На правах рукописи

Шатравин Владислав

АЛГОРИТМЫ ВЫЧИСЛЕНИЯ ОТКЛИКА НЕЙРОННЫХ СЕТЕЙ НА ДИНАМИЧЕСКИ ПЕРЕСТРАИВАЕМЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕДАХ

2.3.8. Информатика и информационные процессы

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Томск – 2023

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский Томский государственный университет».

Научный руководитель: доктор технических наук
Шидловский Станислав Викторович

Официальные оппоненты:

Курносов Михаил Георгиевич, доктор технических наук, профессор, федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет телекоммуникаций и информатики», кафедра вычислительных систем, профессор

Катаев Михаил Юрьевич, доктор технических наук, профессор, федеральное государственное бюджетное образовательное учреждение высшего образования «Томский государственный университет систем управления и радиоэлектроники», кафедра автоматизированных систем управления, профессор

Жданов Дмитрий Сергеевич, кандидат технических наук, федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет», лаборатория медицинского приборостроения (CyberMed), заведующий лабораторией

Защита состоится 19 октября 2023 г. в 10 час. 35 мин. на заседании диссертационного совета «НИ ТГУ.2.3.01», созданного на базе Института прикладной математики и компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет», по адресу: 634050, г. Томск, пр. Ленина, 36 (учебный корпус № 2 ТГУ, аудитория 104).

С диссертацией можно ознакомиться в Научной библиотеке и на официальном сайте федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет» www.tsu.ru.

Материалы по защите диссертации размещены на официальном сайте ТГУ: https://dissertations.tsu.ru/PublicApplications/Details/ec4bf598-fac1-4d50-8105-365738ef37a3

Автореферат разослан «____» сентября 2023 г.

Ученый секретарь
диссертационного совета,
доктор физико-математических
наук, доцент

Воробейчиков
Сергей Эрикович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Последние десятилетия характеризуются всё более активным применением алгоритмов машинного обучения, в особенности нейронных сетей. Их способность эффективно решать плохо формализованные задачи открывает широкие возможности для построения интеллектуальных информационных систем. На сегодняшний день нейронные сети применяются для задач распознавания образов, обработки естественного языка, классификации объектов, предсказательной аналитики, восстановления изображений и многих других. Целевые системы, опирающиеся на применение нейронных сетей, варьируются от маломощных мобильных устройств до облачных систем и центров обработки данных.

Значительная вычислительная сложность нейронных сетей делает неэффективным применение вычислительных устройств классических архитектур. Всё большую роль играют устройства на основе многоядерных графических процессоров, а также специализированных вычислителей, построенных на основе программируемых логических интегральных схем (ПЛИС, FPGA) и интегральных схем специального назначения (ASIC). Это связано с их способностью выполнять вычисления с высокой степенью параллелизма, что позволяет значительно ускорить наиболее трудоёмкие операции нейросетевых алгоритмов. Вычислительные устройства, специализированные на реализацию нейросетевых алгоритмов, называют аппаратными ускорителями нейронных сетей.

На сегодняшний день наблюдается явная тенденция к усложнению архитектур нейронных сетей, что приводит к необходимости постоянного совершенствования программно-аппаратных средств их реализации. В связи с этим разработка новых моделей и алгоритмов функционирования аппаратных ускорителей играет большую роль в обеспечении прикладного применения современных архитектур нейронных сетей. Особенно остро обозначенная проблема стоит для маломощных мобильных и автономных устройств, для которых характерно наличие ограничений по массе и энергоэффективности.

Степень разработанности темы исследования. Наиболее перспективным направлением развития специализированных под алгоритмы машинного обучения вычислительных устройств являются вычислители с архитектурой параллельно-конвейерного типа. Такие вычислители поддерживают несколько параллельных потоков исполнения команд, каждый из которых использует конвейеризацию, то есть выполнение последовательности из нескольких команд над разными фрагментами данных.

В развитие и применение вычислительных устройств с параллельно-конвейерной обработкой данных неоценимый вклад внесли выдающиеся отечественные и зарубежные учёные: А. А. Болотов, Э. В. Евреинов, А. В. Каляев, И. А. Каляев, В. Б. Кудрявцев, И. И. Левин, А. С. Подколзин, И. В. Прангишвили, В. Г. Хорошевский, А. А. Шалыто, Y. H. Chen, T. Hoefler, S. Matsuoka, R. Duncan, M. T. Sterling, J. L. Traff и другие.

В области теории искусственных нейронных сетей большую роль сыграли такие видные учёные, как И. А. Бессмертный, М. В. Воронов, А. И. Галушкин, А. Г. Ивахненко, А. С. Кронрод, С. Николаенко, Я. З. Цыпкин, Y. Bengio, I. J. Goodfellow, S. Haykin, G. E. Hinton, J. J. Hopfield, N. P. Jouppi, Y. A. LeCun, F. Rosenblatt и другие.

В связи с обширной сферой применения нейронных сетей, архитектуры используемых для их расчёта вычислительных устройств сильно различаются. Настольные, серверные и облачные системы опираются на мощные энергозатратные реализации аппаратных ускорителей, повышение производительности которых достигается не только качественным улучшением вычислительных устройств, но и масштабированием вычислительной системы.

Совсем иначе обстоит дело для маломощных устройств, к которым, в частности, относится широкий класс мобильных и автономных интеллектуальных систем. Примерами таких систем являются мобильные роботы, БПЛА, мобильные телефоны, умные датчики интернета вещей, спутниковые системы и многие другие. Их объединяют жесткие ограничения энергопотребления, массы, высокие требования к надёжности, а также частичная или полная недоступность их аппаратного обеспечения для перенастройки после начала эксплуатации. При этом подавляющее большинство предлагаемых архитектур аппаратных ускорителей ориентированы только на улучшение быстродействия и энергоэффективности вычислений в рамках конкретных архитектур нейронных сетей, жертвуя при этом гибкостью, которая может являться решением многих упомянутых ранее проблем.

Ключевой особенностью функционирования автономных мобильных устройств является долговременное пребывание в потенциально изменчивой среде. В некоторых случаях определить условия функционирования устройства до начала эксплуатации невозможно или крайне затруднительно. Более того, изменяться могут и поставленные перед устройством задачи. Эти два фактора оказывают непосредственное влияние на параметры и архитектуры используемых устройством моделей нейронных сетей. Как следствие, применяемые в маломощных информационных системах вычислительные устройства должны поддерживать удалённое изменение реализуемых алгоритмов. Одним из путей обеспечения таких изменений является применение перестраиваемых устройств.

На сегодняшний день предлагаются различные подходы к построению перестраиваемых аппаратных ускорителей, имеющие свои достоинства и недостатки. С учётом упомянутых преимуществ параллельно-конвейерной архитектуры, а также в связи с необходимостью обеспечения динамической перенастройки вычислителя, большой интерес вызывает разработка моделей аппаратных ускорителей нейронных сетей на основе концепции перестраиваемых вычислительных сред. Эта концепция предлагает построение вычислительного устройства в виде геометрически правильной сетки, узлами которой являются простые вычислительные элементы. Каждый элемент настроен на выполнение определённой операции и способен обмениваться данными с соседними элементами. Благодаря независимой настройке и функционированию каждого

отдельного элемента, устройствам на основе перестраиваемых сред свойственна естественная параллельность и динамическая реконфигурируемость.

Несмотря на преимущества перестраиваемых вычислительных сред, на сегодняшний день вопрос их применения для улучшения характеристик специализированных под задачи машинного обучения вычислительных устройств не проработан в должной степени. Данная работа направлена на устранение этого пробела в целях расширения возможностей применения искусственных нейронных сетей в широком классе прикладных задач.

Соответствие паспорту специальности. Результаты исследования соответствуют научной специальности 2.3.8. «Информатика и информационные процессы»:

Пункт 9. Разработка архитектур программно-аппаратных комплексов поддержки цифровых технологий сбора, хранения и передачи информации в инфокоммуникационных системах, в том числе, с использованием «облачных» интернет-технологий и оценка их эффективности.

Пункт 13. Разработка и применение методов распознавания образов, кластерного анализа, нейросетевых и нечетких технологий, решающих правил, мягких вычислений при анализе разнородной информации в базах данных.

Цель исследования. Повышение гибкости вычисления отклика нейронных сетей в программно-аппаратных комплексах при помощи перестраиваемых вычислительных сред.

Задачи:

1. Сформировать концепцию синтеза алгоритмов вычисления отклика нейронных сетей предварительно заданных архитектур на перестраиваемых вычислительных средах.
2. Разработать алгоритм вычисления отклика нейронных сетей заданного множества архитектур на перестраиваемой вычислительной среде.
3. Синтезировать для предложенной модели перестраиваемой вычислительной среды алгоритмы, реализующие функции активации нейрона: линейный выпрямитель (ReLU), сигмоида, гиперболический тангенс, softmax.
4. Синтезировать для предложенной модели перестраиваемой вычислительной среды алгоритмы, реализующие слои нейронной сети: полносвязный, свёрточный, субдискретизации, преобразования матриц в вектор.
5. Оценить корректность и быстродействие синтезированных в рамках исследования алгоритмов посредством проведения экспериментов на имитационных моделях и физическом вычислительном устройстве.

Научная новизна работы:

1. Предложена концепция синтеза алгоритмов вычисления отклика нейронных сетей, отличающаяся применением перестраиваемых вычислительных сред с динамической подстройкой тактовой частоты, что обеспечивает параллельное вычисление всех нейронов одного слоя, высокую эффективность использования ресурсов и масштабируемость.

2. Разработан алгоритм вычисления отклика нейронных сетей на перестраиваемых вычислительных средах, отличающийся вычислением нескольких слоёв сети и локальной перенастройкой за один такт работы среды, что позволяет повысить скорость.

3. Разработаны алгоритмы вычисления функций активации нейронов (сигмоиды, гиперболического тангенса, softmax) на основе операций заданного базиса, выполняемые на независимых вычислителях за один такт, что позволяет вычислять отклик нейронных сетей.

4. Разработаны алгоритмы вычисления отклика слоёв нейронных сетей (полносвязных, свёрточных, субдискретизации, преобразования набора матриц в вектор), адаптированные к вычислительным средам, что обеспечивает выполнение всех этапов вычисления отклика сетей соответствующих архитектур на одной вычислительной среде.

Теоретическая и практическая значимость диссертации. Полученные в рамках работы результаты расширяют сферу применения теории однородных структур в задачах, решаемых с применением искусственных нейронных сетей.

Предложенные алгоритмы открывают перспективы разработки аппаратных ускорителей нейронных сетей на основе перестраиваемых сред. Подобные ускорители не только будут обладать высоким быстродействием и энергоэффективностью при компактных размерах, но и обеспечат функционирование нейронных сетей разнообразных архитектур с произвольным числом слоёв и нейронов, допуская их динамическое изменение, подстройку тактовой частоты, а также восстановление работоспособности через перераспределение вычислений.

Полученные результаты демонстрируют высокую эффективность применения перестраиваемых сред для вычисления моделей нейронных сетей полносвязных, свёрточных и рекуррентных архитектур. Разработанные алгоритмы могут быть применены на специализированных вычислительных устройствах, удовлетворяющих требованию полной или локальной однородности и реализующих операции заданного базиса. Предложенные модели и алгоритмы были использованы при разработке информационной системы анализа дорожного движения, что подтверждено актами внедрения.

Связь работы с научными проектами. Диссертационные исследования выполнены в рамках научного проекта № 20-37-90034 Аспиранты при финансовой поддержке ФГБУ «Российский фонд фундаментальных исследований» по теме «Исследование и разработка моделей и алгоритмов перестраиваемых вычислительных сред для задач машинного обучения» (2020-2022); научного проекта № 14.578.21.0241 в рамках федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы» по теме «Разработка системы автономного интеллектуального функционирования беспилотным летательным аппаратом на базе реконфигурируемых алгоритмов управления, навигации и обработки информации и создание на ее основе аппаратно-программного комплекса защиты от малогабаритных летательных аппаратов» (2017-2020); научного проекта № 19-29-06078 мк при финансовой поддержке ФГБУ

«Российский фонд фундаментальных исследований» по теме «Разработка и исследование конфигурируемых быстродействующих алгоритмов распознавания изображений для оценки дорожной ситуации на базе специализированных мобильных устройств с параллельно-конвейерной архитектурой» (2019-2022); научного проекта № 21-71-00012 при финансовой поддержке «Российского научного фонда» по теме «Исследование и разработка бинарных нейронных сетей для классификации и распознавания изображений» (2021-2023); научного проекта № 5.4.4.22 ПИШ «Разработка общей концепции и требований к цифровой инфраструктуре для управления сельскохозяйственными процессами в области точного земледелия» в рамках программы развития Передовой инженерной школы Томского государственного университета «Агробиотек» (2022); стипендиальной программы НИЦ компании «Huawei» (2020-2021).

Методология и методы исследования. Для решения поставленных задач использовались методы теории нейронных сетей, теории булевой алгебры, теории вычислительных систем. Экспериментальные исследования проводились с использованием методов имитационного моделирования на ПК и прототипирования на программируемых логических интегральных схемах (FPGA).

Положения, выносимые на защиту:

1. Концепция синтеза алгоритмов вычисления отклика нейронных сетей на перестраиваемых вычислительных средах.
2. Алгоритм вычисления отклика нейронных сетей, рассматривающий слои сети как шаги конвейера со скользящим окном настройки, предназначенный для перестраиваемых вычислительных сред.
3. Алгоритмы вычисления функций активации нейронов (сигмоиды, гиперболического тангенса, softmax) за один такт, предназначенные для вычислительных сред, поддерживающих операции заданного базиса.
4. Алгоритмы реализации слоёв нейронных сетей (полносвязных, свёрточных, субдискретизации, преобразования набора матриц в вектор), разработанные для применения на вычислительных средах.
5. Результаты имитационного моделирования искусственного нейрона, функций активации сигмоида и softmax, а также опытной реализации нейронной сети.

Степень достоверности и апробация результатов исследования. Достоверность полученных в рамках данной диссертационной работы результатов подтверждается экспериментами, проведёнными при помощи имитационного моделирования и с применением испытательного стенда.

Результаты работы представлялись и обсуждались на научных конференциях различного уровня, среди которых международные: школа-конференция «Инноватика-2019» (г. Томск), «Инноватика-2020» (г. Томск); третий форум «Интеллектуальные системы 4-й промышленной революции» (г. Томск, 2019); четвёртый форум «Интеллектуальные системы 4-й промышленной революции» (г. Томск, 2021); конференция «International Conference on Information Technology» (г. Амман, Иордания, 2021); конференция «Distributed Computer and Communications Networks: Control,

Computation, Communications» (г. Москва, 2021); конференция «Информационные технологии и математическое моделирование» (г. Карши, Узбекистан, 2022), а также Всероссийская научно-практическая конференция «Перспективные системы и задачи управления» (п. Домбай, Карачаево-Черкессия, 2022).

Публикации. По теме диссертации опубликовано 16 работ, в том числе 4 статьи в журналах, включенных в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (из них 1 статья в российском научном журнале, входящем в Web of Science, 1 статья в зарубежном научном журнале, входящем в Scopus), 3 статьи в сборниках материалов конференций, представленных в изданиях, входящих в Scopus и Springer, 2 свидетельства о государственной регистрации программ для ЭВМ, 6 публикаций в сборниках материалов международных конференции и форумов, школ-конференций, всероссийской научно-практической конференции. В опубликованных работах достаточно полно изложены материалы диссертации.

Структура работы. Диссертационная работа изложена на 143 страницах печатного текста. Состоит из введения, четырёх разделов, заключения, списка сокращений, списка использованной литературы, четырёх приложений. Содержит 11 таблиц и 69 рисунков. Список литературы включает 86 источников, из них 50 на иностранном языке.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы исследования; сформулированы цель и задачи диссертационной работы; отражены её научная новизна, теоретическая и практическая ценность; представлены выносимые на защиту положения; перечислены используемые в работе методы; отмечена степень достоверности и данные об апробации результатов исследования.

В первом разделе приводятся основы теории искусственных нейронных сетей (НС), определяются ключевые проблемы их прикладного применения, описываются основные понятия и принципы функционирования перестраиваемых вычислительных сред. По результатам выполненного в рамках данного раздела исследования формулируются основные задачи диссертационной работы, решению которых посвящены следующие разделы.

В подразделе 1.1 приводятся базовые понятия искусственных нейронных сетей, их преимущества, а также цель и методы их обучения. Описываются ключевые архитектуры НС, которым уделяется основное внимание в последующих разделах – сети прямого распространения, свёрточные и рекуррентные.

В подразделе 1.2 описываются ключевые проблемы прикладного применения НС, приводится обзор и анализ существующих решений, а также осуществляется постановка проблемы исследования.

В подразделе 1.3 описываются основные понятия и принципы перестраиваемых вычислительных сред (ПВС) – дискретных математических моделей широкого класса вычислительных устройств, реализованных в виде геометрически правильной решётки, узлами которой являются простые вычислительные элементы (рисунок 1). Каждый вычислительный элемент (ВЭ) динамически настраивается на реализацию одной операции из заранее заданного базиса и функционирует независимо от других. Это позволяет реализовывать на ПВС сложные алгоритмы потоковой обработки данных. Показаны основные преимущества применения ПВС.

Во втором разделе обсуждаются вопросы разработки алгоритмов вычисления отклика НС на ПВС: описывается предлагаемая в рамках работы концепция синтеза требуемых алгоритмов; задаются основные параметры архитектуры среды; приводится методика формирования базиса операций и функций выходов ВЭ; предлагается реализация отдельного нейрона и полносвязного слоя НС; описываются режимы функционирования разработанной ПВС и процесс её настройки на реализацию конкретного нейросетевого алгоритма.

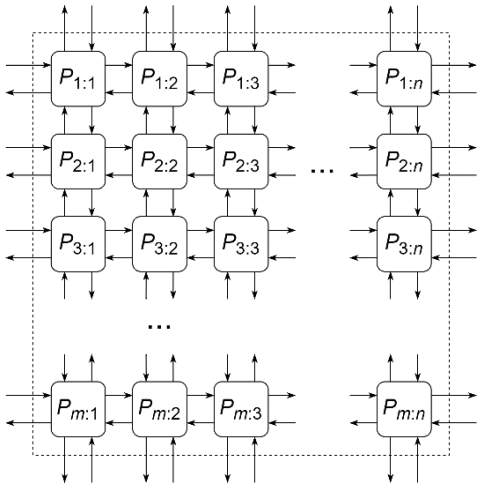


Рисунок 1 – Перестраиваемая вычислительная среда

В подразделе 2.1 описывается концепция синтеза алгоритмов вычисления отклика нейронных сетей заданных архитектур для ПВС, а также разработка самих моделей и алгоритмов ПВС. В качестве целевых архитектур рассматриваются полносвязные сети прямого распространения, свёрточные сети и рекуррентные сети на примере сети Хопфилда. В основе предложенной концепции лежат следующие принципы:

1) ПВС должна поддерживать выполнение всех шагов вычисления отклика НС. Обращение к внешним вычислительным системам должно быть обоснованно и носить исключительный характер.

2) Высокая детальность алгоритмов, обеспечиваемая декомпозицией вычислений отдельных нейронов и их функций активации на коллектив элементов ПВС.

3) Вычисление отдельных нейронов, функций активации и слоёв выполняется асинхронно. Соединение сконфигурированных вычислительных элементов среды формирует комбинационную схему.

4) Промежуточные результаты вычислений хранятся на самой ПВС, что позволяет устранить накладные расходы на обращение к внешней памяти.

5) Локальная перенастройка и конвейеризация позволяют использовать ПВС для вычисления отклика сетей с произвольным числом слоёв, а также уменьшают накладные расходы на перенастройку и увеличивают быстродействие.

6) Динамическая подстройка частоты вычислительного конвейера ПВС, возможная благодаря разному использованию ресурсов ПВС на разных слоях НС, позволяет как увеличить быстродействие при вычислении небольших слоёв, так и адаптировать конвейер и алгоритмы для работы среды на требуемой частоте.

В рамках данной работы рассматриваются двумерные ПВС, каждый ВЭ (кроме граничных) соединён с четырьмя соседними (рисунок 1). ВЭ располагает небольшим объёмом памяти для хранения собственной настройки. Среда функционирует в смешанном (синхронном и асинхронном) режиме. Анализ литературы позволил выделить базис ВЭ из десяти простых операций: передача сигнала (TRS), источник (SRC), умножение с накоплением (MAC), максимум (MAX), минимум (MIN), параметрический выпрямитель (функция активации PReLU, PRL), затвор (GAT), объединение (U), задержка (DEL) и блок (BLK). Также ВЭ настраивается на работу в одном из четырёх направлений, что определяет используемые при выполнении операций входы и выходы. Код направления (z_{dir}), операции (z_{op}) и её дополнительный аргумент (z_{arg}) образуют настроечный сигнал Z вычислительного элемента. Обозначим коды направлений как dir_f, где f принимает одно из четырёх значений: l (слева), t (сверху), r (справа), b (снизу). Аналогичным образом введём обозначение входных сигналов ВЭ $X = (x_l, x_t, x_r, x_b)$, выходных сигналов $Y = (y_l, y_t, y_r, y_b)$, хранимого в памяти элемента значения S, а также функцию $op(x, z, s)$, соответствующую операции, заданной текущей настройкой. Тогда значение на выходе ВЭ может быть описано таблицей 1.

При помощи автоматно-структурного метода были получены автоматные отображения выходов ВЭ. К примеру, выражение

$$\begin{aligned} y_l = {} & s_4(x_r, eq_4(z_{op}, TRS)) \vee ((eq_4(z_{op}, SRC) \vee eq_4(z_{op}, PRL) \vee eq_4(z_{op}, DEL)) \cdot \overline{eq_2(z_{dir}, dir_r)}) \vee \\ & \vee ((eq_4(z_{op}, MAC) \vee eq_4(z_{op}, MAX) \vee eq_4(z_{op}, MIN) \vee eq_4(z_{op}, GAT)) \cdot \overline{eq_2(z_{dir}, dir_l)}) \vee \\ & \vee (eq_4(z_{op}, U) \cdot \overline{eq_2(z_{dir}, dir_b)}))) \vee s_4(0, eq_4(z_{op}, BLK)) \vee \\ & \vee s_4(op(X, Z), ((eq_4(z_{op}, SRC) \vee eq_4(z_{op}, DEL) \vee eq_4(z_{op}, PRL)) \cdot eq_2(z_{dir}, dir_r)) \vee \\ & \vee ((eq_4(z_{op}, MAC) \vee eq_4(z_{op}, MAX) \vee eq_4(z_{op}, MIN) \vee eq_4(z_{op}, GAT)) \cdot eq_2(z_{dir}, dir_l)) \vee \\ & \vee (eq_4(z_{op}, U) \cdot eq_2(z_{dir}, dir_b))) \end{aligned} \quad (1)$$

описывает значение на левом выходе ВЭ в зависимости от его входов и настройки. При этом используется вспомогательная функция равенства n-разрядных двоичных чисел

$$eq_n(a_n, b_n) = \wedge_i^n eq_1(a^i, b^i) = \wedge_i^n \left((a^i \cdot b^i) \vee \overline{(a^i \vee b^i)} \right) \quad (2)$$

и функция выбора

$$s_n(x, st) = (x^n \cdot st,\ x^{n-1} \cdot st\ , \dots, x^1 \cdot st), \quad (3)$$

где m^k обозначает k-й разряд двоичного числа m, отсчёт разрядов ведётся с единицы.

Таблица 1 – Табличная форма функции выходов ВЭ

z_{op}	z_{dir}	$Y = (y_l, y_t, y_r, y_b)$
TRS	$dir_l, dir_t, dir_r, dir_b$	(x_r, x_b, x_l, x_t)
SRC PRL DEL	dir_l	$(x_r, x_b, op(X,Z,S), x_t)$
	dir_t	$(x_r, x_b, x_l, op(X,Z,S))$
	dir_r	$(op(X,Z,S), x_b, x_l, x_t)$
	dir_b	$(x_r, op(X,Z,S), x_l, x_t)$
MAC MAX MIN GAT	dir_l	$(x_r, x_b, x_l, op(X,Z))$
	dir_t	$(\ op(X,Z), x_b, x_l, x_t)$
	dir_r	$(x_r, op(X,Z), x_l, x_t)$
	dir_b	$(x_r, x_b, op(X,Z), x_t)$
U	dir_l	$(x_r, U(X,Z), x_l, x_t)$
	dir_t	$(x_r, x_b, U(X,Z), x_t)$
	dir_r	$(x_r, x_b, x_l, U(X,Z))$
	dir_b	$(U(X,Z), x_b, x_l, x_t)$
BLK	$dir_l, dir_t, dir_r, dir_b$	$(0,0,0,0)$

Соединяя ВЭ последовательно, можно получить нейрон с требуемым количеством входов, а объединение нескольких таких цепочек образует полносвязный слой сети, как показано на рисунке 2. В правом углу ВЭ изображено хранимое в памяти значение. Символ w_{ij} обозначает вес j-й связи i-го нейрона; b_i – величина смещения i-го нейрона.

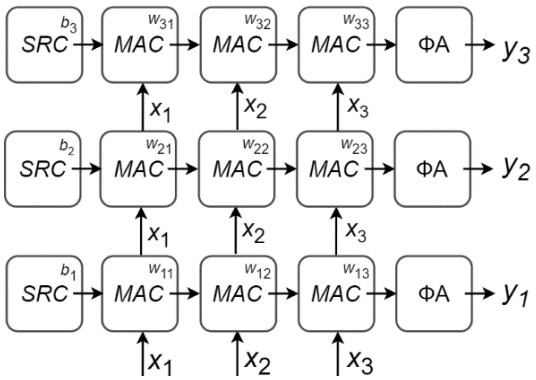
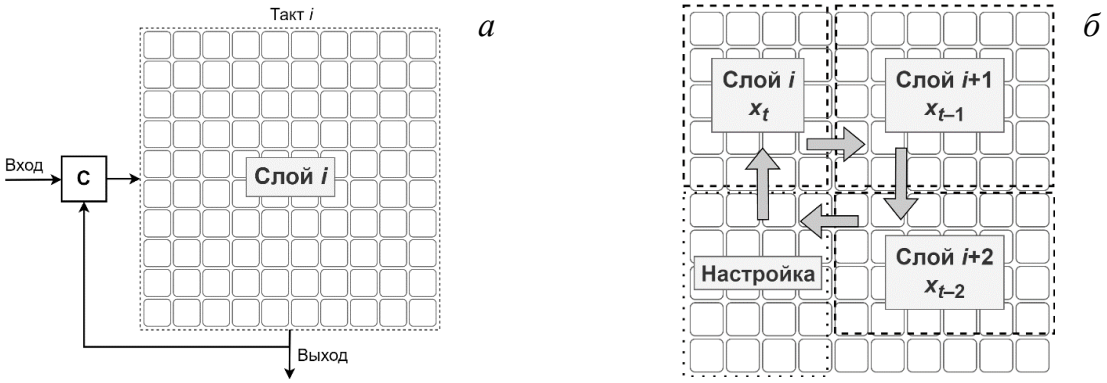


Рисунок 2 – Полносвязный слой из трёх нейронов на ПВС

В работе предлагаются интегральный и сегментированный режимы функционирования среды (рисунок 3). В интегральном вся среда реализует один слой сети, что позволяет поддерживать слои большого размера. Сегментированный режим обеспечивает быстродействие и энергоэффективность средних и малых слоёв благодаря применению скользящего окна перенастройки и конвейеризации. Его ключевым преимуществом является хранение промежуточных результатов на самой среде, что значительно увеличивает пропускную способность.

Для реализации на ПВС нейросетевых алгоритмов необходимо осуществить её настройку. В данной работе предлагается координатный метод на основе настроечной сетки, что обеспечивает гибкость и высокую скорость настройки. Коды настройки

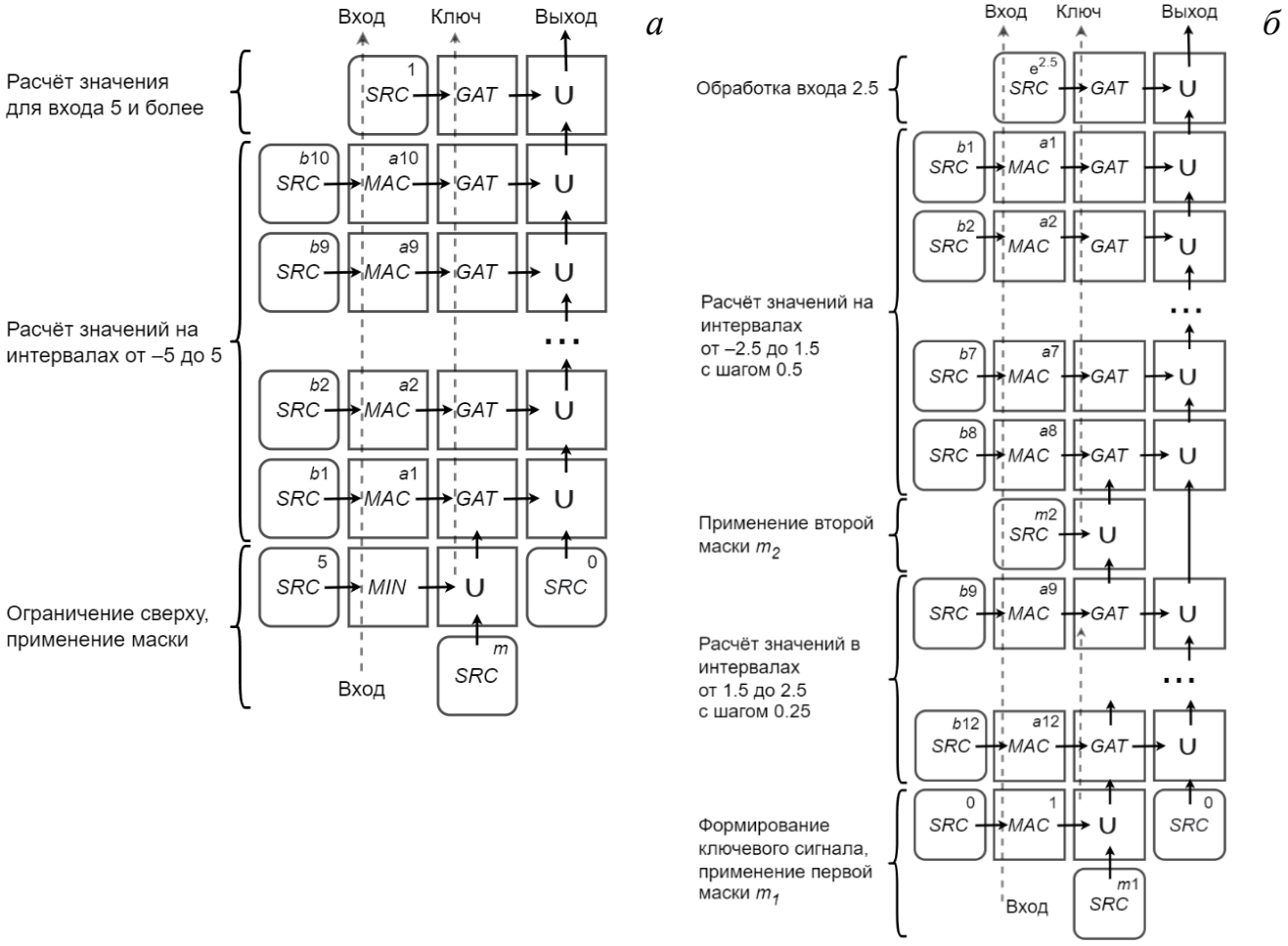
передаются по основным связям между ВЭ, что упрощает строение ПВС. Для передачи настроечного сигнала большой разрядности он разбивается на две части, передаваемые независимо по горизонтальным и вертикальным связям.



а – интегральный; б – сегментированный

Рисунок 3 – Режимы функционирования среды

В подразделе 2.2 описываются алгоритмы, обеспечивающие вычисление на среде функций активации нейрона: сигмоиды, гиперболического тангенса и softmax (рисунки 4-5). Их реализация основана на применении метода кусочно-линейной аппроксимации, что позволяет снизить вычислительную сложность и распараллелить вычисления по элементам среды, увеличив тем самым общее быстродействие.



а – сигмоида; б – реализация экспоненты для softmax

Рисунок 4 – Функций активации на ПВС

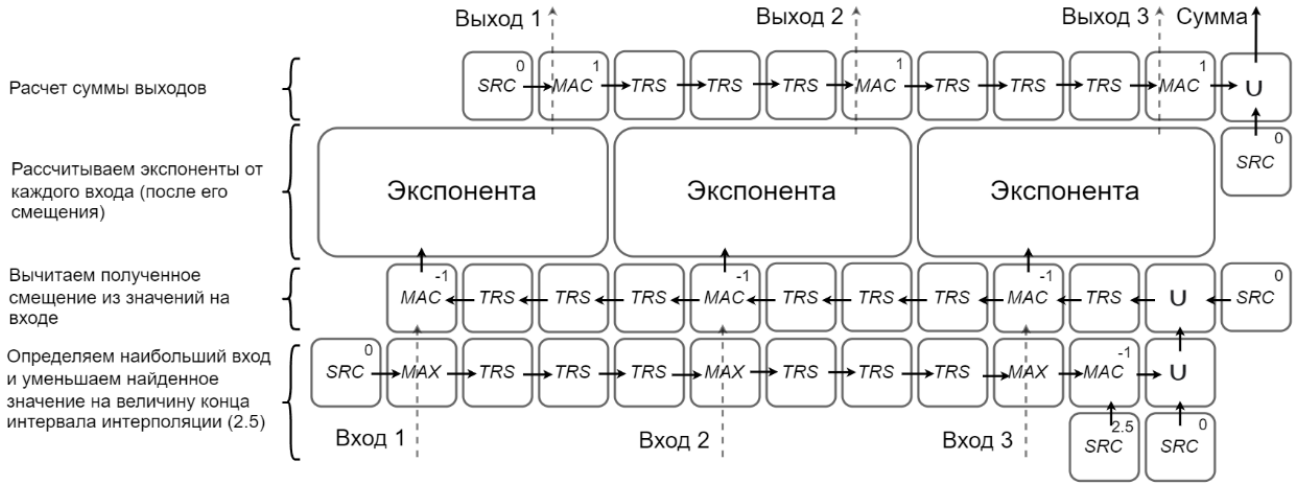
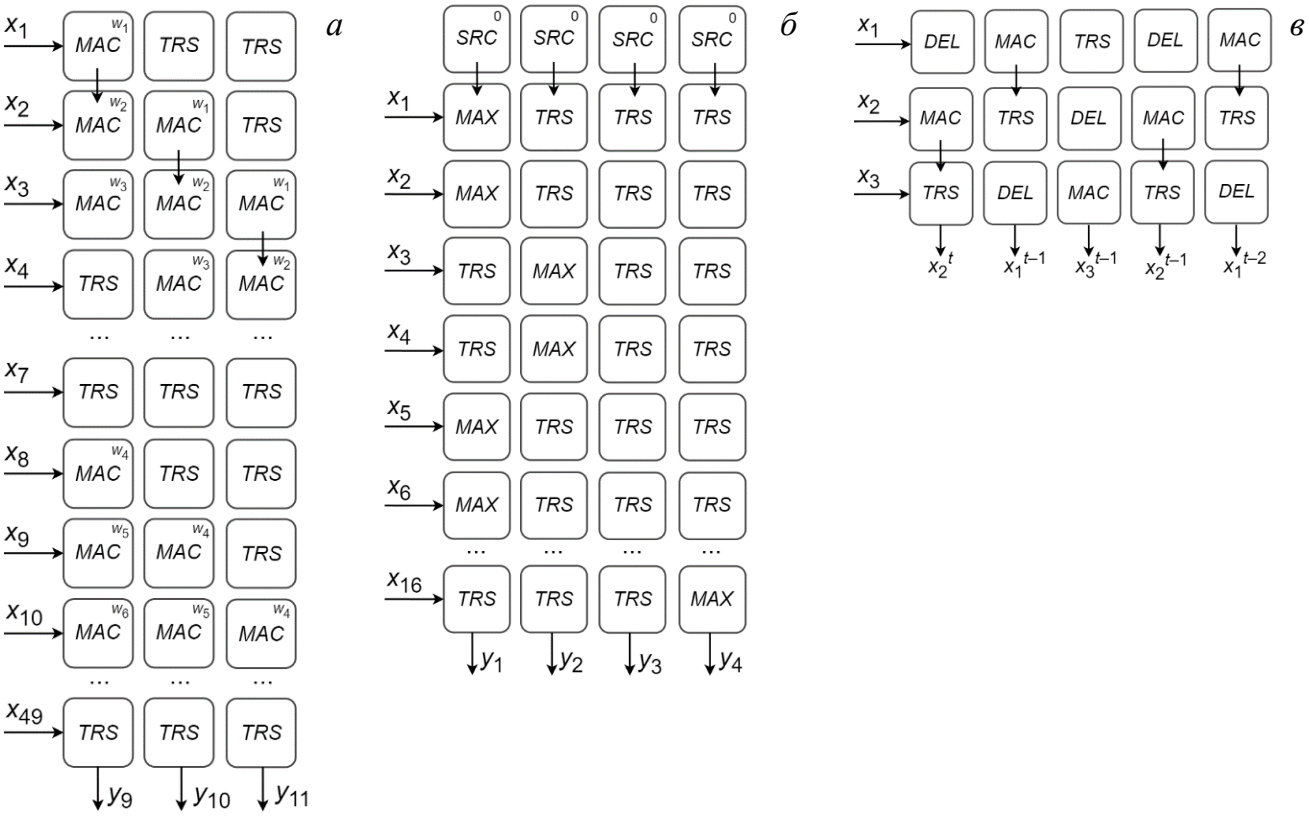


Рисунок 5 – Функция активации softmax с тремя входами на ПВС

Символы a_i, b_i на рисунке 4 обозначают параметры аппроксимации i-го интервала.

В подразделе 2.3 описано применение ПВС для свёрточных и рекуррентных нейронных сетей (рисунок 6). В частности, предложена реализация свёрточного слоя, слоя субдискретизации (pooling), слоя преобразования матрицы в вектор (flatten). Показано, как на ПВС могут быть реализованы рекуррентные сети на примере сети Хопфилда и сети долгой краткосрочной памяти (LSTM).

В подразделе 2.4 приведена типовая архитектура программно-аппаратного комплекса на основе предложенных в диссертационной работе моделей и алгоритмов.



а – свёрточный; б – субдискретизации; в – преобразования матриц в вектор

Рисунок 6 – Реализация слоёв свёрточной сети на ПВС

В третьем разделе описаны построенные в рамках исследования имитационные модели разработанных алгоритмов, а также полученные экспериментальные результаты. Моделирование осуществлялось в программном пакете MATLAB Simulink, а также в используемом для разработки на FPGA программном пакете Quartus и его инструменте временного анализа Timing Analyzer при помощи языка описания аппаратуры Verilog.

В подразделе 3.1 приводятся результаты имитационного моделирования предложенных вычислительных элементов среды в программном пакете MATLAB Simulink. Представлена тестовая модель для ручной проверки ВЭ при разных настроечных и входных сигналах. Показан результат применения сценария автоматического построения Simulink-модели ПВС требуемого размера при помощи встроенных в MATLAB инструментов. Испытания описанных в данном подразделе имитационных моделей подтвердили корректность разработанных алгоритмов.

В подразделе 3.2 приводятся результаты временной симуляции полносвязных слоёв разного размера. В ходе исследований была определена длительность вычисления нейрона при разном количестве входов (задержка обработки), а также длительность перемещения сигнала между нейронами одного слоя (задержка передачи). Эти величины позволяют оценить полное время обработки сигнала для полносвязного слоя на ПВС. Эксперименты показали, что расчёт нейрона с 25 входами занимает 27 нс, длительность передачи сигнала по 25 элементам среды составляет 12,8 нс. Иными словами, вычисление на ПВС полносвязного слоя из 25 нейронов с 25 входами каждый занимает около 40 нс. Полученные результаты показывают высокое быстродействие предложенных алгоритмов, хотя и уступают неперестраиваемым аналогам. Ключевым преимуществом предложенных моделей является способность динамически изменять реализуемые алгоритмы, что открывает широкие возможности для построения сложных многофункциональных устройств.

В подразделе 3.3 приводятся результаты временной симуляции предложенных реализаций функций активации нейрона – сигмоиды и softmax. Благодаря распределению алгоритма по коллективу вычислителей, параллельно-конвейерной архитектуре среды и низкоуровневым оптимизациям, предложенные реализации показывают высокое быстродействие, сравнимое с неперестраиваемыми аналогами, однако отличаются большой площадью вычислителя. Имитационная модель сигмоидной функции использует 13175 логических элементов FPGA, наибольшая задержка составляет 18,5 нс, а средняя абсолютная ошибка аппроксимации 4×10^{-3}. Полученные показатели близки к наиболее быстродействующим аналогам (10 нс), которые при этом не обладают способностью к перестраиваемости. Реализация функции softmax с тремя входами требует 54718 логических элементов (из них по 15069 на каждую экспоненту). Длительность расчёта составляет 43 нс. Сравнение с существующими аналогами показывает в шесть раз меньшую тактовую частоту (около 25 МГц) при соизмеримой пропускной способности (1,1 Гбит/с), что достигается благодаря параллельно-конвейерной архитектуре.

В четвёртом разделе показано применение предложенных алгоритмов на вычислительном устройстве на основе FPGA для решения классической задачи классификации ирисов Фишера.

В подразделе 4.1 описана подготовка модели нейронной сети для решения поставленной задачи. При помощи стандартных инструментов (Python, TensorFlow, Keras, Scikit-learn) была получена модель полносвязной сети прямого распространения с двумя скрытыми слоями (5 и 12 нейронов в каждом) с функцией активации ReLU, показавшая точность классификации 90%.

В подразделе 4.2 описана реализация разработанной модели НС на ПВС предложенной в данной работе архитектуры. Для этого потребовалась среда из 190 ВЭ, разделённая на четыре сегмента, функционирующих в одном такте.

В подразделе 4.3 показана реализация полученной модели ПВС на FPGA Cyclone V в составе системы на кристалле DE10-Nano. В связи с недостаточным количеством портов ввода/вывода, исходные данные и полученный н выходе среды результат сохранялись в памяти устройства. Эксперименты показали работоспособность и корректность предложенных алгоритмов. Полученные на выходе ПВС результаты совпадают с ожидаемыми с учётом потерь, связанных со снижением разрядности.

В заключении сформулированы основные результаты выполнения диссертационной работы, даны рекомендации по дальнейшим исследованиям.

В приложениях приведены акты внедрения результатов работы, блок-схемы разработанных алгоритмов, краткая информация по рассматриваемым функциям активации, параметры аппроксимации экспоненциальной функции на ПВС.

ЗАКЛЮЧЕНИЕ

Диссертационная работа является научно-квалификационной работой, в которой изложены новые научно обоснованные технические решения применения перестраиваемых вычислительных сред для реализации нейросетевых алгоритмов в программно-аппаратных комплексах для развития технологий искусственного интеллекта.

Ключевые результаты, полученные в данной работе:

1. Предложена концепция синтеза алгоритмов вычисления отклика нейронных сетей заданных архитектур для перестраиваемых вычислительных сред. Применение перестраиваемых сред обеспечивает параллельное вычисление всех нейронов одного слоя и динамическую реконфигурацию реализуемой сети на уровне отдельных нейронов. Синтезируемые алгоритмы ориентированы на низкоуровневую аппаратную реализацию, конвейеризацию вычислений и распределение операций по коллективу независимых вычислителей. Каждый нейрон реализуется строкой динамически перестраиваемых элементов среды, что обеспечивает высокую гибкость требуемых алгоритмов.

2. Разработан алгоритм вычисления отклика нейронных сетей полносвязных, свёрточных и некоторых рекуррентных архитектур на перестраиваемых вычислительных средах, обладающий рядом преимуществ: вычислением нескольких слоёв сети за один такт при помощи конвейеризации вычислений; уменьшенным внешним обменом

промежуточными результатами; возможностью точечной и групповой настройки элементов среды; поддержкой нейронных сетей с произвольным числом слоёв. Определены 10 операций, которые необходимо поддерживать элементам таких сред.

3. Предложены алгоритмы реализации распространённых слоёв (полносвязного, свёрточного, субдискретизации, преобразования матриц в вектор) и функций активации (сигмоиды, гиперболического тангенса, softmax) нейронных сетей на вычислительных средах. Разработанные алгоритмы выполняются за один такт работы среды и учитывают особенности систем на основе коллектива вычислителей. Сложные функции опираются на метод кусочно-линейной аппроксимации, который отличается вычислительной простотой и допускает распределение операций между элементами среды.

4. Проведён ряд экспериментов над разработанными алгоритмами при помощи имитационного моделирования и с применением испытательного стенда на основе программируемых логических интегральных схем. Для имитационного моделирования использовались программные пакеты MATLAB Simulink и Quartus. Результаты симуляций подтвердили работоспособность и высокое быстродействие предложенных моделей, соизмеримое с аналогами, лишёнными способности к перестраиваемости. Вычисление на предложенной модели среды функции активации softmax с тремя входами занимает 43 нс, а полносвязного слоя из 25 нейронов с 25 входами каждый – 40 нс. Негативной особенностью разработанных моделей является большая площадь на полупроводнике, что объясняется пространственным распределением алгоритмов и накладными расходами на обеспечение перестраиваемости вычислительной среды.

Автор выражает благодарность научному руководителю доктору технических наук С. В. Шидловскому за неоценимый вклад в выполнение научных исследований, связанных с написанием данной работы.

Автор благодарит кандидата технических наук Д. В. Шашева за оказанную им помощь в проведении исследований и анализе полученных результатов.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в журналах, включенных в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук:

1. **Шатравин В.** Применение вычислительных сред для ускорения рекуррентных нейронных сетей / В. Шатравин, Д. В. Шашев // Цифровая экономика. – 2023. – № 22 (1). – С. 27–35. – DOI: 10.34706/DE-2023-01-04. – 0,89 / 0,44 а.л.

2. **Шатравин В.** Разработка алгоритма настройки перестраиваемой вычислительной среды в составе аппаратного ускорителя искусственных нейронных сетей / В. Шатравин, Д. В. Шашев // Цифровая экономика. – 2022. – № 20 (4). – С. 11–18. – DOI: 10.34706/DE-2022-04-02. – 0,58 / 0,29 а.л.

3. **Shatravin V.** Sigmoid Activation Implementation for Neural Networks Hardware Accelerators Based on Reconfigurable Computing Environments for Low-Power Intelligent Systems / V. Shatravin, D. Shashev, S. Shidlovskiy // Applied Sciences (Switzerland). – 2022. – Vol. 1, № 10. – Article number 5216. – 16 p. – URL: https://www.mdpi.com/2076-3417/12/10/5216 – DOI: 10.3390/app12105216. – 1,15 / 0,38 а.л. *(Scopus).*

4. Шашев Д. В. Реализация сигмоидной функции активации с помощью концепции перестраиваемых вычислительных сред / Д. В. Шашев, **В. Шатравин** // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2022. – № 61. – С. 117-127. – DOI: 10.17223/19988605/61/12. – 0,86 / 0,43 а.л.

Web of Science: Shashev D. V. Implementation of the sigmoid activation function using the reconfigurable computing environments / D. V. Shashev, **V. V. Shatravin** // Vestnik Tomskogo Gosudarstvennogo Universiteta–Upravlenie Vychislitelnaja Tehnika I Informatika–Tomsk State University Journal Of Control And Computer Science. – 2022. – № 61. – P. 117-127.

Статьи в сборниках материалов конференций, представленных в изданиях, входящих в Scopus и Springer:

5. **Shatravin V.** Application of the Piecewise Linear Approximation Method in a Hardware Accelerators of a Neural Networks Based on a Reconfigurable Computing Environments / V. Shatravin, D. V. Shashev // Distributed Control and Communication Networks. – 2023. – Vol. 1748 : Communications in Computer and Information Science : selected papers of 25th International Conference, DCCN 2022. Moscow, Russia, September 26-29, 2022. – P. 63–74. – DOI: 10.1007/978-3-031-30648-8_6. – 0,63 / 0,32 а.л. *(Springer).*

6. **Shatravin V.** Applying the Reconfigurable Computing Environment Concept to the Deep Neural Network Accelerators Development / V. Shatravin, D. Shashev, S. Shidlovskiy // International Conference on Information Technology, ICIT 2021 : materials of International Conference. Amman, Jordan, July 14–15, 2021. – 2021. – P. 842–845. – DOI: 10.1109/ICIT52682.2021.9491771. – 0,44 / 0,15 а.л. *(Scopus).*

7. **Shatravin V.** Designing high performance, power-efficient, reconfigurable compute structures for specialized applications / V. Shatravin, D. V. Shashev // Journal of Physics: Conference Series. – 2020. – Vol. 1611 : materials of the XVII International Conference on Prospects of Fundamental Sciences Development. – Tomsk, Russian Federation, April 21–24, 2020. – Article number 012071. – 6 p. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/1611/1/012071. – DOI: 10.1088/1742-6596/1611/1/012071. – 0,41 / 0,21 а.л. *(Scopus).*

Свидетельства о государственной регистрации программ для ЭВМ:

8. Свидетельство о государственной регистрации программы для ЭВМ № RU 2022660840. Программа реализации перестраиваемой вычислительной среды для кусочно-линейной интерполяции экспоненциальной функции / Д. В. Шашев (RU), **В. В. Шатравин** (KZ); правообладатель: федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет» (RU). – Заявка № 2022660339, дата поступления – 09.06.2022; дата государственной регистрации в Реестре программ для ЭВМ – 10.06.2022.

9. Свидетельство о государственной регистрации программы для ЭВМ № RU 2021610860. Программная модель ячейки перестраиваемой вычислительной среды для реализации полносвязной искусственной нейронной сети / Д. В. Шашев (RU), **В. В. Шатравин** (KZ), С. И. Пославский (KZ); правообладатель: федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет» (RU). – Заявка № 2020667609; дата поступления – 28.12.2020; дата государственной регистрации в Реестре программ для ЭВМ – 19.01.2021.

Публикации в прочих научных изданиях:

10. **Шатравин В.** Разработка аппаратных ускорителей искусственных нейронных сетей на основе перестраиваемых вычислительных сред для интеллектуальных робототехнических систем / В. Шатравин, Д. В. Шашев, С. В. Шидловский // Перспективные системы и задачи управления : материалы XVII Всероссийской научно-практической конференции. Домбай, 04–08 апреля 2022 г. – Таганрог, 2022. – С. 173–179. – 0,44 / 0,15 а.л.

11. **Шатравин В.** Режимы функционирования многотактных перестраиваемых вычислительных сред в задачах машинного обучения / В. Шатравин, Д. В. Шашев // Интеллектуальные системы 4-й промышленной революции : сборник материалов IV Международного форума. Томск, 15–16 декабря 2021 г. – Томск, 2022. – С. 52–54. – 0,18 / 0,09 а.л.

12. **Shatravin V.** Developing of models of dynamically reconfigurable neural network accelerators based on homogeneous computing environments / V. Shatravin, D. V. Shashev, S. V. Shidlovskiy // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2021) : материалы XXIV Международной научной конференции. Москва, 20–24 сентября 2021 г. – 2021. – P. 102–107. – DOI: 10.25728/dccn.2021.015. – 0,34 / 0,11 а.л.

13. **Шатравин В. В.** Подготовка модели искусственной нейронной сети и модуля её обучения для реализации на ПЛИС / В. В. Шатравин, С. В. Шидловский // Интеллектуальные системы 4-й промышленной революции : сборник материалов III Международного форума. Томск, 26–27 ноября 2019 г. – Томск, 2020. – С. 122–126. – 0,19 / 0,10 а.л.

14. Шашев Д. В. Перестраиваемые вычислительные среды в задачах построения нейросетевых алгоритмов / Д. В. Шашев, **В. Шатравин** // Инноватика–2020 : сборник материалов XVI Международной школы-конференции студентов, аспирантов и молодых ученых. Томск, 23–25 апреля 2020 г. – Томск, 2020. – С. 84–87. – 0,14 / 0,07 а.л.

15. **Шатравин В.** Применение концепции перестраиваемых вычислительных сред в задачах построения новых архитектур искусственных нейронных сетей / В. Шатравин, Д. В. Шашев // Телекоммуникации. – 2020. – № 6. – С. 30–38. – 0,53 / 0,26 а.л.

16. **Шатравин В.** Моделирование полносвязной искусственной нейронной сети в среде MATLAB / В. Шатравин // Инноватика–2019 : сборник материалов XV Международной школы-конференции студентов, аспирантов и молодых ученых. Томск, 25–27 апреля 2019 г. – Томск, 2019. – С. 441–445. – 0,19 а.л.

Издание подготовлено в авторской редакции.
Отпечатано на участке цифровой печати
Издательства Томского государственного университета
Заказ № 7551 от «6» сентября 2023 г. Тираж 100 экз.
г. Томск, Московский тр. 8, тел. 53-15-28, publish.tsu.ru