого был выполнен сравнительный анализ модели нейронной сети для визуальной локализации.

Предлагаемый ПГО позволяет достичь высокого параллелизма и энергоэффективности, что делает его эффективным средством для реализации приложений СНС. Он не предназначен для какой-либо конкретной модели или архитектуры СНС. Этой эффективности удалось достичь благодаря специализированному набору команд, обеспечивающих работу ПГО. Для

реализации моделей СНС на ПГО необходимо привести их к специальному формату, такому как Tensorflow, Caffe, Pytorch и др.

Внутренняя архитектура разработанного ускорителя (рис. 6) в основном состоит из тактового генератора, ОЗУ, блоков команд и глобальной памяти. Процессор имеет архитектуру ARM. Он выполняет прерывания в ПГО и из него, а также передачу данных.

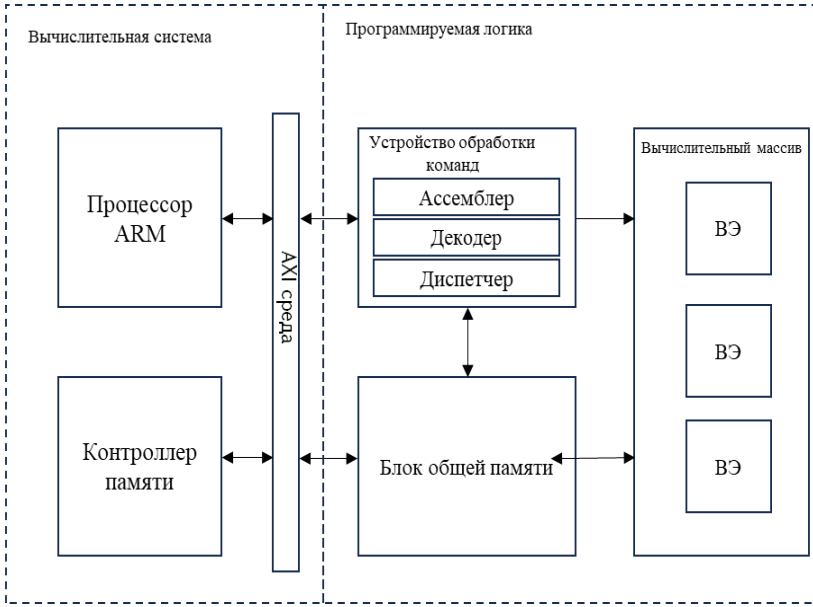


Рис.6. Внутренняя архитектура ПГО

Блок обработки команд выполняет чтение и исполнение команд, связанных с различными операциями ускоренной СНС. Основная роль ассемблера заключается в извлечении команд, связанных с моделью, из памяти ПГО. Затем декодер с использованием арифметического логического устройства (АЛУ) производит декодирование команд. Диспетчер отвечает за управление передачей данных/инструкций между АЛУ и памятью. Блок общей памяти действует в качестве буфера для входных и выходных данных, а также для промежуточного вывода из ПГО, что обеспечивает высокую пропускную способность.

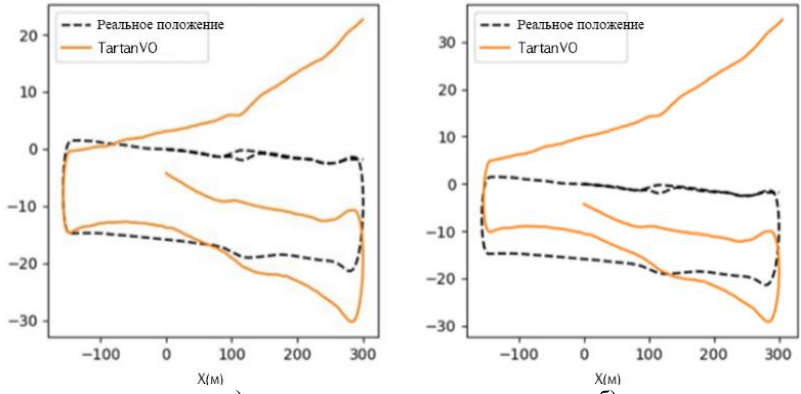
ПГО-ускоритель был реализован на аппаратном обеспечении ПВМ Xilinx Zynq UltraScale+ MPSoC ZCU104.

Далее было проведено развертывание модели визуальной одометрии (ВО) на разработанном ускорителе. В качестве модели ВО использовалась НС TartanVO. Это алгоритм ВО, который был обучен на наборе данных TartanAir.

Вывод алгоритма TartanVO был протестирован с использованием ЦП, ГП и разработанного аппаратного обеспечения ПВМ. Было проведено сравнение, как

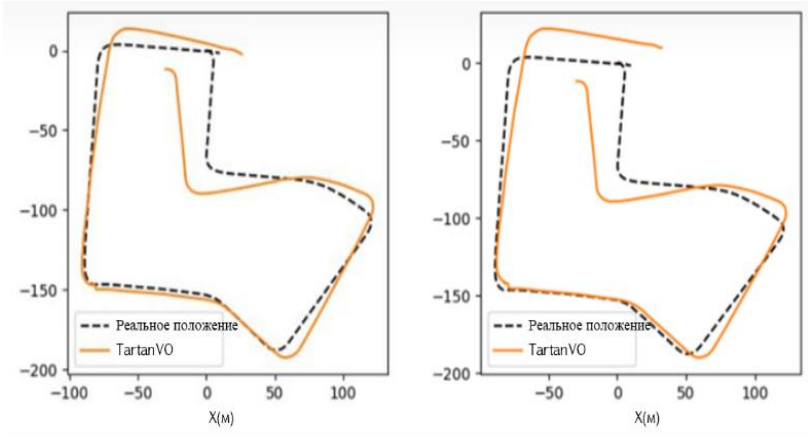
ПВМ помогает достичь аппаратного ускорения, а также скорости и точности вывода на наборе данных KITTI.

Чтобы оценить точность определения положения и результаты реализации на ПВМ, вывод TartanVO был протестирован на последовательностях 06 (рис. 7) , 07 (рис. 8) набора данных KITTI. Сначала было проведено полное выведение модели на ЦП и ГП. Затем вывод был также выполнен на разработанном ПГО. Полученные траектории были сравнены с реальными положениями.



а) б)

Рис. 7. Визуализация полной модели на ГП (а) и на ПГО (б) в случае последовательности 06 набора данных KITTI



а) б)

Рис. 8. Визуализация полной модели на ГП (а) и на ПГО (б) в случае последовательности 07 набора данных KITTI

Производительность моделей ВО оценивается средним квадратичным отклонением пространственного смещения и средним квадратичным отклонением изменения угловой ошибки (табл. 1).

Таблица 1

Результаты тестирования модели ВО

Последовательность	06	07
Средняя квадратичная ошибка позиционирования (%) (ЦП/ГП)	4,72	4,32
Средняя квадратичная ошибка вращения (%) (ЦП/ГП)	2,95	3,41
Средняя квадратичная ошибка позиционирования (%) (ПГО)	4,75	4,35
Средняя квадратичная ошибка вращения (%) (ПГО)	2,95	3,41

Также было проведено сравнение общей производительности сети путем вычисления времени вывода данных (табл. 2) для каждого аппаратного обеспечения. Результаты показывают что, по сравнению с ЦП и ГП, ПГО обеспечивает ускорение времени вывода соответственно в 30 и 2,6 раза.

Таблица 2

Результаты продолжительности тестирования модели

	Длительность работы (*с*)	
Последовательность	06	07
ЦП (AMD RYZEN 7)	683	662
ГП (RTX 2060)	56	54
ПГО (ZCU104)	20	20

Процедура метода распределенной обработки нейронной сети в микроэлектронных вычислительных системах.

Для точного и систематического выполнения процессов преобразования модели НС и развертывания на ПГО была разработана процедура которая состоит из следующих этапов:

- обучение модели НС, для чего требуются наборы данных для обучения и тестирования;
- обработка обученной модели для реализации на ускорителе, для чего требуется набор данных для валидации;
- уменьшение вычислительной сложности НС;

- генерация формата, специфичного для ускорителя;
- выбор аппаратного обеспечения и синтез ускорителя;
- оценка требуемых ресурсов для синтезированного ускорителя;
- реализация модели НС на ускорителе;
- вывод полученной модели;
- исследование и сравнение полученных результатов.

Если уже имеется предварительно обученная модель НС, то этап обучения модели (фрагмент в рамке на рис. 9) можно пропустить.

Для проверки эффективности разработанной процедуры была выполнена реализация модели распознавания объектов. Модели глубокого обучения достигли значительного прогресса в задачах компьютерного зрения, в частности в обнаружении объектов в реальном времени. Архитектура YOLO позволяет одновременно предсказывать классы объектов и координаты ограничивающих рамок за один проход, а также реализовывать такие приложения, как автономное вождение, наблюдение и интерактивная робототехника.

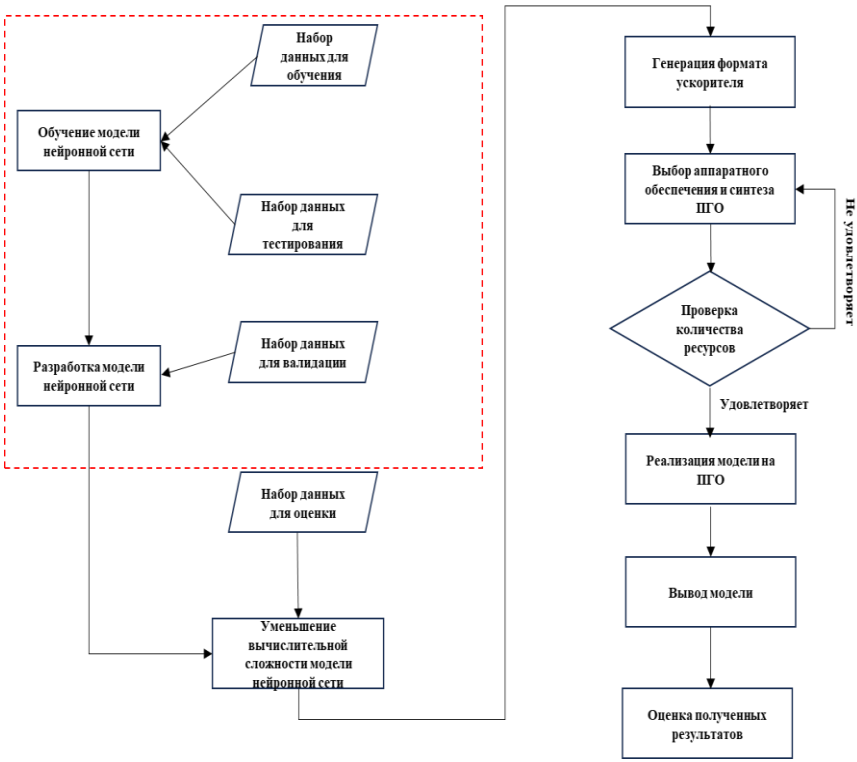


Рис. 9. Процедура обработки НС и реализации на аппаратном обеспечении

Выполняется синтез модели для получения предварительного проекта ПГО и количества требуемых ресурсов (рис. 10). После оценки требуемого количества ресурсов реализуется проект на аппаратном обеспечении ПВМ (рис. 10). Для реализации данной задачи была выбрана платформа ПВМ ZCU102.

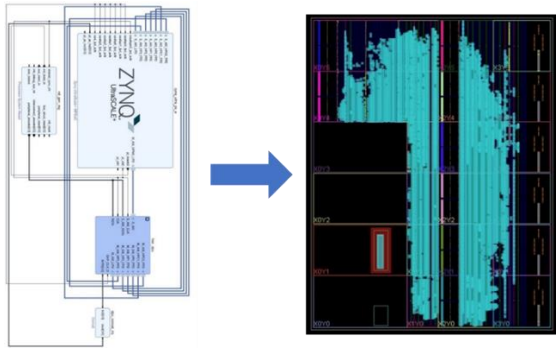


Рис. 10. Синтез и реализация модели на ПВМ

Точность моделей обнаружения объектов оценивалась с использованием стандартных эталонных наборов данных, включая COCO. Модели оценивались по способности точно обнаруживать объекты различных классов и размеров. Один из общих способов проверки - определение средней усредненной точности (mAP) (1)

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \,, \qquad (1)$$

где N - количество классов, а AP_i - средняя точность для i-го класса.

Для обнаружения объектов моделью (табл. 3) были выбраны изображения, содержащие человека, автомобиль, велосипед, собаку, стул. Изображения были взяты из набора данных COCO. Разработанная процедура позволяет выполнять обнаружение, используя сразу несколько моделей YOLO, в данном случае YOLOv3, YOLOv4, YOLOv5 и YOLOv6.

Таблица 3

Средняя усредненная точность для ряда классов модели обнаружения объектов, развернутой на спроектированном ПГО - *ускорителе с помощью разработанной процедуры*

Класс	yolov3	yolov4	yolov5	yolov6
Среднее качество (%)	75,59	76,66	71,97	76,81
Человек (%)	91,78	89,63	81,84	87,86
Автомобиль (%)	75,26	77,84	71,37	74,51
Велосипед (%)	77,42	77,84	71,37	74,51
Собака (%)	86,52	87,69	77,99	86,64
Стул (%)	68,43	69,73	64,94	69,67

Разработка метода снижения вычислительной сложности нейронной сети в микроэлектронных вычислительных системах.

Для снижения вычислительной сложности НС предлагается метод квантования ее модели (рис. 11).

Во время обучения НС обычно используются 32-битные веса и значения активации с плавающей запятой. Благодаря квантованию можно снизить вычислительную сложность без потери точности прогнозирования, заменив 32-битные веса и активации с плавающей запятой на 8-битный целочисленный формат (INT8).

Рис. 11. Квантование НС

Внедрение модели сети с фиксированной запятой требует меньше памяти, обеспечивая более высокую производительность и энергоэффективность по сравнению с моделью с плавающей запятой.

Для оценки эффективности метода квантования НС сначала проводится реализация полной модели. Реализация осуществляется на проектированном ПГО. Затем выполняется снижение вычислительной сложности модели НС предложенным методом квантования.

Процедура метода квантования (рис. 12) состоит из следующих этапов:

1. Разработка полной 32-битной модели с плавающей запятой.
2. Квантование весов и масштабирование значений активации с использованием специального набора данных.
3. Генерация формата соответствующей модели для ускорителя.
4. Проектирование ускорителя в соответствии с разработанной моделью.

Рис. 12 Процедура метода квантования

В качестве модели использовалась НС для распознавания рукописных цифр MNIST.

Было проведено реализация сети на ПГО. Для проверки результатов в качестве метрик были приняты скорость вывода, точность выходного результата и требуемая мощность для ПГО (табл. 4).

Таблица 4

Результаты до и после внедрения метода квантования НС

Аппаратное обеспечение	ПГО (модель до квантования)	ПГО (модель после квантования)
Скорость вывода (с)	2,2	1,5
Точность вывода (%)	98,9	98
Требуемая мощность (Вт)	13,59	10,70

В третьей главе представлено разработанное программное средство “HADT”, которое дает возможность реализовать предложенные решения, выполнить моделирование, проанализировать полученные результаты. Программное обеспечение снижает вероятность ошибки, вносимой дизайнерами.

ПО в начале работы позволяет задать рабочие условия, вводя модель НС, а также то аппаратное обеспечение (ПВМ или ИССН), на котором должен быть реализован проект (рис. 13).

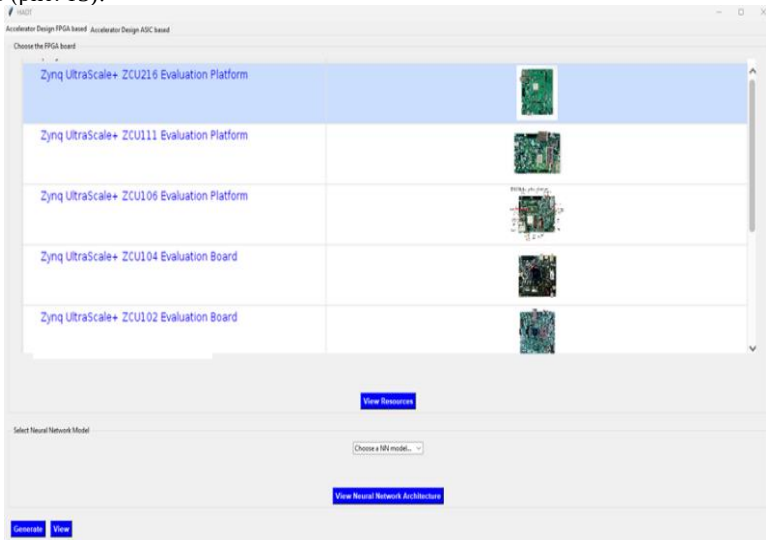


Рис. 13. Главное окно ПО

Результаты проектирования с помощью ПО представлены в табл. 5 и 6.

Таблица 5

Оценка эффективности метода квантования на СНС MNIST с помощью программного обеспечения

Метод	Скорость вывода (с)	Точность вывода (%)
Существующий	2,2	98,9
Предлагаемый	1,5	98
С помощью ПО	1,51	97,98

Таблица 6

Оценка эффективности проектирования СНС YOLOv3 с помощью программного обеспечения

Реализация	Использование ресурсов (%)	Мощность (Вт)
Предлагаемый метод (ПВМ)	50,93	13,81
Спроектированный с помощью ПО (ПВМ)	51,05	14,25

Наибольшее расхождение между результатами оптимизации, полученными с помощью ПО, и результатами, описанными во второй главе, не превышает 0,02%. Таким образом, можно заключить, что данное программное решение при существенной экономии времени проектирования и приемлемой потере точности применимо для задач проектирования современных ИС.

ОСНОВНЫЕ ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. Предложены подходы к созданию систем распределенной обработки нейронных сетей посредством программируемых вентильных матриц, благодаря которым за счет снижения вычислительной сложности нейронных сетей повышается эффективность распознавания на программируемых вентильных матриц в соответствии с современными требованиями.
2. Спроектирован аппаратный ускоритель для распределенной обработки нейронных сетей, который в результате ускорения распознавания обеспечил соответственно 30- и 2,6-кратное ускорение времени распознавания, а также в 3 раза меньшую ошибку локализации по сравнению с центральным и графическим процессорами за счет роста погрешности пространственного смещения на 0,06% для сверточной нейронной сети визуальной локализации.
3. Разработана процедура, которая благодаря высокому параллелизму позволила одновременно выполнять разработку четырех и более моделей нейронных сетей при увеличении требуемых ресурсов на 24,375% и необходимой мощности на 24,96%.
4. Предложен метод благодаря которому путем снижения вычислительной

сложности нейронной сети, квантования весов и активаций обеспечил уменьшение необходимой мощности для работы ускорителя на 21,23% и улучшение скорости распознавания на 32% за счет роста погрешности точности последнего на 0,9%.

5. Создана процедура реализации нейронной сети с помощью интегральной схемы, ориентированной на ее применение, которая благодаря характерным для платформы возможностям обеспечила 16,9-кратное уменьшение энергопотребления за счет потери перепрограммируемости.
6. В диссертации способы проектирования ускорителей нейронных сетей в разработанных микроэлектронных вычислительных системах были реализованы в программном средстве "Hardware Accelerator Design Tool", которое было внедрено в ООО "ЭНДЖИН" и позволило сократить время проектирования и проверки ускорителей в 3...4 раза. Реализация предложенных методов с помощью программного инструмента "Hardware Accelerator Design Tool" позволила обеспечить 2,6-кратное ускорение времени распознавания за счет роста погрешности пространственного смещения на 0,06%

Основные результаты диссертации опубликованы в следующих работах:

1. **Melikyan V., Grigoryan M., Avetisyan A. and Khachatryan T.** Accelerating CNN models for visual odometry: Design and FPGA Implementation for Efficient Hardware Acceleration // 2023 IEEE East-West Design & Test Symposium (EWDTS).- Batumi, Georgia, 2023.- p. 1-5, doi: 10.1109/EWDTS59469.2023.10297049.

2. **Grigoryan M. T.** Investigating the performance indices of YOLO models implemented on a DPU: A Comparative Analysis // Proceedings of the RA NAS and NPUA. Series of Technical Sciences ISSN:0002-306X. – 2023. - Vol. 76, № 3. - P. 343-350.

3. Power supply noise rejection improvement method in modern VLSI Design / **Melikyan V., Mkhitaryan A., Grigoryan H., Grigoryan M.,** et al // 2019 IEEE East-West Design & Test Symposium (EWDTS). - 2019. - P. 1-4, doi: 10.1109/EWDTS.2019.8884372.

4. **Avetisyan A.A., Grigoryan M.T., Melikyan A.V.** Benchmarking and implementing deep learning algorithms on FPGA and ASIC platforms // Proceedings of the RA NAS and NPUA. Series of Technical Sciences. ISSN:0002-306X. – 2024. - Vol. 77, № 1. - P. 87-96.

5. **Avetisyan A.A., Khachatryan T.B., Grigoryan M.T.** Photorealistic and synthetic stereo-dataset generation method for visual odometry and depth estimation // Proceedings of the RA NAS and NPUA. Series of Technical Sciences. ISSN:0002-306X. – 2023. - Vol. 76, № 1. - P. 12-21.

ԱՄՓՈՓԱԳԻՐ

Հավելվածների արագ զարգացմամբ և կիրառության ոլորտների շարունակական ընդլայնմամբ արհեստական բանականության և մեքենայական ուսուցման պահանջներն անընդհատ աճում են: Նեյրոնային ցանցերի բարդության աճը նոր մարտահրավերներ է առաջացնում դրանք արդյունավետորեն իրագործելու համար անհրաժեշտ սարքավորումների զարգացման առումով:

Ավանդական սարքային ապահովումները, ինչպիսիք են կենտրոնական և գրաֆիկական պրոցեսորները, դեռևս կարևոր դեր են խաղում այս ոլորտում, սակայն դրանք այլևս չեն կարող բավարարել խոշոր նեյրոնային ցանցերի մշակման ծանրաբեռնվածությունը: Կենտրոնական պրոցեսորները կարող են զուգահեռ կատարել գործողություններ, բայց դրանց հաշվարկային ռեսուրսները սահմանափակ են, իսկ գրաֆիկական պրոցեսորները՝ թանկ և ոչ էներգաարդյունավետ:

Մասնագիտացված սարքավորումները, ինչպիսիք են թեմատիկ ծրագրավորվող փականների զանգվածները և կիրառությանը կողմնորոշված ինտեգրալ սխեմաները, առաջարկում են բարձր կատարողականություն և ճկունություն: Թեմատիկ ծրագրավորվող փականների զանգվածները կարող են վերափոխվել նոր ալգորիթմներին համապատասխանելու համար, իսկ կիրառությանը կողմնորոշված ինտեգրալ սխեմաները նախագծված են կոնկրետ խնդիրների համար՝ առավելագույն արդյունավետություն ապահովելու նպատակով: Սակայն այս մասնագիտացված սարքավորումներն ունեն իրենց սահմանափակումները: Թեմատիկ ծրագրավորվող փականների զանգվածները որոշ խնդիրների համար կարող են ցածր արդյունավետություն ունենալ և ավելի շատ էներգիա սպառել, քան կիրառությանը կողմնորոշված ինտեգրալ սխեմաները, իսկ վերջիններս իրենց ֆիքսված ճարտարապետության պատճառով պակաս կիրառելի են նոր ալգորիթմների պարագայում:

Հետևաբար, նեյրոնային ցանցերի սարքավորումների զարգացումը միտված է արդյունավետության, էներգախնայողության և ճկունության օպտիմալ համադրության որոնմանը՝ համապատասխանելու արհեստական բանականության և մեքենայական ուսուցման բնագավառների արագընթաց զարգացմանը: Լուծումները կարող են գալ հիբրիդ մոտեցումներից կամ կոնկրետ կիրառությունների համար նոր սարքավորումների մշակումից:

Ատենախոսությունը նվիրված է առերեսման ճշտության նվազագույն անկման հաշվին միկրոէլեկտրոնային հաշվողական համակարգերում

նեյրոնային ցանցի իրագործման եղանակների, միջոցների և ընթացակարգերի մշակմանը։

Առաջարվել են թեմատիկ ծրագրավորող փականների զանգվածների միոցով նեյրոնային ցանցի բաշխված մշակման համակարգերի ստեղծման մոտեցումներ, որոնք նեյրոնային ցանցերի հաշվարկային բարդության նվազեցման շնորհիվ բարձրացնում են թեմատիկ ծրագրավորող փականների զանգվածների վրա առերեսման արդյունավետությունը ժամանակակից պահանջներին համապատասխան։

Մշակվել է նեյրոնային ցանցերի բաշխված մշակման համար սարքային արագացուցիչ, որն առերեսման արագացման շնորհիվ ապահովել է առերեսման ժամանակի համապատասխանաբար 30 և 2,6 անգամ արագացում, ինչպես նաև 3 անգամ ավելի քիչ տեղորոշման սխալանք կենտրոնական պրոցեսորի և գրաֆիկական պրոցեսորի հետ համեմատած՝ տեսողական տեղորոշման փաթույթային նեյրոնային ցանցի 0,06% տարածքային շեղման սխալանքի աճի հաշվին։

Մշակվել է ընթացակարգ, որը բարձ զուգահեռության շնորհիվ թույլ է տվել կատարել մի անգամից չորս և ավել նեյրոնային ցանցերի մոդելների մշակում պահանջվող ռեսուրսների 24,375% և անհրաժեշտ հզորության 24,96% աճի հաշվին։

Առաջարկվել է նեյրոնային ցանցի հաշվարկային բարդության նվազեցման եղանակ , որը կշիռների և ակտիվացումների քվանտացման միջոցով ապահովել է արագացուցիչի աշխատանքի համար անհրաժեշտ հզորության 21,23% նվազեցում և առերեսման արագության 32% բարելավում , վերջինիս սխալանքի 0,9% աճի հաշվին։

Ստեղծվել է նեյրոնային ցանցի կիրառությանը կողմնորոշված ինտեգրալ սխեմայի միջոցով իրականացման ընթացակարգ, որը հարթակին բնորոշ հնարավորությունների շնորհիվ ապահովել է էներգասպառման 16,9 պատիկ փոքրացում վերածրագրավորելիության կորստի հաշվին։

Ատենախոսությունում մշակված միկրոէլեկտրոնային հաշվողական համակարգերում նեյրոնային ցանցերի արագացուցիչների նախագծման միջոցներն իրագործվել են “Hardware Accelerator Design Tool” ծրագրային միջոցում, որը ներդրվել է «ԷՆԶԻՆ» ՍՊԸ-ում և թույլ է տվել կրճատել արագացուցիչների նախագծման և ստուգումների ժամանակը 3-4 անգամ։ Առաջարկված մեթոդների իրագործումը՝ “Hardware Accelerator Design Tool” ծրագրային գործիքի միջոցով, թույլ է տվել 0,06% տարածքային շեղման սխալանքի հաշվին, ապահովել առերեսման ժամանակի 2,6 անգամ արագացում։

MUSHEGH TARON GRIGORYAN

Distributed Neural Network Processing and Research in Microelectronic Computing Systems

SUMMARY

With the rapid development of applications and continuous expansion of fields of application, the demands of artificial intelligence and machine learning are constantly growing. The increasing complexity of neural networks poses new challenges in the development of hardware necessary for their efficient implementation.

Traditional hardware such as central processing units (CPUs) and graphics processing units (GPUs) still play an important role in this field, but they can no longer meet the computational demands of large neural networks. CPUs can perform operations in parallel, but their computational resources are limited, while GPUs are expensive and not energy-efficient.

Specialized hardware such as field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) offer high performance and flexibility. FPGAs can be reconfigured to adapt to new algorithms, while ASICs are designed for specific problems to ensure maximum efficiency. However, these specialized hardware have their limitations. FPGAs may have low efficiency and consume more energy than ASICs for certain problems, while ASICs, due to their fixed architecture, are less applicable for new algorithms.

Therefore, the development of neural network hardware aims to find the optimal combination of efficiency, energy efficiency, and flexibility to keep up with the rapid advances in the fields of artificial intelligence and machine learning. Solutions may come from hybrid approaches or the development of new hardware for specific applications.

The dissertation is dedicated to developing methods, means, and procedures for implementing neural networks in microelectronic computing systems with minimal accuracy degradation.

Approaches for creating distributed neural network processing systems using FPGAs have been proposed, which increase the inference efficiency on FPGAs by reducing the computational complexity of neural networks in accordance with modern requirements.

A hardware accelerator for distributed neural network processing has been designed, which, due to inference acceleration, has provided a 30 and 2.6 times speedup in inference time respectively, compared to a CPU and GPU, as well as a 3 times lower localization error at a 0.06% spatial deviation penalty for a visual localization convolutional neural network.

A procedure has been developed that, due to high parallelism, allowed processing of four or more neural network models at once, with a 24.375% increase in required resources and a 24.96% increase in required power.

A method for reducing the computational complexity of neural networks has been proposed, which, through weight and activation quantization, provided a 21.23% reduction in power required for accelerator operation and a 32% improvement in inference speed, at the cost of a 0.9% increase in error.

A procedure for implementing a neural network using an application-specific integrated circuit has been created, which, due to platform-specific capabilities, provided a 16.9-fold reduction in power consumption at the cost of losing reconfigurability.

The microelectronic computing system neural network accelerator design tools developed in the dissertation have been implemented in the "Hardware Accelerator Design Tool" software, which has been introduced at "NGENE" LLC and has allowed reducing the design and verification time of accelerators by 3-4 times. The implementation of the proposed methods using the "Hardware Accelerator Design Tool" software has enabled a 2.6 times acceleration in inference time at the cost of a 0.06% spatial deviation error.