На правах рукописи

Лимонова Елена Евгеньевна

Биполярная морфологическая аппроксимация нейрона для уменьшения вычислительной сложности глубоких сверточных нейронных сетей

Специальность 1.2.2 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2022

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук в отделе №92.

Научный руководитель: **Арлазаров Владимир Викторович**
кандидат технических наук

Официальные оппоненты: **Соловьев Роман Александрович**
доктор технических наук, чл.-корр. РАН, ФГБУН Институт проблем проектирования в микроэлектронике РАН, главный научный сотрудник отдела методологии проектирования интегральных схем

Мясников Евгений Валерьевич
кандидат технических наук, ФГАОУ ВО «Самарский национальный исследовательский университет имени академика С.П. Королева» (Самарский университет), доцент кафедры геоинформатики и информационной безопасности

Ведущая организация: Федеральное государственное учреждение «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук»

Защита состоится 20 февраля 2023 г. в 11.00 на заседании диссертационного совета 24.1.224.01 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 117312, г. Москва, проспект 60-летия Октября, 9 (конференц-зал, 1-й этаж).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: г. Москва, ул. Вавилова, д. 40 и на официальном сайте ФИЦ ИУ РАН: http://www.frccsc.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 119333, г. Москва, ул. Вавилова, д. 44, кор. 2, ученому секретарю диссертационного совета 24.1.224.01.

Автореферат разослан « » _____ 20___ года.
Телефон для справок: +7 (499) 135-51-64.

Ученый секретарь
диссертационного совета 24.1.224.01,
канд. физ.-мат. наук, доцент И.В. Смирнов

Общая характеристика работы

Актуальность темы. Современные технологии нейросетевого распознавания используются в различных сферах жизнедеятельности человека. Они способны облегчить решение ряда прикладных задач, однако их внедрение ограничивается не только точностью распознавания и скоростью работы, но и соображениями безопасности и конфиденциальности данных пользователей. Именно эти вопросы выходят на первый план при распознавании идентификационных документов, банковских карт и платежных документов, обработке медицинской информации. Один из наиболее эффективных способов обеспечения безопасности пользовательских данных предлагает концепция граничных вычислений, при которой вычисления выполняются в точке, максимально приближенной к конечному пользователю, в идеале — на конечном устройстве, где эти результаты и будут использоваться далее.

Однако конечные устройства чаще всего обладают ограниченной вычислительной мощностью и объемом доступной оперативной памяти. Кроме того, повышенные требования предъявляются к их энергоэффективности, так как часто они работают от аккумулятора (например, смартфоны и различные носимые устройства) или являются составной частью других систем, также ограниченных в энергопотреблении (например, беспилотных транспортных средств или элементов интернета вещей). Также вопрос энергоэффективности нейросетевого распознавания в последнее время привлекает все большее внимание из-за возможного негативного влияния на экологию вследствие затрат энергии на обучение и многократные запуски глубоких нейронных сетей уже после их внедрения.

Таким образом, разработчикам распознающих систем с одной стороны необходимо обеспечить достаточно высокую точность распознавания для успешного решения поставленных задач, которая обычно достигается за счет усложнения нейросетевых моделей, а с другой — выполнить требования по энергоэффективности и скорости работы.

Особенно важной является эта задача в случае распознавания в режиме реального времени, например, при обработке видеопотока: прослеживании траекторий объектов, сегментации меняющейся сцены или извлечении текстовой информации в произвольных условиях.

Для визуального распознавания, как правило, используются модели, имеющие сверточную архитектуру, то есть состоящие из большого количества последовательно расположенных сверточных слоев, между которыми могут включаться слои субдискретизации, нормализации или слои других типов. Основную вычислительную сложность таких сетей составляют именно вычисления в сверточных слоях. Для обеспечения высокой точности распознавания такие модели выполняют несколько миллиардов

операций аккумулирующего умножения на запуск. Современные мощные вычислительные устройства имеют частоту в несколько гигагерц и несколько вычислительных ядер, однако даже они могут рассчитать всего несколько таких запусков в секунду. Таким образом, задача исследования вычислительной эффективности сверточных нейросетевых моделей на сегодняшний день крайне актуальна. В разное время ей занимались отечественные и зарубежные ученые, такие как Ю. И. Журавлев, В. Л. Арлазаров, В. А. Сойфер, Ю. В. Визильтер, И. Б. Гуревич, В. Б. Бетелин, Д. П. Николаев, а также H. Wen, M. Rastegari, A. Farhadi, Y. Lecun, Y. Bengio, G. Hinton и другие.

Повышение вычислительной эффективности таких моделей возможно из-за наличия в них неявной вычислительной избыточности. Исследования показывают, что эта избыточность в большей степени связана с несовершенством существующих методов обучения, а не конкретным числом нейронов и способом их организации в слои. Процесс обучения нейросетевых моделей заключается в поиске минимума некоторой функции потерь, которая в общем случае является невыпуклой и имеет множество экстремумов. С теоретической точки зрения такая задача крайне сложна и не имеет общего решения. Вследствие этого поиск методов снижения вычислительной избыточности нейросетевых моделей носит экспериментальный характер. Есть множество методов, снижающих число тех или иных арифметических операций в нейросетевых моделях, таких как тензорные декомпозиции, обрезка моделей, применение дистилляции знаний для создания более простых моделей. Эти методы позволяют в разы или даже на порядки снизить число операций, однако все еще не позволяют достичь желаемой вычислительной эффективности глубоких нейросетевых моделей при сохранении удовлетворительной точности распознавания.

Одним из наиболее перспективных направлений повышения вычислительной эффективности нейросетевых моделей последнего времени является совместная оптимизация архитектуры нейросетевой модели и архитектуры вычислительного устройства. Оно требует высокой квалификации специалиста как в области искусственного интеллекта, так и в области проектирования вычислительных устройств, или создания междисциплинарной команды ученых. Ведь разработчики нейросетевых моделей, ограниченные конкретным вычислительным устройством или классом устройств, вынуждены проектировать модели, опираясь на доступный объем вычислительных ресурсов. Как правило, современные нейросетевые модели направлены на исполнение на графических процессорах. С другой стороны, разработчики специализированных устройств выполняют большую работу по низкоуровневому проектированию и обычно рассматривают лишь одну нейросетевую архитектуру, позволяющую получить высокую точность распознавания. Результирующее устройство

при этом отличается высокой эффективностью, но может требовать модификации при малейших изменениях модели. В качестве компромисса были созданы специализированные тензорные процессоры (например, Google TPU или Intel VPU), которые могут эффективно исполнять отдельные классы нейросетевых моделей. Однако они также потребуют модификации при создании новых классов распознающих архитектур, например, в настоящее время они не поддерживают исполнение моделей с бинарными или тернарными весами.

В таких условиях особый интерес представляет смена модели вычислений в элементарных логических элементах нейронной сети — отдельных слоях или отдельных нейронах. Такие изменения не затрагивают архитектуру сети и все также позволяют строить и использовать модели разных типов, но могут сделать аппаратную реализацию модели гораздо эффективнее, поскольку различные типы нейронов требуют разных аппаратных и энергетических затрат при реализации и в процессе работы. Поскольку существующие модели слоев и нейронов уже доказали свою эффективность в решении практических задач и позволяют добиться высокой точности распознавания, данная работа посвящена исследованию их аппроксимаций, упрощающих последующее создание вычислителя, но при этом сохраняющих высокую точность работы.

Основные результаты диссертации были получены в процессе выполнения работ по следующим научным грантам РФФИ:

1. 18-07-01384 — «Исследование применимости методов нелинейных аппроксимаций для оптимизации быстродействия искусственных нейронных сетей на современнных микропроцессорных архитектурах»
2. 17-29-03297 — «Исследование возможности создания энергоэффективных аппаратных устройств для мобильных устройств комплексов идентификации и верификации личности в составе систем технического зрения наземных робототехнических комплексов»
3. 17-29-03240 — «Глубокие нейронные сети с вычислительно упрощенной моделью нейрона»

Целью данной работы является разработка и исследование вычислительно-эффективных аппроксимаций нейросетевых моделей, методов их обучения и оптимизации их вычисления на существующих и перспективных вычислителях.

Для достижения этой цели были поставлены следующие **задачи**:

1. Разработать метод аппроксимации вычислительно-интенсивных частей нейросетевых моделей, исследовать его вычислительную эффективность и точность.
2. Оценить вычислительную эффективность на различных платформах.

3. Разработать методы обучения предложенной аппроксимирующей структуры.
4. Провести экспериментальную оценку точности предложенного метода обучения аппроксимированных нейросетевых моделей для различных нейросетевых архитектур.
5. Разработать комплекс программ, позволяющий моделировать аппроксимацию нейросетевых моделей, обучение полученных структур и проверку результирующего качества работы.

Научная новизна:

1. Предложена новая аппроксимация классического нейрона нейроном с морфологической структурой, позволяющая создавать глубокие нейронные сети с морфологическими слоями и обеспечивающая высокую точность распознавания.
2. Предложен новый метод обучения произвольных, в том числе биполярных морфологических и целочисленных, аппроксимаций классических нейросетевых моделей путем послойного преобразования и дообучения, позволяющий повысить их качество.
3. Впервые показано, что для предложенной аппроксимации метод послойного преобразования и дообучения позволяет добиться более высокого качества работы нейросетевой модели, чем прямое обучение с помощью метода обратного распространения ошибки и градиентных методов оптимизации.
4. Проведено оригинальное исследование точностных характеристик нейросетевых моделей LeNet- и ResNet-подобных архитектур, использующих предложенную морфологическую аппроксимацию.
5. Впервые теоретически показано, что нейросетевая модель с достаточным числом нейронов биполярного морфологического вида может приблизить произвольную непрерывную на компакте функцию с любой заранее заданной точностью.

Практическая значимость. Предложенная аппроксимация позволяют создать нейросетевые модели, подобные по архитектуре классическим глубоким моделям, но в то же время обладающие принципиально новыми теоретическими свойствами. Она снижает вычислительную сложность исходных моделей и потенциально способна повысить их эффективность.

Разработанные в рамках диссертации методы были реализованы в виде программных компонентов и внедрены в программное обеспечение «Smart ID Engine», «Smart Code Engine», «Smart Document Engine», а также «Smart IDReader» компании ООО «Смарт Энджинс Сервис». Данные продукты интегрированы в информационную инфраструктуру и мобильные приложения АО «Тинькофф Банк», а также в ряд информационных решений государственных структур Российской Федерации. Кроме того, полученные оценки и результаты моделирования демонстрируют,

что включение специализированных модулей для элементарных арифметических операций при создании устройств для исполнения нейросетевых моделей способно повысить эффективность их работы и используются в АО «МЦСТ» при проектировании новых устройств.

Соответствие диссертации паспорту научной специальности. В соответствии с формулой специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ» (технические науки) в работе выполнены разработка, исследование и реализация модели вычислительно-эффективного биполярного морфологического нейрона как аппроксимации классического математического нейрона. Работа соответствует следующим пунктам паспорта специальности: п. 2 «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий», п. 3 «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента», п. 7 «Качественные или аналитические методы исследования математических моделей (технические науки)» и п. 9 «Постановка и проведение численных экспериментов, статистический анализ их результатов, в том числе с применением современных компьютерных технологий (технические науки)».

Методология и методы исследования. В диссертационной работе использовались методы математического анализа, линейной алгебры, методы численного моделирования и нелинейной теории оптимизации.

Основные положения, выносимые на защиту:

1. Разработана аппроксимация модели математического нейрона и сверточного слоя: биполярные морфологические нейрон и сверточный слой, не задействующие умножений в своих вычислительно-интенсивных частях.
2. Доказано, что нейронная сеть из биполярных морфологических нейронов может с любой заранее заданной точностью приблизить любую непрерывную на компакте функцию.
3. Предложен метод обучения аппроксимаций классических нейросетевых моделей путем послойного преобразования и дообучения, позволяющий повысить их качество.
4. Экспериментально показано, что предложенный метод послойного преобразования и дообучения позволяет добиться высокого качества работы аппроксимированных нейросетевых моделей на примере линейно квантованных малобитных и биполярных морфологических нейронных сетей.
5. Разработан комплекс программ, реализующий предложенную в работе модель биполярного морфологического нейрона, метод послойного дообучения для этой модели и позволяющий оценивать точностные характеристики результирующих сетей.

Достоверность полученных результатов подтверждается соответствием теоретических и экспериментальных результатов, продемонстрированных в работе, успешной апробацией результатов и внедрением в коммерческие системы распознавания документов.

Апробация работы. Основные результаты работы докладывались и обсуждались на следующих семинарах и конференциях:

1. Междисциплинарной школе-конференции Института проблем передачи информации им. А. А. Харкевича Российской академии наук (ИППИ РАН) «Информационные технологии и системы» (ИТиС) в 2015 году.
2. Международной конференции «International Conference on Machine Vision» (ICMV) в 2016, 2019, 2020 годах.
3. международной конференции «International Conference on Pattern Recognition» (ICPR) в 2020 году.
4. Научном семинаре Лаборатории №11 ИППИ РАН в 2021 году.
5. Международном научно-исследовательском семинаре «Анализ и понимание изображений (Математические, когнитивные и прикладные проблемы анализа изображений и сигналов)» в 2022 году.

Личный вклад. Все основные результаты диссертационной работы получены и обоснованы автором самостоятельно. Постановка задач и обсуждение результатов проводились совместно с научным руководителем. В [2—5] автором предложена аппроксимация классического нейрона морфологической структурой, методы для ее обучения, а также выполнено экспериментальное исследование ее точности для нейросетевых моделей различных архитектур, оценки вычислительной эффективности и выразительной способности предложенной структуры. В [6; 7] автор осуществил анализ вычислительной эффективности рассматриваемых алгоритмов на VLIW-платформах, выполнил их доработку и оценки производительности. В [10; 8] автор предложил методы квантования нейросетевых моделей, метод послойного преобразования и дообучения таких моделей и провел экспериментальную оценку их вычислительной эффективности. В [9] автору принадлежит идея разработанной аппаратной архитектуры и план проведения экспериментов. Исследование аппроксимаций функций активации биполярных морфологических моделей было выполнено и опубликовано в [1] без соавторства.

Публикации. Основные результаты по теме диссертации изложены в 10 печатных изданиях, из которых 1 работа издана в журнале, рекомендованном ВАК, 8 — в научных изданиях, индексируемых Web of Science и Scopus, 1 — в сборнике трудов конференции. Зарегистрировано 2 программы для ЭВМ.

Содержание работы

Во **введении** обсуждается тема диссертационного исследования, обосновывается ее актуальность и научная новизна, ставятся цели и задачи работы, а также показывается ее теоретическая и практическая значимость, и приводятся положения, выносимые на защиту. Введение завершается кратким содержанием диссертационной работы.

Первая глава начинается с рассмотрения нейросетевых моделей с вычислительной точки зрения. В ней изложены модели отдельного нейрона, такие как классический математический, морфологический и спайковый нейроны, модели слоев нейронной сети, а также приведены примеры нейросетевых архитектур. Рассматриваются основные классы вычислительных устройств, используемых для вычисления нейросетевых моделей, их преимущества и недостатки. На практике это чаще всего бывают:

1. Высокопроизводительные серверы, получающие данные с удаленных устройств. В этом случае сложность нейросетевой модели не представляет проблемы, однако возникают задержки, связанные с каналами связи, а также вопросы к безопасности пересылаемых данных, которые могут содержать персональную, медицинскую или финансовую информацию, которая должна оставаться конфиденциальной.
2. Непосредственно конечные устройства; чаще всего это будут пользовательские рабочие станции, мобильные или встраиваемые устройства. Их использование не несет дополнительных угроз для конфиденциальности, однако осложняется ограниченностью вычислительных ресурсов подобных устройств.

Проблема повышения вычислительной эффективности нейросетевых моделей на конечных устройствах имеет большое научное и практическое значение и именно ей посвящена данная работа. Поэтому далее вводятся две модели конечных устройств и методики оценки вычислительной эффективности для них, а именно модель логической цепи, которая соответствует программируемым/специализированным логическим интегральным схемам (ПЛИС/СЛИС), и модель Single Instruction Multiple Data процессора, отвечающая центральным процессорам мобильных устройств и рабочих станций.

Оставшаяся часть первой главы посвящена анализу существующих методов снижения вычислительной трудоемкости нейросетевых моделей и их классификации. Эти методы можно разделить на две практически не пересекающиеся группы. Методы первой группы уменьшают число коэффициентов и вычислительных операций в сети, за счет чего она также может работать быстрее независимо от вида вычислителя. К ним

можно отнести тензорные разложения сверток, обрезку моделей, дистилляцию знаний. Методы второй группы ориентируются на возможности конкретных вычислителей и заменяют одни вычислительные операции и структуры другими, которые могут быть оптимальнее реализованы и требуют меньше времени для подсчета. Например, малобитные вычисления эффективны как для центральных процессоров, так и для логических интегральных схем, а альтернативные эффективно-вычислимые модели отдельных нейронов и слоев предназначены в основном для логических интегральных схем. При этом методы из разных групп могут комбинироваться и дополнять друг друга. Проведенный анализ показал, что постоянно усложняющиеся задачи и нейросетевые модели, используемые для их решения, а также соображения энергоэффективности, безопасности и качества обслуживания заставляют ученых искать все новые методы повышения вычислительной эффективности для конечных устройств и встраиваемых систем. Несмотря на то, что существуют достаточно эффективные решения частных задач, в общем случае они не позволяют достичь баланса между желаемым качеством распознавания и скоростью работы. Поэтому цель диссертационной работы актуальна, а задачи по разработке и исследованию вычислительно-эффективных аппроксимаций существующих нейросетевых моделей, методов их обучения и оптимизации их вычисления на существующих и перспективных вычислителях представляют научный и практический интерес.

Во второй главе была предложена новая структура нейрона, имеющая морфологический вид и являющаяся аппроксимацией математического нейрона: биполярный морфологический (БМ) нейрон. При этой аппроксимации положительные и отрицательные значения коэффициентов и входных сигналов нейрона рассматриваются отдельно, формируя отдельные вычислительные пути для каждой комбинации знаков коэффициентов и входов. Такая структура позволяет явным образом моделировать процессы возбуждения и торможения в нейроне.

Биполярный морфологический нейрон имеет следующий вид:

$$f_{BM}(\mathbf{x}, V, v) = \varphi\left(\exp\max_{j=1}^{N}(\ln x_j^+ + v_j^+) - \exp\max_{j=1}^{N}(\ln x_j^+ + v_j^-) - \right.$$
$$\left. - \exp\max_{j=1}^{N}(\ln x_j^- + v_j^+) + \exp\max_{j=1}^{N}(\ln x_j^- + v_j^-) + v_0\right),$$

$$x_j^+ = \max(x_j, 0), \qquad x_j^- = \max(-x_j, 0)$$

где $\mathbf{x}$ — вектор входных значений длины N, v^+, v^- — векторы весовых коэффициентов длины N, v_0 — смещение, $\varphi(\cdot)$ — нелинейная функция активации.

Структура БМ нейрона показана на рисунке 1. Функция max позволяет отбрасывать отрицательные значения и сформировать четыре ветки

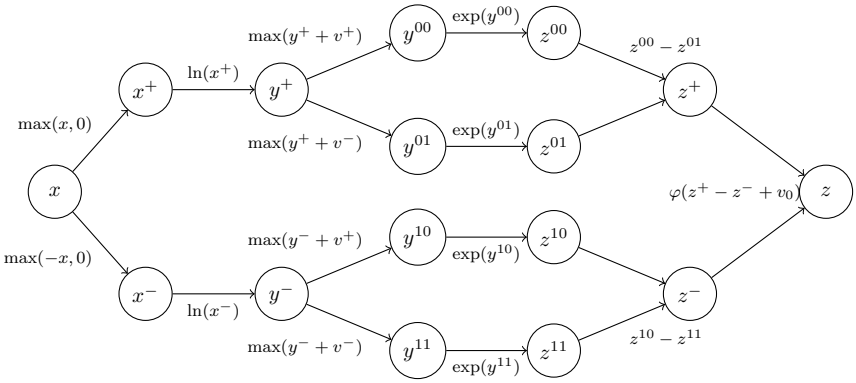


Рис. 1 — Структура БМ нейрона с вектором входных значений x, весовыми коэффициентами v^+, v^-, v_0 и вектором выходных значений z.

вычислений для различных комбинаций знаков входных значений и весовых коэффициентов. Затем выполняется логарифмирование и основная морфологическая операция внутри слоя. Ее результаты потенцируются и аккумулируются для получения выходного значения.

Далее для БМ сетей формулируется и доказывается теорема 1, демонстрирующая, что БМ сети имеют такую же выразительную способность, как и классические многослойные персептроны.

Определение 1. *Будем говорить, что функция $g(x)$ равномерно приближает $f(x)$ на компакте C с точностью $\epsilon > 0$, если*

$$\forall x \in C : |f(x) - g(x)| < \epsilon. \tag{1}$$

Теорема 1. *Любая непрерывная на компакте функция N переменных $f(x_1, ..., x_N)$ может быть равномерно приближена с любой заранее заданной точностью $\epsilon > 0$ некоторой нейронной сетью, состоящей только из БМ нейронов.*

Приведем схему доказательства. Рассмотрим одномерный случай, когда $f(x)$ определена и непрерывна на отрезке $[\alpha, \beta]$. Построим трехслойную нейронную сеть следующего вида.

1. Расположим на первом слое $2n$ нейронов, вычисляющие линейные функции
$$\xi_i^{\pm} = \pm x + a_i^{\pm}, \quad i = 1, \ldots, n,$$
где $a_i^+, a_i^- \in \mathbb{R}$ — произвольные действительные числа.
2. Расположим на втором слое n нейронов, вычисляющих прямоугольные импульсы η_i с основаниями $(a_i^-, -a_i^+)$ высоты 1.

3. Введем последовательность $x_0 = \alpha$, $x_1 = \alpha + \delta, ..., x_n = \beta$. Возьмем $a_i^+ = -x_i$, $a_i^- = x_{i-1}$, где $i = 1, ..., n$. Расположим на третьем слое один нейрон, вычисляющий кусочно-постоянную функцию $\zeta(x) = \max\limits_{i=1}^{n} f(x_i)\eta_i$.
4. Докажем, что можно подобрать такие n и δ, что $\zeta(x)$ равномерно приближает $f(x)$ на отрезке $[\alpha, \beta]$ с заранее заданной точностью.

Благодаря тому, что БМ нейрон является аппроксимацией классического, он может применяться в произвольных нейросетевых архитектурах. Поэтому далее приводятся оценки числа арифметических операций и вычислительной эффективности БМ слоев. Для современных центральных процессоров использование БМ нейронов с вещественными коэффициентами не позволяет повысить эффективность вычислений, поскольку они не предусматривают эффективных модулей для реализации отдельных арифметических операций, однако БМ нейроны могут быть эффективно реализованы для ПЛИС или СЛИС. Для создания такой реализации использовались аппроксимации для функций активации в БМ нейроне: аппроксимация Митчелла для реализации логарифмирования и аппроксимация Шраудольфа для реализации потенцирования.

Логарифм, аппроксимированный методом Митчелла, можно задать следующим выражением:

$$\widehat{\log 2}(x) = \lfloor \log_2 x \rfloor + \frac{x - 2^{\lfloor \log_2 x \rfloor}}{2^{\lceil \log_2 x \rceil} - 2^{\lfloor \log_2 x \rfloor}}.$$

Экспоненту, аппроксимированную методом Шраудольфа, можно задать следующим выражением:

$$\widehat{\exp 2}(x) = float(ax + (b - c)),$$
$$a = 2^{23},\ b = 127 \cdot 2^{23},\ c = 486411,$$

а функция $float$ интерпретирует целочисленный аргумент как число, записанное в вещественном формате $binary32$.

Оценки числа вентилей и латентностей для реализации основных арифметических модулей БМ нейрона приведены в таблице 1. Для их получения было создано описание основных операций (сложение, взятие максимума, умножение) на языке Verilog HDL, и была синтезирована вентильная реализация устройств с помощью программного пакета Synopsys Design Compiler для 16 нм технологических библиотек. Оценки сложности и эффективности реализации двоичных логарифма и экспоненты сделаны на основе оценок для основных операций. Далее были промоделированы сверточные БМ слои и проведено сравнение с классическими слоями с помощью полученных вентильных сложностей и латентностей отдельных операций. Результаты моделирования для некоторых параметров сверточных слоев и аппроксимированных функций активации приведены

Таблица 1 — Оценка числа элементарных арифметических операций, логических вентилей и латентности (в тактах) для операций в БМ слоях.

Операция	#add	#max	#mul	Вент.	Лат.
add	1	0	0	2659	3
max	0	1	0	563	2
mul	0	0	1	3247	4
$\widehat{\log 2}$	1	0	0	2659	3
$\widehat{\exp 2}$	1	0	1	5906	7

в таблице 2. Они показывают, что для слоев с достаточно большим числом входных и выходных каналов БМ слои используют практически столько же вентилей, сколько и классические слои, однако имеют латентность на 30-40% ниже. При этом для сверточных слоев с достаточно малым числом входных каналов и размером фильтров 3×3 при использовании аппроксимированных функций активации латентность на 12-40% меньше, чем для классических слоев.

Изложенные в этой главе результаты опубликованы в [1-2; 4-5; 9].

Таблица 2 — Оценка отношения числа вентилей и латентности для классического и БМ сверточных слоев для двухветочной структуры слоя, где $K \times K$ – размер ядра свертки, F – число выходных, а C – входных каналов.

K	F	C	V_{std}/V_{BM}	L_{std}/L_{BM}
1	16	16	0.74	1.19
1	32	32	0.82	1.29
1	64	64	0.87	1.34
1	128	128	0.89	1.37
1	256	256	0.90	1.38
1	512	512	0.91	1.39
3	16	16	0.89	1.37
3	32	32	0.90	1.39
3	64	64	0.90	1.39
3	128	128	0.91	1.40
3	256	256	0.92	1.40
3	512	512	0.92	1.40

Третья глава посвящена исследованию точностных характеристик БМ нейронных сетей. В ней показано, что прямое преобразование к БМ виду и использование классических методов обучения неэффективно, и предлагается оригинальный метод послойного преобразования и дообучения, использующий аппроксимационную природу БМ нейрона. Он опирается на два соображения:

1. Классические нейронные сети способны легко адаптироваться к небольшим изменениям входных сигналов при обучении, а значит можно постепенно заменять классические нейроны на БМ и выполнять обучение классической части сети.
2. БМ нейрон аппроксимирует классический нейрон, а значит значения коэффициентов классического нейрона можно использовать в качестве начальных значений при обучении БМ модели.

В качестве части нейронной сети, которая будет приближаться на каждой итерации, рассматривается элементарный блок современных нейросетевых архитектур: один слой. Таким образом, в предложенном методе сначала обучается классическая нейронная сеть, а затем выполняется преобразование нейронов каждого слоя к БМ модели, аппроксимируя весовые коэффициенты классического слоя. Далее полученная нейронная сеть дообучается классическими методами, что позволяет нивелировать падение качества. Преобразование выполняется послойно от первого к последнему слою. Подробно данный подход показан в Методе 1.

Метод 1: Обучение БМ сети

Входные данные: Обучающая выборка, валидационная выборка.
Выходные данные: БМ нейронная сеть.

1 Обучить классическую нейронную сеть стандартными методами.
2 **для всех** *сверточных и полносвязных слоев* **выполнить**
3 Заменить классические математические нейроны с весовыми коэффициентами w БМ нейронами с весовыми коэффициентами $\{v^+, v^-, v_0\}$, где:

$$v_j^+ = \begin{cases} \ln w_j, & \text{если } w_j > 0, \\ -\infty, & \text{иначе,} \end{cases}$$

$$v_j^- = \begin{cases} \ln |w_j|, & \text{если } w_j < 0, \\ -\infty, & \text{иначе,} \end{cases}$$

$$v_0 = w_0.$$

4 Обучить полученную нейронную сеть стандартными методами.

Экспериментально показано, что наилучших результатов можно достичь, если на этапе дообучения выполнять дообучение всей сети, а не только ее классической части. При этом в случае сверточных нейросетевых моделей, следует преобразовывать к БМ виду только сверточные слои. В результате показано, что для LeNet-подобных моделей на выборке MNIST точность распознавания сопоставима с точностью классических моделей.

Кроме того, предложенный метод послойного преобразования и дообучения может успешно использоваться и с другими аппроксимациями, например, малобитными целочисленными аппроксимациями нейросетевых моделей. Такие аппроксимации являются дискретными и не могут быть обучены с помощью классических градиентных методов, предполагающих непрерывность оптимизируемых параметров. В случае использования 8-битных целочисленных коэффициентов, представляющих вещественные числа с фиксированной точкой, и 16-битных аккумуляторов, применение метода послойного преобразования и дообучения позволило обеспечить точность распознавания 98.7% при преобразовании сверточных слоев модели, в то время как точность преобразованной модели без дообучения составила лишь 48.6%. Время работы преобразованной модели на мобильном процессоре архитектуры ARM снизилось на 20%.

Таким образом, метод послойного преобразования и дообучения не ограничен БМ моделями и может успешно применяться в задачах распознавания, например, при обучении квантованных нейронных сетей, которые широко используются для повышения скорости работы реальных приложений.

Далее БМ модели были апробированы в двух основных категориях задач технического зрения, которые решаются с помощью нейросетевых моделей: задачах визуальной классификации и семантической сегментации. На практике сложность классифицирующих моделей может значительно варьироваться в зависимости от задачи. Поэтому было рассмотрено несколько задач, задействующих модели различной сложности. Первая из них это задача классификации символов машиночитаемой зоны (МЧЗ) паспортов на реальных данных с помощью LeNet-подобных моделей; сложности таких моделей вполне достаточно для обеспечения высокой точности классификации, и в то же время они достаточно вычислительно-эффективные для использования на мобильных и встраиваемых устройствах. Использованные нейросетевые архитектуры и результаты обучения приведены в таблице 3. Они демонстрируют, что нейросетевые модели с БМ сверточными слоями имеют высокую точность классификации и могут использоваться в практических задачах.

Следующие рассмотренные задачи – это классификация рукописных цифр выборки MNIST и объектов выборки CIFAR-10 с помощью глубоких моделей ResNet с 22 сверточными слоями. Такие модели являются достаточно глубокими, чтобы проиллюстрировать возможность использования БМ слоев в глубоких нейронных сетях, но в то же время достаточно вычислительно эффективны для использования на мобильных и встраиваемых устройствах.

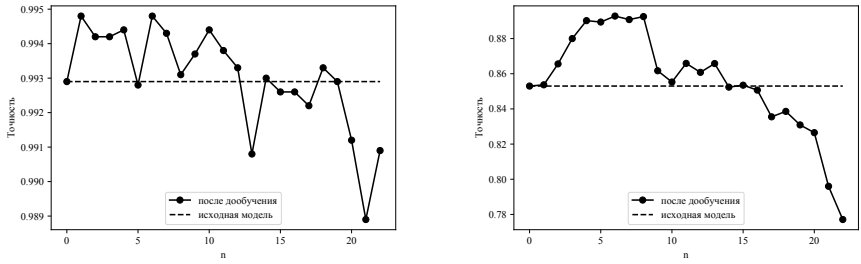
Точность модели в процессе дообучения на выборке MNIST проиллюстрирована на рисунке 2а. По горизонтальной оси отложено число преобразованных к БМ виду слоев, а по вертикальной — точность такой

Таблица 3 — Точность распознавания символов МЧЗ с помощью LeNet-подобных моделей, p_b – точность классической модели, p_{ft} – БМ модели, обученной с помощью предложенного метода.

Модель	Архитектура	p_b, %	p_{ft}
CNN_3	conv1(8, 3, 3) - relu1 - conv2(30, 5, 5) - relu2 - conv3(30, 5, 5) - relu3 - dropout1(0,25) - fc1(37) - softmax1	99.6	99.6
CNN_4	conv1(8, 3, 3) - relu1 - conv2(8, 5, 5) - relu2 - conv3(8, 3, 3) - relu3 - dropout1(0,25) - conv4(12, 5, 5) - relu4 - conv5(12, 3, 3) - relu5 - conv6(12, 1, 1) - relu6 - fc1(37) - softmax1	99.7	99.6

модели. Пунктирной линией показана точность классической сети 99.3%. Численные результаты обучения представлены в таблице 4. Можно видеть, что БМ модель с преобразованными и дообученными 19 слоями не уступает в точности классической, но дальнейшее преобразование снижает точность до 99.1%.

Точность модели на выборке CIFAR-10 при преобразовании и дообучении показана в таблице 4 и проиллюстрирована на рисунке 2б. По горизонтальной оси отложено число преобразованных к БМ виду слоев, а по вертикальной — точность такой модели. БМ модель с преобразованными и дообученными 16 слоями не уступает в точности классической, а при дальнейшем преобразовании точность постепенно снижается до 77.7% у полностью преобразованной модели. Пунктирной линией показана точность классической сети 85.3%. Такое снижение точности не слишком велико и результирующие модели пригодны для классификации. Кроме того, оно происходит постепенно с ростом числа преобразованных слоев.



а) б)

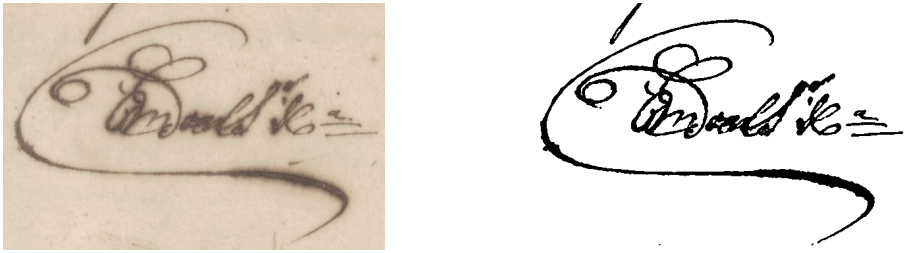
Рис. 2 — Точность классификации БМ ResNet после послойного преобразования и дообучения в зависимости от числа преобразованных слоев а) на выборке MNIST, б) на выборке CIFAR-10.

Таблица 4 — Точность распознавания MNIST и CIFAR-10 с помощью глубоких нейронных сетей на разных этапах послойного дообучения; p_b — после преобразования и до дообучения, p_{ft} — после дообучения.

Преобразовано слоев	Точность, %			
	MNIST		CIFAR-10	
	p_b	p_{ft}	p_b	p_{ft}
до преобразования	99.3		85.3	
1	61.3	99.5	13.7	85.4
2	99.3	99.4	70.1	86.6
3	94.8	99.4	56.5	88.0
4	99.4	99.4	87.2	89.0
5	14.6	99.3	9.9	88.9
6	98.7	99.5	43.2	89.3
7	73.5	99.4	23.6	89.1
8	98.9	99.3	63.8	89.3
9	94.3	99.4	28.9	86.2
10	91.6	99.4	30.2	85.5
11	99.2	99.4	85.3	86.6
12	95.9	99.3	13.4	86.1
13	80.2	99.1	35.5	86.6
14	67.0	99.3	12.7	85.2
15	49.5	99.3	11.1	85.4
16	80.3	99.3	22.3	85.1
17	11.4	99.2	9.2	83.6
18	73.2	99.3	45.4	83.9
19	91.3	99.3	11.9	83.1
20	59.6	99.1	17.2	82.7
21	11.4	98.9	20.7	79.6
22	85.2	99.1	33.7	77.7

Поэтому, чтобы полностью избавиться от него и сохранить преимущества БМ моделей, предлагается использовать гибридные (то есть частично преобразованные) модели.

Далее была рассмотрена задача семантической сегментации на примере задачи бинаризации исторических документов конкурса Document Image Binarization Competition (DIBCO) 2017. Примеры фрагментов изображений и эталонной бинаризации показаны на рисунке 3а, б. Для решения этой задачи в диссертационной работе используется модель U-Net, продемонстрировавшая лучший результат на конкурсе. Она показана на рисунке 4. Эта модель была преобразована к БМ виду методом послойного преобразования и дообучения. Фрагмент обработанного с помощью БМ U-Net изображения по сравнению с результатом работы классического U-Net показан на рисунке 3в, г.



а) б)

в) г)

Рис. 3 — Пример бинаризации: а) входное изображение, б) эталонное изображение, в) с помощью U-Net, г) с помощью БМ U-Net.

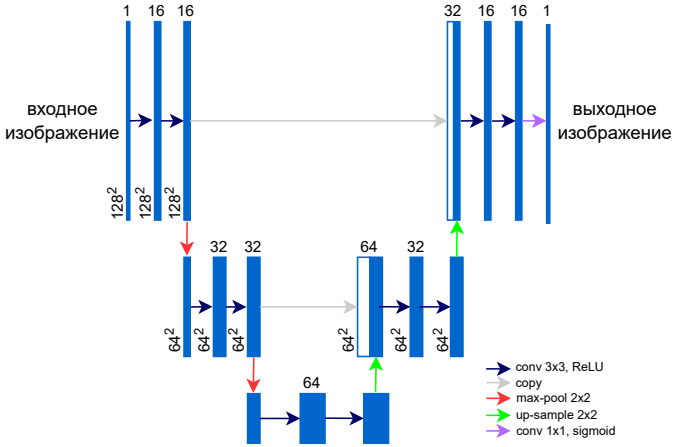


Рис. 4 — Нейросетевая архитектура U-Net, conv — сверточный слой, copy — копирование промежуточных результатов, max-pool — слой субдискретизации с операцией максимума, up-sample — сверточный слой, повышающий размерность, sigmoid — сигмоидальная функция активации.

Можно видеть, что с визуальной точки зрения качество бинаризации на разных участках изображения неоднородно: в некоторых зонах БМ U-Net демонстрирует лучший результат, а в некоторых – несколько проигрывает классической модели.

Количественная оценка качества бинаризации с помощью различных методов представлена в таблице 5. Методы Отсу и Сауволы не являются нейросетевыми и демонстрируют достаточно посредственное качество бинаризации по всем приведенным метрикам. Метод 17a, использующий глубокую полносветочную сеть, и 12, использующий ансамбль из трех глубоких сетей, заняли второе и третье место в конкурсе DIBCO 2017 соответственно. Можно видеть, что БМ U-Net уступает классической модели U-Net, однако все еще значительно превосходит методы Отсу и Сауволы. Кроме того, в достаточно широком диапазоне параметров он сопоставим по качеству с вторым и третьим решениями конкурса. При этом качество бинаризации достаточно равномерно снижается с ростом числа преобразованных сверточных слоев. Поэтому в этой задаче также можно использовать на практике гибридные модели, позволяющие достичь оптимального соотношения эффективности и точности.

Таким образом, полученные результаты показывают, что БМ нейроны могут успешно применяться в рассмотренных категориях задач технического зрения. Изложенные в третьей главе результаты опубликованы в [1-8; 10].

Таблица 5 — Сравнение качества различных методов бинаризации.

Метод	FM	Fps	PSNR	DRD
Отсу	77.7	77.9	13.9	15.5
Сауволa	77.1	84.1	14.3	8.9
17a (полносверточная глубокая сеть)	89.7	91.0	17.6	4.4
12 (ансамбль из 3 глубоких сетей)	89.4	91.5	17.6	3.6
U-Net	90.9	92.8	18.2	3.3
БМ-U-Net (10 БМ слоев)	85.8	88.0	17.0	5.1
БМ-U-Net (9 БМ слоев)	87.7	89.5	17.1	4.9
БМ-U-Net (8 БМ слоев)	87.2	89.4	17.3	4.7
БМ-U-Net (7 БМ слоев)	89.0	90.6	17.5	4.2
БМ-U-Net (6 БМ слоев)	88.2	90.2	17.4	4.5
БМ-U-Net (5 БМ слоев)	89.3	91.1	17.5	4.4
БМ-U-Net (4 БМ слоя)	90.5	92.1	18.0	3.6
БМ-U-Net (3 БМ слоя)	90.4	92.5	18.0	3.5
БМ-U-Net (2 БМ слоя)	90.4	92.6	18.0	3.5
БМ-U-Net (1 БМ слой)	90.9	92.4	18.0	3.4

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработана аппроксимация модели математического нейрона – биполярный морфологический нейрон, – которая может применяться в сверточных и полносвязных слоях нейросетевых моделей для упрощения их внутренней структуры. При такой аппроксимации в вычислительно-интенсивных частях слоя остаются лишь операции взятия максимума и сложения, однако слой дополняется функциями активации на основе операций потенцирования и логарифмирования.
2. Аналитическими методами доказано, что нейросетевая модель с достаточным числом нейронов биполярного морфологического вида может приблизить произвольную непрерывную на компакте функцию с любой заранее заданной точностью. Это означает, что биполярные морфологические нейронные сети имеют ту же выразительную способность, что и классические модели.
3. Вычислительно-емкие сверточные биполярные морфологические слои могут быть эффективно реализованы для ПЛИС/СЛИС. Оценка числа вентилей и латентности ПЛИС-реализации для БМ сверточных слоев по сравнению с классическими сверточными слоями показала, что для слоев с достаточно большим числом входных и выходных каналов БМ слои используют практически столько же вентилей, сколько и классические слои, однако имеют латентность на 30-40% ниже; для слоев с достаточно малым числом входных каналов и размером фильтров 3×3 при использовании аппроксимированных функций активации латентность на 12-40% меньше, чем для классических слоев;
4. Для обучения аппроксимированных нейросетевых моделей предложен оригинальный метод послойного дообучения, позволивший получить лучшее качество по сравнению с обучением стандартными методами для биполярных морфологических нейросетевых моделей и квантованных 8-битных нейросетевых моделей по результатам численных экспериментов.
5. Вычислительным экспериментом показано, что биполярная морфологическая аппроксимация сверточных слоев позволяет снизить вычислительную избыточность глубоких нейросетевых моделей в задачах классификации изображений и семантической сегментации без снижения качества распознавания для гибридных моделей и ряда полностью преобразованных моделей.
6. Разработан комплекс программ, позволяющий выполнить послойную и обучение аппроксимацию классической модели: обучить классическую модель, выполнить послойное преобразование к БМ виду, провести дообучение и оценить результирующее качество.

Разработанные в рамках диссертации методы были реализованы в виде программных компонентов и внедрены в программное обеспечение «Smart ID Engine», «Smart Code Engine», «Smart Document Engine», а также «Smart IDReader» компании ООО «Смарт Энджинс Сервис». Данные продукты интегрированы в информационную инфраструктуру и мобильные приложения АО «Тинькофф Банк», а также в ряд информационных решений государственных структур Российской Федерации. Кроме того, полученные оценки и результаты моделирования демонстрируют, что включение специализированных модулей для элементарных арифметических операций при создании устройств для исполнения нейросетевых моделей способно повысить эффективность их работы и используются в АО «МЦСТ» при проектировании новых устройств.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

1. *Limonova E. E.* Fast and gate-efficient approximated activations for bipolar morphological neural networks // Информационные технологии и вычислительные системы. — 2022. — № 2. — С. 3—10.

В изданиях, входящих в международные базы цитирования Scopus и Web of Science

2. Bipolar Morphological Neural Networks: Gate-Efficient Architecture for Computer Vision / E. E. Limonova [и др.] // IEEE Access. — 2021. — Т. 9. — С. 97569—97581.
3. *Limonova E.*, *Nikolaev D.*, *Arlazarov V. V.* Bipolar Morphological U-Net for Document Binarization // ICMV 2020. Т. 11605. — International Society for Optics, Photonics, 2021. — С. 1—9.
4. ResNet-like Architecture with Low Hardware Requirements / E. E. Limonova [и др.] // ICPR 2020. — IEEE. 2021. — С. 6204—6211.
5. Bipolar morphological neural networks: convolution without multiplication / E. Limonova [и др.] // ICMV 2019. Т. 11433. — International Society for Optics, Photonics, 2020. — С. 1—8.
6. *Limonova E. E.*, *Neyman-Zade M. I.-O.*, *Arlazarov V. L.* Special aspects of matrix operation implementations for low-precision neural network model on the Elbrus platform // Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software. — 2020. — Т. 13, № 1. — С. 118—128.

7. Performance Evaluation of a Recognition System on the VLIW Architecture by the Example of the Elbrus Platform / E. E. Limonova [и др.] // Programming and Computer Software. — 2019. — Т. 45, № 1. — С. 12—17.

8. Fast Integer Approximations In Convolutional Neural Networks Using Layer-By-Layer Training / D. Ilin, E. Limonova, V. Arlazarov, D. Nikolaev // ICMV 2016. Т. 10341. — International Society for Optics, Photonics, 2017. — С. 1—5.

9. *Tsoy M. O.*, *Alfonso D. M.*, *Limonova E. E.* Hardware Implementation of Classical and Bipolar Morphological Models for Convolutional Neural Network // En&T-2021. — IEEE. 2022. — С. 1—5.

В сборниках трудов конференций

10. *Николаев Д.*, *Лимонова Е.*, *Ильин Д.* Ускорение нейросетевого распознавания образов на SIMD архитектурах // 39-я междисциплинарная школа-конференция ИТиС 2015. — ИППИ РАН, 2015. — С. 472—483.

Зарегистрированные программы для ЭВМ

11. Программа для распознавания идентификационных карт личности «Smart IDReader»: свидетельство о государственной регистрации программ для ЭВМ № 2016616961, опубл. 22.06.2016 по заявке № 2016612014 от 01.03.2016 / В. В. Арлазаров [и др.]. — 2016.

12. *Лимонова Е. Е.* Программа для обучения сверточных биполярных морфологических нейронных сетей: свидетельство о государственной регистрации программ для ЭВМ №2022660269, опубл. 01.06.2022 по заявке №2022618573 от 05.05.2022. — 2022.

Лимонова Елена Евгеньевна

Биполярная морфологическая аппроксимация нейрона для уменьшения вычислительной сложности глубоких сверточных нейронных сетей

Автореф. дис. на соискание ученой степени канд. техн. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____