ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ, ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Ավետիսյան Աշոտ Ազատի

ԱՌԵՐԵՍՎՈՂ ԱՐՀԵՍՏԱԿԱՆ ԲԱՆԱԿԱՆՈՒԹՅԱՄԲ ԾՐԱԳՐԱՎՈՐՎՈՂ ՓԱԿԱՆՆԵՐԻ ՄԱՏՐԻՑԻ ՃԱՐՏԱՐԱՊԵՏՈՒԹՅԱՆ 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

Аветисян Ашот Азатович

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АРХИТЕКТУРЫ МАТРИЦЫ ПРОГРАММИРУЕМЫХ ВЕНТИЛЕЙ С ВЫВОДЯЩИМ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

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՝ տ.գ.դ. Սուրիկ Խաչիկի Խուդավերդյան
տ.գ.թ. Տիգրան Արայիկի Հախվերդյան

Առաջատար կազմակերպություն՝ Երևանի Պետական Համալսարան

Ատենախոսության պաշտպանությունը կայանալու է 2024 հուլիսի 29-ին, ժամը 12^{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 2024 թ. 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.т.н. Сурик Хачикович Худавердян
к.т.н. Тигран Араикович Ахвердян

Ведущая организация: Ереванский Государственный Университет

Защита диссертации состоится 29-го июля 2024 года в 12^{00} 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 года.

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

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Внедрение искусственного интеллекта (ИИ) в архитектуры матриц программируемых вентилей (МПВ) в последнее пятилетие привлекает все больше внимания со стороны многих компаний, занимающихся информационными технологиями. Причина кроется в ряде преимуществ таких устройств, в частности, в высокой степени параллелизма и гибкой специализации архитектур.

Специализированные МПВ сегодня внедряются в центрах обработки данных и суперкомпьютерах с целью ускорения работы выводящего ИИ. По данным исследования компании Xilinx, внедрение их платформы Alveo в центрах обработки данных позволило увеличить количество обрабатываемых за секунду изображений в 192 раза по сравнению с ранее используемыми центральными процессорами (ЦП) и снизить энергопотребление одного изображения в 1,6 раза по сравнению с графическими процессорами (ГП).

Кроме того, специализированные МПВ обеспечивают минимальную задержку между входом и полученным выходом, что важно в таких приложениях, где компьютер должен принимать быстрые решения. Хорошими примерами являются автономные транспортные средства, навигация беспилотных летательных аппаратов, телемедицина и др. Все указанные области в настоящее время, наряду с ИИ, имеют экспоненциальный рост, что создает соответствующий спрос на применение, совершенствование и специализацию обсуждаемых архитектур. Существующие архитектуры МПВ ограничены в своих возможностях и не способны обслуживать современные крупномасштабные модели ИИ. В связи с этим совершенствование архитектур и их исследование стали крайне востребованными.

Диссертация посвящена основным вопросам разработки и исследованию архитектуры МПВ для выводящего ИИ.

Объект исследования. Факторы, влияющие на структуру архитектуры МПВ для выводящего ИИ. Исследование методов оптимизации архитектуры в зависимости от типа нейронных сетей (НС) и используемого аппаратного обеспечения.

Цель работы. Исследование и разработка способов компенсации между занятостью ресурсов платформы, энергопотреблением и производительностью в зависимости от параметров специализированного МПВ и НС.

Методы исследования. В ходе исследования были использованы современные методы оценки, моделирования и оптимизации архитектуры МПВ для выводящего ИИ, а также методы разработки соответствующего программного обеспечения (ПО).

Научная новизна:

- Предложены способы разработки архитектуры МПВ, которые позволяют существенно повысить эффективность и оптимизировать ресурсные затраты систем выводящих ИИ, доводя их до уровня современных практических требований.
- Разработан метод сокращения неэффективных соединений умножителей, который за счет неизменности весовых коэффициентов НС и, следовательно, коэффициентов умножителей на этапе вывода обеспечивает сокращение

используемых регистров на 38% и цифровых сигнальных процессоров на 27% при росте количества таблиц истинности (ТИ) всего на 5%.

- Разработан способ регулирования работы сумматоров, благодаря которому с применением каскадных и древовидных соединений обеспечивается рост максимальной рабочей частоты в 2,2 раза при увеличении занятости устройства на 53,5%, а также снижение занятости до 43% при уменьшении производительности до 87,7%.
- Создана процедура сжатия НС, благодаря которой за счет применения методов квантования весовых коэффициентов, активаций и отсечения нейронов снижается энергопотребление специализированного МПВ на 63,21% при потере точности вывода на 4,9%.
- Предложена последовательность реализации архитектуры НС с помощью ориентированной на ее применение ИС, которая благодаря характерным для данного аппаратного обеспечения особенностям позволила сэкономить более 90% энергопотребления и повысить производительность на 67,5% за счет потери гибкости устройства.

<u>Практическая ценность работы.</u> Разработанные в диссертации средства проектирования архитектуры МПВ для вывода ИИ были реализованы в программном инструменте "Neural Network Circuit Designer", который был внедрен в компании ООО “ЭНДЖИН” и позволил сократить время проектирования и проверки ускорителей в 2...3 раза.

Реализация предложенных методов с помощью программного инструмента "Neural Network Circuit Designer" позволила обеспечить снижение занятости элементов в среднем на 15,57%, снижение энергопотребления – на 13,7% при потере точности вывода – 4,9% и снижении максимальной частоты на 7,37%.

<u>На защиту выносятся:</u>

- метод сокращения неэффективных соединений умножителей;
- средство регулирования работы сумматоров;
- процедура сжатия НС;
- последовательность реализации архитектуры НС с помощью ориентированной на ее применение ИС;

<u>Достоверность научных положений</u> в диссертации была подтверждена экспериментальными результатами моделирования и математическими обоснованиями, представленными в работе.

<u>Внедрение.</u> Разработанный программный инструмент "Neural Network Circuit Designer" был внедрен в компании ООО “ЭНДЖИН”. Он используется при проектировании специализированных МПВ для НС с целью повышения их надежности и производительности вывода.

<u>Апробация работы.</u> Основные научные и практические результаты диссертации докладывались на:

- 21-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, трёх глав, основных выводов, списка литературы из 126 наименований и четырёх приложений. В первом приложении представлен акт о внедрении диссертации, во втором - фрагменты Verilog-описания исследуемых в реализаций НС, в третьем - фрагменты описания программного средства " Neural Network Circuit Designer ", а в четвертом - списки использованных рисунков, таблиц и сокращений. Основной объем диссертации составляет 105 страниц, а вместе с приложениями - 137 страниц, включая 77 рисунков и 10 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 проблемы вывода НС на современных ЦП и ГП. Обсуждены решения проблем производительности и энергопотребления сетей при их реализации на программируемых вентильных матрицах (ПВМ) и интегральных схемах специального назначения (ИССН). Изучены ограничения вышеперечисленных платформ и возможные улучшения архитектур.

Необходимость применения ИИ с использованием ПВМ и ИССН вытекает из постоянно растущих требований приложений ИИ, включая высокую производительность, энергоэффективность, гибкость и возможность обработки данных в реальном времени.

Спрос на вычислительные мощности для ИИ со временем растет экспоненциально, особенно ускорившись в 2012-2014 годах из-за распространения использования ГП и их комбинаций (рис. 1). Увеличение возможностей ИИ приводит к несоразмерному росту технических характеристик ГП, т.е. развитие ИИ в настоящее время ограничено скоростью роста возможностей ГП.

ИССН и ПВМ предлагают аппаратное ускорение алгоритмов ИИ за счет специализированной обработки данных. Создание новых специализированных архитектур ускорителей ИССН и ПВМ также обусловлено постоянно растущими энергетическими потребностями ИИ. Ежегодный отчет института «Uptime» показывает, что, хотя вычислительная эффективность центров обработки данных растет из года в год, коэффициент эффективности использования общей мощности практически не изменился с 2018 года.

Это особенно подчёркивает важность использования ПВМ в исследованиях и разработке ИИ. Ускорители НС - это специализированное аппаратное обеспечение, разработанное для эффективного выполнения вычислений, требуемых НС.

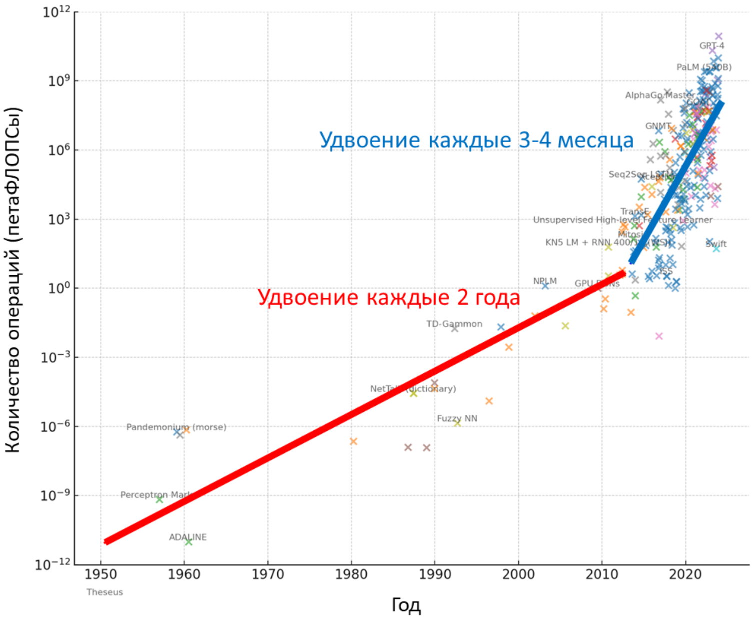

Рис.1. Увеличение вычислительных ресурсов, требуемых для машинного обучения по годам

При рассмотрении ускорения сверточных нейронных сетей (СНС) необходимо сосредоточиться на измерении нескольких характеристик, предоставляемых платформами ПВМ:

- вычислительная пропускная способность;
- точность классификации или обнаружения;
- энергопотребление.

С точки зрения вышеупомянутых величин, современная литература предлагает три основных подхода к задачам глубокого обучения на основе ПВМ: преобразование данных, улучшение путей передач информации и усреднение вычислений.

Выделяются три основных метода преобразования данных. Так называемый метод “im2col” преобразует массивы особенностей и весовых коэффициентов таким образом, что трехмерные свертки становятся умножениями матриц. Метод быстрого преобразования Фурье (БПФ) работает в частотной области, преобразуя свертки в умножения. Наконец, во время фильтрации Винограда свертки упрощаются до поэлементных умножений благодаря группировке данных и линейному преобразованию.

Принятый подход к реализации СНС на ЦП и ГП - отображение сверточных и полносвязных слоев (ПС) как "общих умножений матриц" (ОУМ). Большая часть весовых коэффициентов СНС используется в ПС слоях (рис. 2). Вместо того, чтобы многократно загружать весовые коэффициенты (W) для классификации нескольких входных данных, набор признаков, полученных из пакета входных данных (X), преобразуется в матрицу CHW x B. В этом случае веса загружаются только один раз для данного пакета.

Эта оптимизация позволяет сократить количество операций и требуемую пропускную способность памяти при выводе на наборе входных карт особенностей (КО) для полносвязных слоев. Вместо повторной загрузки весов для каждого входного примера веса загружаются один раз, а затем выполняются вычисления матричного умножения весов на преобразованный набор признаков от всех входных примеров одновременно.

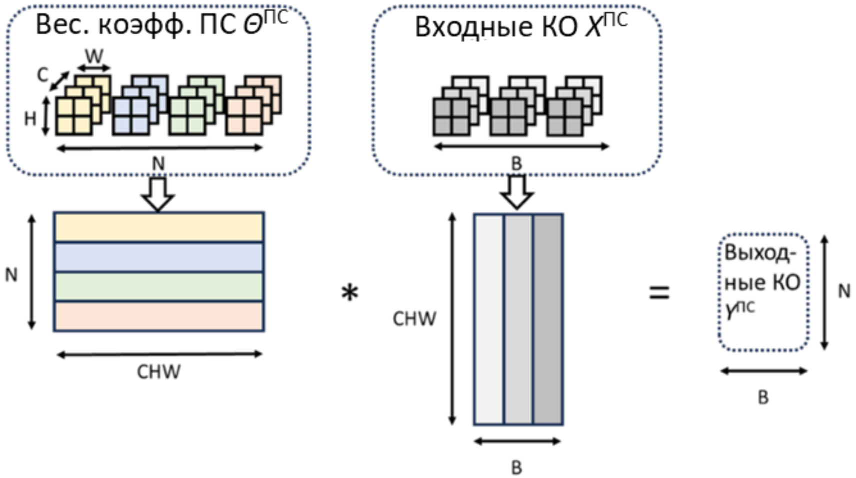

Рис. 2. Вычисление полносвязных слоев с применением ОУМ

Минимальная фильтрация Винограда - это вычислительное преобразование, которое может применяться к архитектурам СНС во время однотактной свертки (рис. 3). Этот алгоритм особенно эффективен при обработке небольших сверток (K<3). В то время, как обычная фильтрация требует $u^2 \times k^2$ умножений, алгоритм Винограда, обозначаемый как F(u x u, k x k), требует $(u+k-1)^2$ умножений. Когда размер группы u=2, а размер ядра k=3, порядок алгебраического упрощения равен x2.25.

Метод БПФ известен как алгоритм, преобразующий двумерные свертки в поэлементные умножения матриц. Использование БПФ для обработки двумерных сверток уменьшает вычислительную сложность с $O(W^2 \times K^2)$ до $O(W^2 \log 2(W))$. По сравнению с классической и фильтрацией Винограда, метод БПФ дает преимущество при свертках с большими размерами ядра (K > 5).

Работа СНС имеет множество возможностей для параллелизма. Однако из-за ограниченных ресурсов устройств ПВМ невозможно полностью распараллелить все конкурирующие процессы.

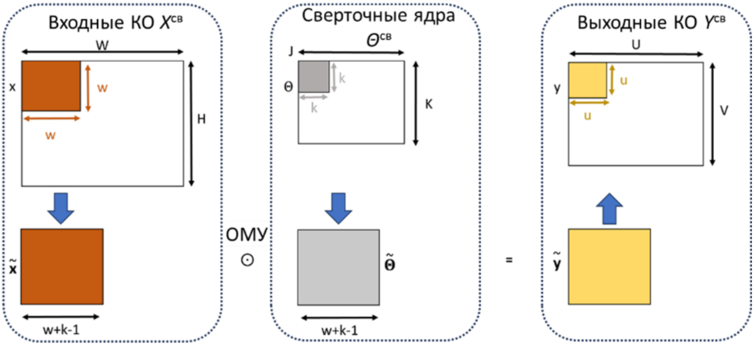

Рис. 3. Преобразование Винограда

Чтобы преодолеть это препятствие, общий подход в лучших реализациях - отобразить ограниченное количество обрабатывающих элементов (ОЭ) на ПВМ.

Для СНС первые ускорители на базе ПВМ реализовывались с помощью систолических массивов. Конфигурация систолических массивов не зависит от моделей СНС, что делает их неэффективными при обработке больших сетей. Из-за неэффективности систолических массивов на ПВМ были разработаны гибкие и специализированные пространственные архитектуры (рис. 4).

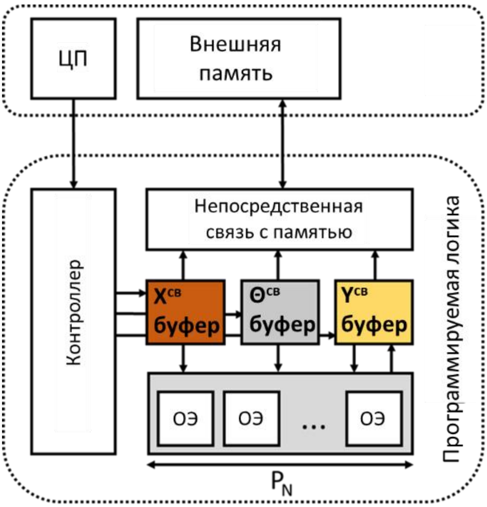

Рис. 4. Архитектура специализированного ускорителя

Производительность СНС может быть ускорена с использованием приближенных вычислений, в которых известны устройства ПВМ. Небольшая точность СНС компенсируется улучшением вычислительной пропускной способности и энергоэффективности ускорителя. При этом применяются два основных подхода. Первый использует приближенную арифметику для обработки слоев СНС с пониженной точностью. Второй подход сокращает количество операций, возникающих в СНС, без большого влияния на точность моделирования.

Обучение и вывод СНС с крайне сжатым представлением данных в последнее время стали предметом исследовательского интереса. Предлагается концепция "двоичных нейронных сетей", где активации и весовые коэффициенты представлены всего одним битом. В этих сетях отрицательные значения представлены как 0, а положительные - как 1.

В дополнение к приближенной арифметике, в некоторых исследованиях делались попытки сократить количество операций, выполняемых в СНС. В реализациях на базе ПВМ изучались два основных подхода. Это прореживание весовых коэффициентов, которое увеличивает разреженность модели и фильтров. При этом также уменьшается количество необходимых умножений во время вывода.

СНС являются сверхпараметризованными сетями, и значительное количество весовых коэффициентов может быть удалено или прорежено без существенного влияния на точность классификации. В своей простейшей форме прореживание выполняется по величине значений. Таким образом, весовые коэффициенты с низкими значениями обрезаются, становясь 0.

Перечисленные способы реализации ИИ на основе ПВМ существенно ускоряют работу таких систем, уменьшают необходимый объем памяти и аппаратных ресурсов. Однако они все же не полностью соответствуют современным практическим требованиям и могут быть реализованы только для НС с ограниченными размерами, при использовании передовых и дорогих устройств. Следовательно, необходимо разработать новые архитектуры ПВМ, которые позволят лучше удовлетворить современные практические требования.

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

Методология преобразования нейронных сетей в электрические схемы

Слои свертки являются одними из самых сложных частей СНС с точки зрения вычислений. Для ускорения работы этих слоев предлагается использовать так называемый подход прямого картографирования устройства (ПКУ), который полностью развертывает заданный свернутый слой.

Подход ПКУ имеет два основных преимущества. Он обеспечивает высокую вычислительную производительность и позволяет отказаться от использования какой-либо внешней памяти. Основным недостатком данного способа является отсутствие гибкости ускорителя. При предложенном подходе ПКУ отображение СНС может быть ограничено ресурсами, имеющимися в данной ПВМ. ПКУ СНС подходит для применения к таким сетям и слоям, которые содержат небольшое количество операций при большом количестве данных.

В противоположность ПКУ, существует принцип фон Неймана, который хорошо работает для большого количества операций и небольшого количества входных данных. Для различных нагрузок данного слоя должны применяться разные принципы. Для понимания типа нагрузки применяется коэффициент "вычисление-коммуникация".

Для выполнения вышеперечисленных действий был разработан соответствующий алгоритм, в котором также проверяется результативность НС на разных стадиях процесса (рис. 5).

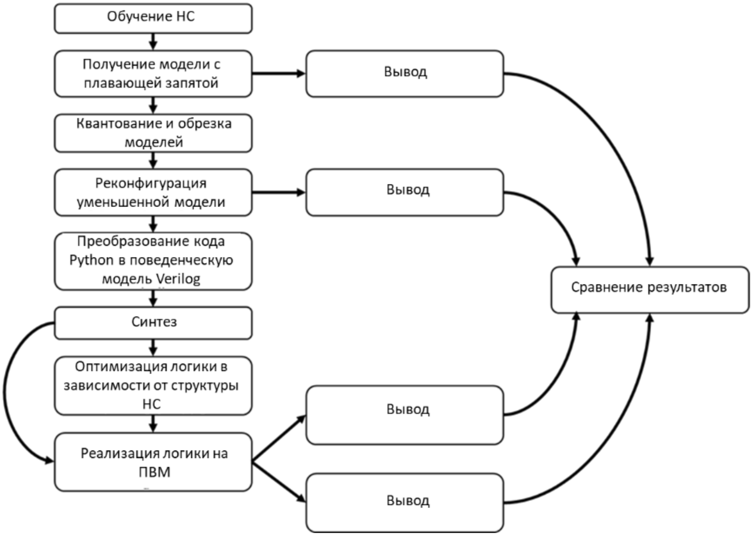

Рис.5. Пошаговый процесс от обучения НС до вывода через устройство

Сокращение структур "Первый вход - первый выход"

Во время конвейеризации потоки “Первый вход - первый выход” (ПВПВ) связывают между собой два зависимых элемента. Однако из-за большого количества элементов в СНС возникающие ПВПВ приводят к высокой занятости аппаратных ресурсов. Для изученных СНС YOLO структуры ПВПВ составляют до 44% от общего количества двоичных элементов (ТИ и процессоры цифровых сигналов), и до 76% регистров. Было проведено исследование занятости ресурсов до и после удаления ПВПВ. В результате экспериментов было собрано количество используемых ресурсов СНС Yolo (табл 1).

Оптимизация умножительных элементов

Цепочка умножителей оптимизируется в зависимости от значений произведений. По умолчанию при синтезе схемы создаются умножители с динамическими входами. Однако при реализации НС можно воспользоваться тем, что все производные статичны и не всегда необходимо создавать целый умножитель для операции произведения. Выполняют следующие оптимизации (рис. 6):

- цепочка умножителей удаляется, если одно из произведений равно 0;
- умножитель заменяется прямым соединением, если одно из произведений равно 1;

- умножитель заменяется регистром сдвига, если одно из произведений является степенью 2;

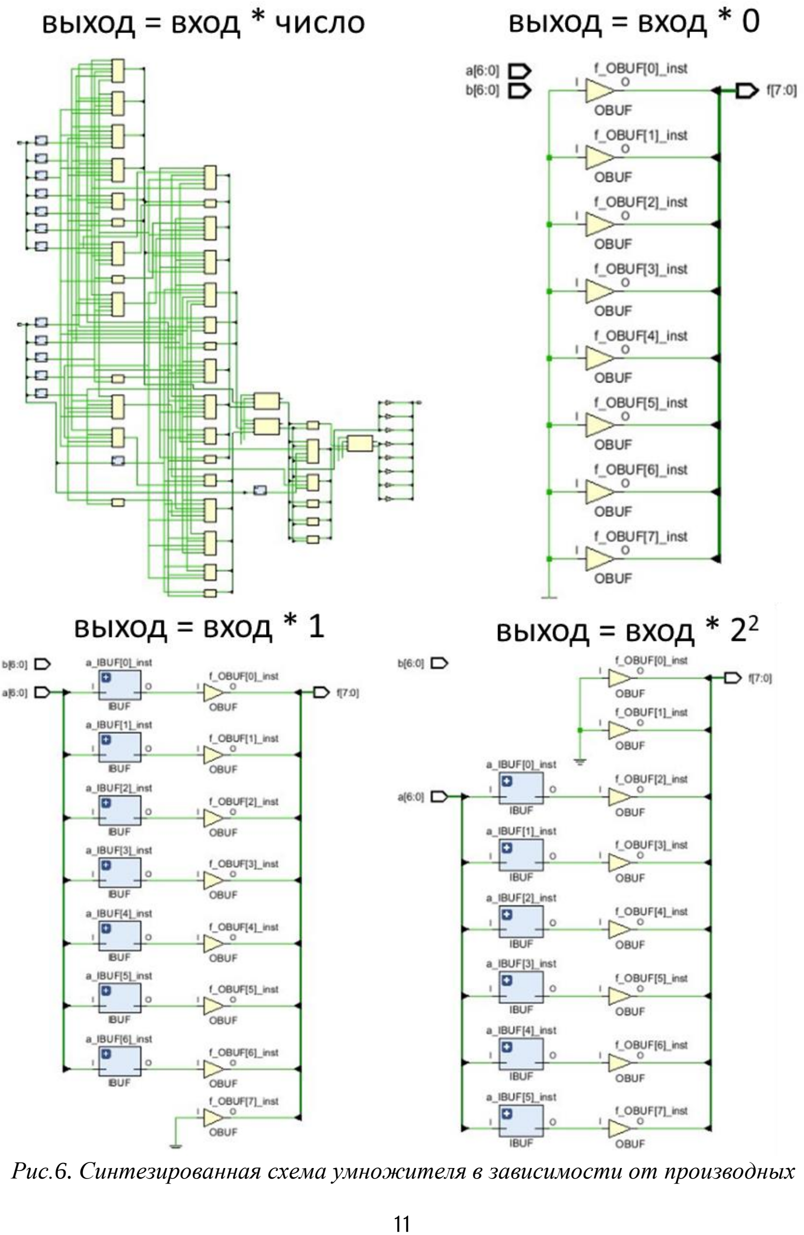

Рис.6. Синтезированная схема умножителя в зависимости от производных

Таблица 1

Занятость ПВМ при реализации сети YOLO

Модель НС	С ПВПВ			Без ПВПВ		
	ТИ	Регистр	ПЦС	ТИ	Регистр	ПЦС
YOLO3	67254	114831	590	74503	71194	486
YOLO4	74412	85916	476	48512	48113	263
YOLO5	71213	75320	433	81602	57476	280
YOLO6	64888	101904	519	39196	59108	233

Эти оптимизации позволяют сократить аппаратные ресурсы, необходимые для реализации умножителей, а также уменьшить задержку и повысить производительность цепочки умножителей. Кроме того, инструменты синтеза также могут выполнять другие оптимизации, такие как объединение констант, распараллеливание операций и др. в зависимости от структуры цепочки умножителей и целей оптимизации (площадь, задержка, энергопотребление).

Архитектура ускорителя фиксируется, значительно снижая его гибкость. Прямое отображение на ПВМ требует повторного синтеза всей архитектуры каждый раз, когда топология или весовые коэффициенты СНС меняются. Тем не менее, экономия ресурсов, достигаемая с помощью ПВМ, остается существенной (табл. 2).

Таблица 2

Процент весовых коэффициентов нулевых умножителей и умножителей на степень 2 в свёрточных слоях

Сеть	AlexNet	DeepComp	SqueezeNet	VGG16	VGG16	VGG16
Слой	conv1	conv1	conv1	conv1-1	conv1-2	conv2-1
Вес. коэфф.	34848	34848	14112	1728	36864	73728
Умн. на ноль (%)	36,84	38,32	22,67	5,56	44,66	57,15
Умн. на степень 2 (%)	26,94	26,21	39,34	18,98	26,25	20,64

Оптимизация сумматоров

При создании сумматоров с несколькими операндами (СНО) инструменты синтеза последовательно соединяют несколько сумматоров, каждый из которых строится с помощью комбинационной логики. Создаётся критический путь, который ограничивает рабочую частоту платформы.

Первым предложенным решением проблемы является конвейеризация СНО (рис. 7). После каждого цифрового сумматора размещается регистр. В конвейеризированном варианте добавляются регистры, количество которых линейно зависит от количества входов. Максимальная частота в конвейеризированном случае почти не меняется при изменении количества операндов. Этот подход позволяет обменивать занятость ресурсов устройства на вычислительную производительность.

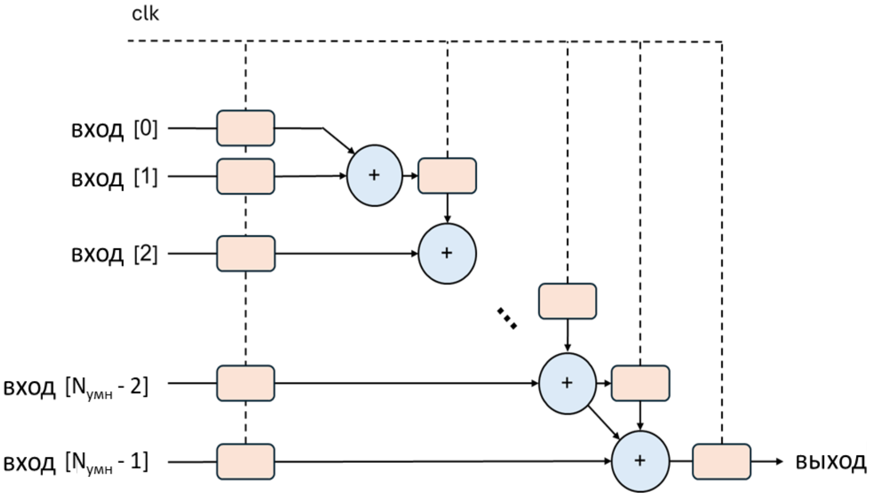

Рис.7. Добавление конвейерных регистров

Еще один способ повышения рабочей частоты - сокращение критичного пути. Для этого необходимо использовать древовидную структуру сумматоров.

В древовидном СНО создается несколько этапов суммирования (рис. 8). На каждом этапе выполняется сложение только двух слагаемых. После каждого этапа количество слагаемых уменьшается вдвое, и так до выходного сумматора. Для СНО с n входами требуемое количество этапов будет log2(n). Это означает, что задержка критического пути будет расти по закону log2(n), а максимальная частота будет уменьшаться как O(1/log2(n)).

Таким образом, древовидная структура СНО позволяет уменьшить задержку критического пути и повысить максимальную рабочую частоту по сравнению с линейной структурой, где все сумматоры соединены последовательно. Однако для реализации древовидной структуры требуется больше аппаратных ресурсов.

Поэтому возникает компромисс между задержкой, максимальной частотой и аппаратными затратами при проектировании СНО. К этому методу также можно добавить конвейеризацию, то есть после каждого этапа древовидной структуры сумматоров добавить регистр, тем самым увеличив рабочую частоту (рис. 9).

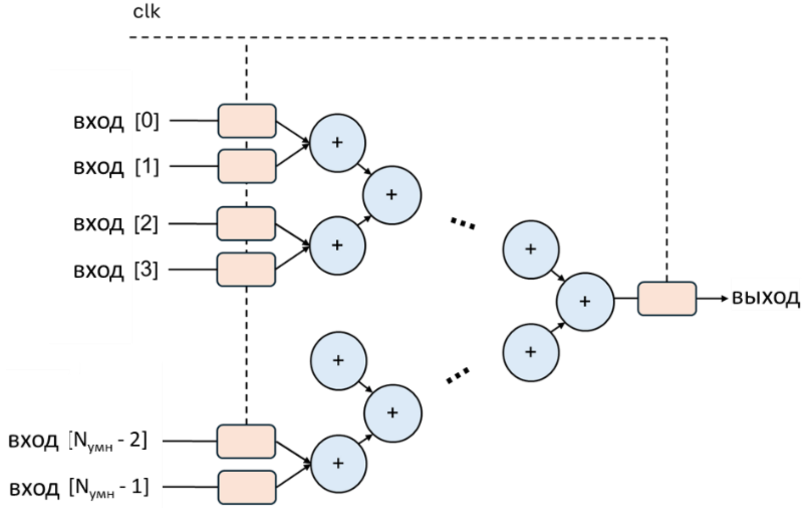

Рис. 8 Древовидное соединение сумматоров

Конечно, подобно каскадному соединению, конвейеризация влияет на занятость ресурсов, при этом сохраняя постоянной рабочую частоту.

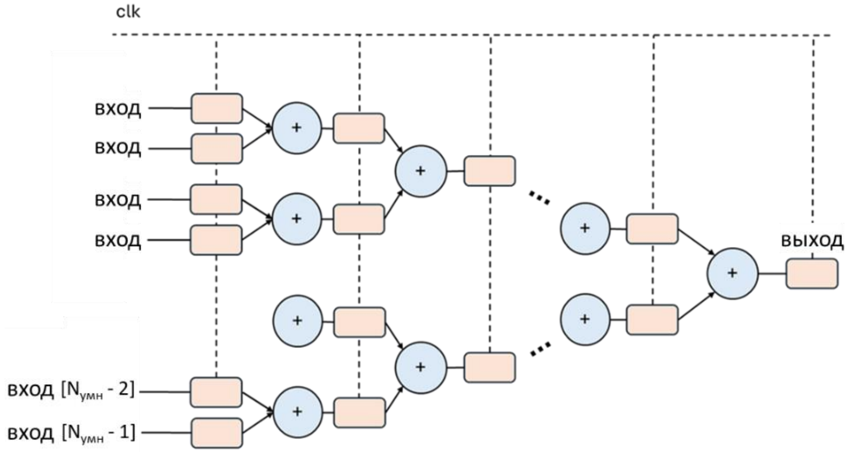

Рис. 9. Конвейерное древовидное соединение сумматоров

Применение квантования и обрезки для снижения энергопотребления

Для количественной оценки сокращения сетей были проведены опыты на сети “Cats vs Dogs”. После обучения точность классификации модели достигает 97,7%.

Затем на обученной модели выполняется операция обрезки. Это повторяющаяся операция, в которой нейроны с наименьшим влиянием удаляются (10% нейронов на каждом этапе), а сеть перенастраивается (переобучается) с целью восстановления точности работы. Процесс устранения и переобучения повторяется шесть раз. Окончательная точность модели снижается до 92,94% (табл. 3).

Таблица 3

Точность распознавания объектов нейронной сетью после квантования и обрезки

Модель НС	Энергопотребление (Вт)	Точность (%)
Модель с плавающей запятой	26,810	97,7
Обрезка 1	23,403	96,4
Обрезка 2	23,648	96,3
Обрезка 3	20,446	94,1
Обрезка 4	19,101	93,8
Обрезка 5	17,252	93,4
Обрезка 6	15,383	92,9
Квантование	11,165	92,8
ПВМ-улучшение	9,864	92,8

Большой интерес также представляет различие в работе одной и той же архитектуры на платформах ПВМ и ИССН. Несмотря на сходные функции, ИССН имеет ряд преимуществ по сравнению с ПВМ. Показано превосходство ИССН в плане энергопотребления при эквивалентных условиях тестирования (табл 4).

Таблица 4

Сравнение энергопотреблений ПВМ и ИССН (Вт)

ПВМ	Энергопотребление	cifar10	vgg16	yolo	Alexnet
Artix-7	Динамическое	2,328	7,128	7,338	7,212
	Статическое	0,244	0,736	0,797	0,752
ИССН	Динамическое	0,189	0,4859	0,5132	0,4808
	Статическое	0,0235	0,0849	0,0927	0,08997

В третьей главе представлено разработанное программное средство “Neural Network Circuit Designer”, которое дает возможность реализовать предложенные решения, выполнить моделирование, проанализировать полученные результаты. ПО снижает вероятность ошибки, вносимой дизайнерами.

Программное средство в качестве входных данных ожидает НС в формате

«*.caffe». Затем пользователь выбирает предпочтительную аппаратную платформу (рис. 10). После выбора пользователь может выполнить сжатие сети (квантование FP32-FP8 / обрезка в 1...6 шагов). При желании этот этап можно также пропустить.

Сжатая НС затем передается инструменту VitisAI, который преобразует НС в поведенческий Verilog-код, а затем выполняет синтез. Синтезированная сеть подается на вход в среду, где выполняются оптимизации.

После оптимизации выходной файл проекта может быть загружен в целевое вычислительное устройство (ПЛИС или КСИА). Полученная аппаратная реализация НС затем может использоваться для выводов в приложениях.

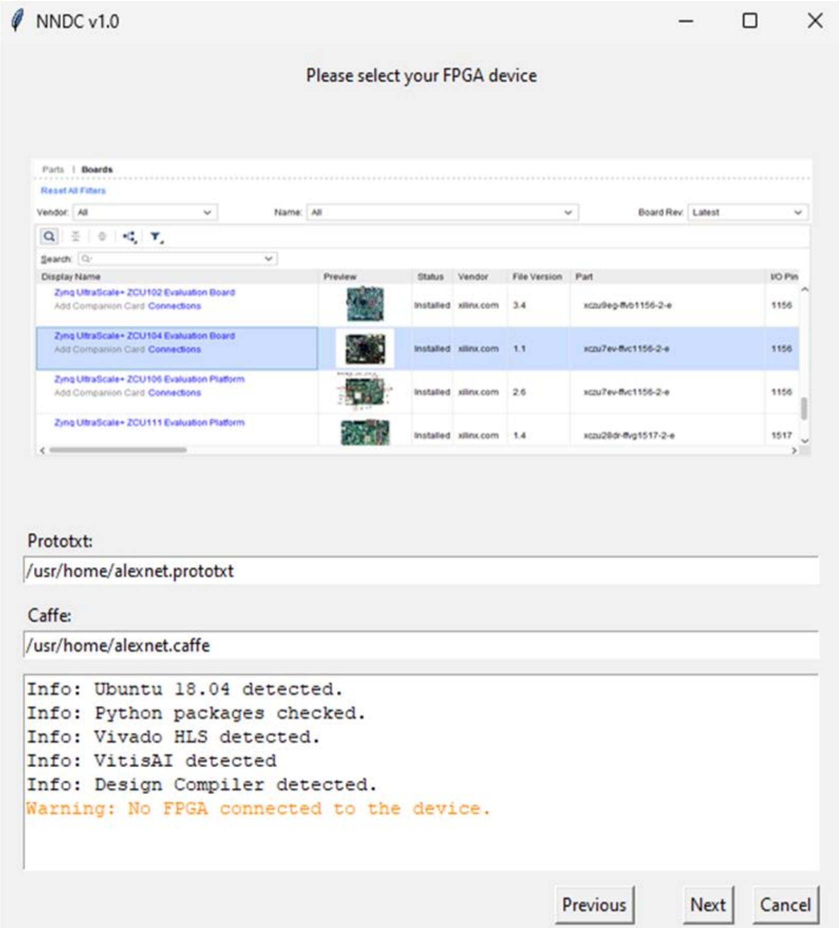

Рис. 10. Главное окно ПО

С помощью программного инструмента NNDC были спроектированы и протестированы реализации НС YoloV6, AlexNet, DeepComp, SqueezeNet и VGG16 на ПВМ и ИССН (табл. 5).

Таблица 5

Оценка эффективности проектирования СНС с помощью программного инструмента

Устройство	Модель	ТИ	Регистры	ПЦС	Энергопот-ребление (Вт)
Существующее	Yolov6	305457	565751	2206	13,326
	AlexNet	180663	334614	1305	10,976
	DeepComp	59361	75112	365	9,125
	SqueezeNet	31837	48499	300	8,923
	VGG16	408713	785923	2814	21,784
С применением предложен-ного метода	Yolov6	319247	349935	1357	10,972
	AlexNet	188819	206969	803	9,445
	DeepComp	56649	58927	326	9,024
	SqueezeNet	32710	44828	297	8,901
	VGG16	394675	481258	2059	16,981

С помощью программного инструмента NNCD удалось достичь в среднем 15,57% занятости ресурсов и снижения энергопотребления на 13,7% за счет уменьшения пропускной способности на 4,9% и снижения максимальной частоты на 7,37%.

Представленные сравнительные данные свидетельствуют об эффективности предложенных подходов к оптимизации и работе программного средства NNDC. Предложенная среда проектирования и предоставленный программным средством интерфейс позволяют сократить время, затрачиваемое на сжатие обученных НС, синтез, оптимизацию синтезированных архитектур и их анализ, в 2...3 раза. Программное средство также позволяет выполнять проектирование регистровых передач на уровне вентилей с использованием различных платформ ПВМ и ИССН.

ОСНОВНЫЕ ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. Предложены методы разработки архитектуры программируемых вентильных матриц, позволяющие существенно повысить эффективность и снизить ресурсопотребление систем реализации исскуственного интеллекта, построенных на их основе, доводя их до уровня современных практических требований.
2. Предложен метод сокращения неэффективных соединений умножителей, который за счет неизменности весовых коэффициентов нейронных сетей и, следовательно, коэффициентов умножителей в процессе распространения обеспечивает соответственно сокращение используемых регистров на 38% и

процессоров цифровых сигналов на 27% при росте количества таблиц истинности всего на 5%.

3. Разработан способ регулирования работы сумматоров, благодаря которому с применением каскадных и древовидных соединений обеспечивается рост максимальной рабочей частоты в 2,2 раза при росте занятости устройства на 53,5%, а также снижение занятости до 43% при снижении производительности до 87,7%.

4. Создана процедура сжатия нейронных сетей, благодаря которой за счет применения методов квантования весовых коэффициентов и активаций, а также удаления нейронов уменьшает энергопотребление массива программируемых логических блоков на 63,21% при потере точности распространения 4,9%.

5. Предложена последовательность реализации архитектуры нейронных сетей с помощью ориентированной на ее применение интегральной схемы, которая благодаря характерным для данного аппаратного обеспечения особенностям позволила сэкономить более 90% энергопотребления и повысить производительности на 67,5% за счет потери гибкости аппарата.

6. Разработанный программный инструмент "Neural Network Circuit Designer" был внедрен в компании ООО “ЭНДЖИН”. Он используется при проектировании специализированных матриц программируемых вентилей для нейронных сетей с целью повышения их надежности и производительности вывода.

Основные результаты диссертации опубликованы в следующих работах:

1. **Melikyan V., Grigoryan M., Avetisyan 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. **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.
3. Real Number Modeling Flow of Digital to Analog Converter / **Melikyan V.Sh., Hovhannisyan V.D., Grigoryan M.T., Avetisyan A.A., Grigoryan H.T.** // Proc. Univ. Electronics. – 2021. - vol. 26, № 2. - P. 144–153, doi: 10.24151/1561-5405-2021-26-2-144-153
4. **Avetisyan A.A.** Investigating Power Metrics of Neural Networks After Pruning, Quantization, DPU Implementation: A Comparative Analysis // Proceedings of the RA NAS and NPUA. Series of Technical Sciences ISSN:0002-306X. – 2023. - Vol. 76, № 4. - P. 506-513.
5. **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. 12-21.

ԱՄՓՈՓԱԳԻՐ

Արհեստական բանականության օգտագործումը թեմատիկ ծրագրավորվող փականների զանգվածների (ԹԾՓԶ) իրականացմամբ վերջին մի քանի տարիների ընթացքում մեծ հետաքրքրություն է առաջացնում բազմաթիվ տեղեկատվական տեխնոլոգիաներով զբաղվող ընկերությունների շրջանում: Այս ուշադրությունը պայմանավորված է նման սարքերի բազմաթիվ առավելություններով, ինչպիսիք են զուգահեռացման մեծ հնարավորությունները, մասնագիտացվող, ճկուն ճարտարապետությունը և ցածր էներգասպառումը: Մյուս կողմից, ԹԾՓԶ հարթակները հեշտ մասշտաբավորելի են և համատեղելի: Այսինքն, մեծ հաշվարկային բեռերի դեպքում հնարավոր է օգտագործել է հարթակների խմբավորում, ընդ որում փոփոխել խմբում ակտիվ տարրերի քանակն ըստ տվյալ խնդրի պահանջների: Դա նշանակում է, որ ԹԾՓԶ-ները հնարավոր է օգտագործել հաշվողական միավորներում՝ կենտրոնական և գրաֆիկական միջուկների նմանությամբ և համակցությամբ:

Ներկայումս սովորական ԹԾՓԶ-ները լայնորեն օգտագործվում են տվյալների կենտրոններում և բարձր արդյունավետության հաշվողական համակարգերում՝ մեքենայական ուսուցման գործընթացներն արագացնելու համար: Microsoft, Google, OpenAI, Intel և մի շարք այլ հայտնի ընկերությունների կատարած հետազոտությունները ցույց են տալիս, որ ԹԾՓԶ-ները առաջադրանքին համապատասխան կարգավորում ստանալու դեպքում կարող են գերազանցել կենտրոնական և գրաֆիկական պրոցեսորներին իրենց արտադրողականությամբ՝ ապահովելով ավելի ցածր էներգասպառում:

Բացի այդ, մասնագիտացված սարքաշարերն ունենում են տվյալների փոխանցման ընթացքում նվազագույն հապաղում, ինչը կարևոր է ակնթարթային արձագանք պահանջող հավելվածների համար: Նման կիրառությունների օրինակներ են ռոբոտային համակարգերը, անօդաչու թռչող սարքերի ինքնավար կառավարումը, շտապ բժշկական օգնության համակարգերը և այլն: Այս ոլորտները, արհեստական բանականության զարգացման հետ մեկտեղ, արագ աճ են ապրում՝ ստեղծելով համապատասխան ճարտարապետությունների զարգացման, կատարելագործման և մասնագիտացման զգալի պահանջարկ: Ժամանակակից ԹԾՓԶ ճարտարապետությունները սահմանափակ են արհեստական բանականության լայնածավալ մոդելներին աջակցելու ունակության մեջ, ինչը նրանց հետագա կատարելագործումն ու հետազոտումը պահանջված խնդիր է դարձնում:

Ատենախոսությունը նվիրված է արհեստական բանականության առաջադրանքների կատարման թեմատիկ ծրագրավորվող փականների զանգվածի ճարտարապետության մշակման և ուսումնասիրության հիմնական հիմնահարցերին:

Առաջարկվել են ծրագրավորվող փականներով մատրիցի ճարտարապետության մշակման մոտեցումներ, որոնք թույլ են տալիս էապես բարելավել դրանց միջոցով կառուցվող առերեսվող արհեստական բանականության համակարգերի արդյունավետությունը և ծախսվող ռեսուրսները՝ հասցնելով դրանք ժամանակակից գործնական պահանջների մակարդակին:

Նախագծվել է բազմապատկիչների ոչ արդյունավետ միացումների կրճատման եղանակ, որը նեյրոնային ցանցերի քաշային գործակիցների և հետևաբար բազմապատկիչների արտադրիչների առերեսման ընթացքում անփոփոխ լինելու շնորհիվ ապահովում է օգտագործված գրանցիչների և թվային ազդանշանների պրոցեսորների քանակների համապատասխանաբար 38% և 27% նվազեցում՝ իսկության աղյուսակների քանակի ընդամենը 5% աճի հաշվին:

Մշակվել է գումարիչների աշխատանքի կարգավորման միջոց, որը կասկադային և ծառային միացումների կիրառմամբ ապահովել է աշխատանքային առավելագույն հաճախության 2,16 անգամ աճ սարքի զբաղվածության 53,5% աճի հաշվին, ինչպես նաև զբաղվածության մինչև 43% նվազեցում կատարողականի մինչև 87,7 % նվազեցման հաշվին:

Ստեղծվել է նեյրոնային ցանցի կրճատման ընթացակարգ, որը կշիռների և ակտիվացումների քվանտացման և նեյրոնների էտման եղանակների կիրառման շնորհիվ փոքրացնում է թեմատիկ ծրագրավորվող փականների զանգվածի էներգասպառումը 63,21%-ով՝ առերեսման ճշտության 4.9% կորստի հաշվին:

Առաջարկվել է նեյրոնային ցանցի ճարտարապետության կիրառությանը կողմնորոշված ինտեգրալ սխեմայի միջոցով իրականացման հաջորդականություն, որը տվյալ սարքային ապահովմանը բնորոշ հատկությունների շնորհիվ թույլ է տվել խնայել ավելի քան 90% էներգասպառում և շահել 67.5% կատարողական, սարքային ճկունության կորստի հաշվին:

Ատենախոսությունում մշակված առերեսվող արհեստական բանականությամբ ծրագրավորվող փականների մատրիցի ճարտարապետության մշակման միջոցներն իրագործվել են «Neural Network Circuit Designer» ծրագրային միջոցում, որը ներդրվել է «ԷՆՋԻՆ» ՍՊԸ-ում և թույլ է տվել կրճատել արագացուցիչների նախագծման և ստուգումների ժամանակը 2-3 անգամ: Առաջարկված մեթոդների իրագործումը՝ «Neural Network Circuit Designer» ծրագրային գործիքի միջոցով, թույլ է տվել ապահովել տարրերի զբաղվածության միջինում 15,57%, և էներգասպառման 13,7% նվազեցում՝ առերեսման 4,9% և առավելագույն հաճախության 7,37% նվազեցման հաշվին:

ASHOT AZAT AVETISYAN

DEVELOPMENT AND RESEARCH OF THE ARCHITECTURE OF A PROGRAMMABLE GATE ARRAY FOR ARTIFICIAL INTELLIGENCE INFERENCE

SUMMARY

The use of artificial intelligence through the implementation of domain-specific programmable gate arrays has been a topic of great interest among many IT companies over the past few years. This attention is due to the numerous advantages of such devices, such as great parallelization opportunities, specialized, flexible architecture, and low power consumption.

Today, conventional FPGAs are widely used in data centers and high-performance computing systems to accelerate machine learning processes. Research by Microsoft, Google, OpenAI, Intel, and a number of other well-known companies shows that domain-specific programmable gate arrays, when properly configured for the task, can outperform central and graphics processors in terms of performance while having lower power consumption.

Furthermore, specialized hardware has minimal latency during data transmission, which is important for applications that require an instant response. Examples of such applications are robotic systems, autonomous control of unmanned aerial vehicles, emergency medical systems, and so on. These areas, along with the development of artificial intelligence, are experiencing rapid growth, creating a significant demand for the development, improvement, and specialization of appropriate architectures. The architectures of modern domain-specific programmable gate arrays are limited in their ability to support large-scale artificial intelligence models, which makes their further improvement and research a pressing issue.

The dissertation is dedicated to the main issues of developing and studying the architecture of domain-specific programmable gate arrays for performing artificial intelligence tasks.

New approaches to developing programmable gate array matrix architectures have been proposed that allow significantly improving the efficiency and resource consumption of built-in artificial intelligence systems, bringing them up to the level of modern practical requirements.

A method for reducing inefficient multiplier connections has been designed, which, due to the unchanged nature of neural network weights and, consequently, multiplier operands during deployment, provides a 38% reduction in the number of registers and 27% reduction in the number of digital signal processors used, with only a 5% increase in the number of truth tables.

A means of regulating the operation of adders has been developed, which, using cascading and tree connections, provided a 2.16-fold increase in the maximum operating frequency with a 53.5% increase in device utilization, as well as up to a 43% reduction in utilization with up to an 87.7% reduction in performance.

A neural network pruning procedure has been created, which, by applying weight and activation quantization and neuron pruning methods, reduces the power consumption of the domain-specific programmable gate array by 63.21% with a 4.9% loss of deployment accuracy.

A sequence for implementing a neural network architecture using an application-specific integrated circuit has been proposed, which, due to the characteristics specific to this hardware, allowed saving more than 90% of power consumption and gaining 67.5% performance, at the expense of hardware flexibility loss.

The methods for developing a programmable gate array matrix architecture with embedded artificial intelligence, developed in the dissertation, have been implemented in the "Neural Network Circuit Designer" software tool, which has been introduced at "NGENE" LLC and allowed reducing the design and verification time of accelerators by 2-3 times. The implementation of the proposed methods using the "Neural Network Circuit Designer" software tool allowed achieving an average reduction of 15.57% in element utilization and 13.7% in power consumption, with a 4.9% reduction in deployment accuracy and a 7.37% reduction in maximum frequency.