На правах рукописи

Гончаренко Александр Игоревич

ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ НЕЙРОННЫЕ СЕТИ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ УСТРОЙСТВ С НИЗКИМИ ВЫЧИСЛИТЕЛЬНЫМИ РЕСУРСАМИ

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

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Новосибирск — 2023

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

Научный руководитель:	доктор технических наук, ведущий научный сотрудник ИАиЭ СО РАН **Нежевенко Евгений Семёнович**
Официальные оппоненты:	**Оселедец Иван Валерьевич,** доктор физико-математических наук, профессор, АНО «Институт искусственного интеллекта», Генеральный директор
	Куликов Виктор Александрович, кандидат технических наук, Picsart Inc., Senior Machine Learning Scientist
Ведущая организация:	Новосибирский государственный технический университет

Защита состоится xxx.

С диссертацией можно ознакомиться в библиотеке ИАиЭ СО РАН и на сайте: https://www.iae.nsk.su/ru/dissertationcouncil.

Автореферат разослан 33 марта 2023 года.

Ученый секретарь
диссертационного совета
24.1.028.01 (Д 003.005.02),
д.ф.-м.н. Ильичёв Л.В.

Общая характеристика работы

Актуальность темы. Нейронные сети достигли значительного прогресса в решении различных практических задач. Однако, их применимость зачастую ограничена использованием мощных графических ускорителей, что не позволяет в полной мере распространить нейросетевые методы решения на мобильные платформы, например, телефоны, планшеты и программируемые логические интегральные схемы.

На сегодняшний день, для оптимизации вычислительной сложности нейронных сетей существуют методы, применение которых сопряжено с рядом трудностей таких как: требование к наличию качественной обучающей выборки, высоких вычислительных ресурсов на тонкую настройку сети, а также снижение скорости разработки итоговых решений, где нейросеть представляет лишь часть программной или программно-аппаратной архитектуры.

Целью диссертационной работы является разработка методов для ускорения вычисления нейронных сетей на мобильных платформах, не требующих от применяющего большого количества размеченных данных и вычислительных ресурсов для тонкой настройки сети, но при этом не допускающие значительного падения качества.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Проанализировать существующие архитектуры нейронных сетей с целью определения возможности вычисления на мобильных платформах;
2. Исследовать существующие методы ускорения нейронных сетей;
3. На основе существующих решений разработать собственный метод ускорения нейронных сетей с незначительным падением точности относительно оригинальной архитектуры и не требующий размеченных данных для тонкой настройки;
4. Экспериментально подтвердить эффективность разработанного метода;

5. Исследовать возможность применения нестандартных типов данных с плавающей запятой;
6. Экспериментально найти подходящие разрядности порядка и мантиссы, не снижающие качественные показатели сетей;

Научная новизна:

1. Предложен и реализован новый алгоритм квантования для моделей произвольного типа на основе тонкой настройки масштабирующих коэффициентов для порогов квантования. При этом время, затраченное на тонкую настройку сети, после применения алгоритма значительно ниже (от 5 до 10 раз, в зависимости от архитектуры нейронной сети), чем в большинстве современных работ в данной области, при незначительном падении точности (менее 1%) относительно оригинальной модели;
2. Разработан и применен алгоритм перемасштабирования весовых коэффициентов для процедуры скалярного квантования для ограниченной функции активации ReLU6, наиболее распространенной в мобильных архитектурах нейронных сетей;
3. Предложена и реализована процедура нахождения разрядности для специализированных типов данных. Предложенный механизм не требует дополнительной тонкой настройки сети, что позволяет упростить внедрение нейронных сетей в сложные программно-аппаратные комплексы. Данная процедура применялась к разнообразным архитектурам сверточных нейронных сетей, что может свидетельствовать о ее универсальности;
4. Предложенный подход нахождения разрядности для специализированных типов данных впервые применен к глубокой рекуррентной сети, что может свидетельствовать о его универсальности и позволяет его использовать в программно-аппаратных комплексах на программируемых интегральных схемах для проектирования аппаратных нейросетевых ускорителей.

Теоретическая значимость полученных результатов заключается в предложенном в диссертации новом подходе устранения влияния выбросов посредством обучения порогов квантования, что ведет к уменьшению шага дискретизации при квантовании и, как следствие, снижению количества ошибок, возникающих в нейронной сети. Предложенный подход был теоретически проанализирован при помощи метода обратного распространения ошибки.

Практическая значимость полученных результатов заключается в реализации разработанного подхода в виде программного комплекса для обучения свёрточных нейронных сетей на ЭВМ. Получаемые таким образом нейросетевые модели могут использоваться для произвольных задач на маломощных вычислителях. В частности, описанный в данной диссертации подход позволил ускорить алгоритм детектирования лица на конечном устройстве пользователя с мобильным ARM процессором. Также разработанный алгоритм стал частью программной платформы EENNT, позволяющей оптимизировать вычислительную сложность нейронных сетей произвольной архитектуры.

Процедура подбора разрядности порядка и мантиссы была использована как один из основных модулей для исследования оптимизации архитектуры аппаратных ускорителей на основе систолического массива.

Технология, основанная на разработанном подходе, внедрена в продуктах компаний:

1. “ООО Диалоговые системы”. Разработанный подход применялся для ускорения нейронной сети для распознавания речи в программной платформе для создания интеллектуальных диалоговых агентов “W11”;
2. “ООО Экспасофт”. Разработанные методы легли в основу программного комплекса ускорения нейронных сетей “Expasoft Embedded Neural Network Technology” или “EENNT”, имеющий свидетельство о государственной регистрации программы для ЭВМ и включенный реестр отечественного ПО с номером реестровой записи 8738;

3. “ООО ИВА ТЕХНОЛОДЖИС”. Разработанные подходы внедрены в программно-аппаратный комплекс “Микропроцессор IVA TPU”.

Основные положения, выносимые на защиту:

1. Использование масштабирующих коэффициентов для порогов квантования в качестве параметров сети и их тонкая настройка позволяет корректировать шаг дискретизации и таким образом более точно квантовать значения, близкие к нулю;
2. Алгоритм квантования нейронных сетей позволяет использовать неразмеченные данные для тонкой настройки;
3. Процедура перемасштабирования последовательных слоев нейронной сети с учетом нелинейной функции активации позволяет производить выравнивание итоговых распределений между каналами и увеличивает точность скалярного квантования до тонкой настройки;
4. Экспериментально подтвержденная процедура подбора порядка и мантиссы для нестандартных типов данных с плавающей запятой позволяет сократить разрядность вычислений до 11 бит для глубоких сверточных и рекуррентных архитектур без применения процедуры тонкой настройки;

Достоверность полученных результатов обеспечивается корректным использованием математического аппарата при разработке и анализе методов и корректным проведением большого числа тестов на реальных данных. Для измерения точности системы использовались объективные метрики, продемонстрировавшие результаты непротиворечивые друг другу и теоретическим выкладкам.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и получили одобрение на Международных конференциях Artificial Neural Networks and Machine Learning (ICANN) и 15th International Work-Conference on Artificial Neural Networks (IWANN).

Используя предложенный алгоритм, автор занял первые места в двух из трех номинациях на конкурсе LPIRC [1].

Личный вклад. Автор разработал и реализовал подходы к оптимизации производительности нейронных сетей. Разработанные подходы описаны в главах: “Глава 2. Метод дообучения порогов как комбинация дистилляции и квантования” и “Глава 4. Процедура нахождения оптимальной разрядности порядка и мантиссы”. Также автором проведен теоретический анализ разработанной процедуры дообучаемых порогов на основе алгоритма обратного распространения ошибки.

Публикации. Материалы диссертации опубликованы в 4 печатных работах в рецензируемых журналах [2—5]. Все работы изданы в индексируемых Scopus журналах. Работа [5] издана в журнале, относящемуся к квартили Q1, а работы [2—4] - в относящихся к квартилям Q3 и Q2.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируются цель и задачи, решению которых посвящена работа.

В **первой главе** приводится обзор на основе литературных источников методов оптимизации скорости работы нейросетевых моделей. Рассматриваются принципы работы и основные характеристики таких методов как прунинг, дистилляция, квантование, канонические разложения и нейросетевой архитектурный поиск. Не оставлен без внимания и процесс изобретения более эффективных архитектур, как один из первых методов оптимизации нейросетевых моделей.

В разделе **1.1 Разработка вычислительно эффективных архитектур на примере задачи классификации** описывается исторический процесс создания мобильных архитектур нейронных сетей на основе разработки вычислительно более эффективных операций.

Раздел **1.2 Нейросетевой поиск архитектур** посвящён описанию алгоритмов построения оптимальных моделей для конкретного набора

обучающих данных на основе заданных операций. В разделе приводится анализ ключевых статей, а также сравнение недифференцируемых методов поиска и дифференцируемых.

Раздел **1.3 Прунинг** посвящён подходу, позволяющему удалить избыточные связи в нейронной сети, не вносящие значимого влияния на процесс распознавания. Приводится классификация методов прунинга по способу удаления связей в нейронной сети (структурированным и неструктурированным образом), анализ различных критериев "полезности" связей, а также недостатки данного метода оптимизации, связанные с появлением дополнительных гиперпараметров.

Раздел **1.4 Квантование** посвящён методу оптимизации вычислительной сложности нейронной сети за счёт дискретизации пространства признаков и весов. Раздел описывает задачу квантования и приводит общую формулу для её решения. Далее в разделе производится классификация методов квантования по следующим признакам: симметричное или ассиметричное, во время или после обучения, а также приводится процедура симуляции квантования, позволяющая адаптировать весовые коэффициенты нейронной сети к шуму, вызванному квантованием.

Раздел **1.5 Дистилляция** посвящён подходу более эффективного обучения моделей на основе передачи знаний между уже обученной сетью и необученной. В разделе анализируются методы переноса знаний с промежуточных слоев, а также описан пример использования дистилляции совместно с процедурой квантования с целью повышения точности квантованной сети.

В разделе **1.6 Использование нестандартных типов данных** описывается практика использования компьютерных типов данных, не регламентированных стандартом IEEE-754, а именно чисел с плавающей запятой с сокращенной разрядностью порядка и мантиссы. В разделе приведен обзор работ, где подобные практики применялись к сетям с большой вычислительной сложности.

Раздел **1.7 Матричные и тензорные разложения** анализирует методы уменьшения вычислительной сложности нейронных сетей, основанных на классических подходах линейной алгебры. В разделе дано описание

канонического разложения, сингулярного и разложение Такера, а также применения данных подходов к сверточным нейронным сетям.

Раздел **1.8 Выводы по первой главе** суммаризирует приведенные выше подходы и анализирует недостатки каждого из них. Делается аргументированный выбор в пользу квантования и нестандартных типов данных в качестве исследуемых в диссертации методов.

Во **второй главе** описан, предложенный в диссертации, метод обучаемых порогов, позволяющий снизить влияние выбросов, возникших при процедуре калибровки, на процедуру квантования. Предлагаемый метод заключается в процедуре тонкой настройки порогов квантования, использовании State Through Estimator для оценки градиентов недифференцируемых функций и применении дистилляции с целью повышения точности.

Раздел **2.1 Описание процесса квантования нейронных сетей в программном пакете TensorflowLite** описывает схему квантования, использованную в программном пакете TensorflowLite, который используется большинством специалистов с целью запуска нейронных сетей на мобильных телефонах. Приводится анализ преимущества подхода с несимметричной процедурой квантования по сравнению с симметричной схемой.

В разделе **2.2 Проблема выбросов в задаче квантования нейронных сетей** освещается проблема выбросов как фактора, снижающего точность нейронных сетей при квантовании.

Раздел **2.3 Дифференцируемый порог квантования** посвящён описанию, предлагаемого в диссертации, метода обучаемых порогов квантования.

В качестве основных переменных, оптимизируемых в процессе тонкой настройки, предлагается использовать пороги квантования, возникающие в операции симуляции квантования.

Для улучшения работы предлагаемого подхода автором диссертации рекомендовано использовать следующие методы:

– Оптимизировать не сам порог квантования, а множитель от 0 до 1, который регулирует масштаб порога квантования, полученного в результате калибровки;

– Выполнить процедуру сплавения слоя пакетной нормализации со сверточными и полносвязными слоями и применять квантование уже к модифицированным весовым коэффициентам;
– Использовать процедуру векторного (поканального) квантования с целью уменьшения шага квантования;
– Использовать процедуру дистилляции на основе квадратичной функции потерь и неквантованной сети, используемой в качестве учителя;

Таблица 1 — Квантование в 8 бит в векторном режиме

Architecture	Symmetric thresholds, %	Asymmetric thresholds, %	Original accuracy, %
MobileNet v2	71.11 ±0.22	71.39 ±0.34	71.55
MNas-1.0	73.96 ±0.25	74.25 ±0.17	74.34
MNas-1.3	75.56 ±0.1	75.72 ±0.11	75.79

В разделе также описан способ применения данного метода к несимметричной процедуре квантования, как основному методу в программном пакете TensorflowLite.

В разделе **2.4 Результаты экспериментов** в свою очередь приводятся результаты тестирования скорости и точности разработанного и реализованного[1] подхода на открытом наборе данных ImageNet-2012.

В результате применения метода удалось ускорить нейронные сети с мобильными архитектурами MobileNet-v2 и MNAS-Net в 1.44 и 1.4 раза при падении точности менее чем на 1%. Также было продемонстрировано снижение требования к обучающим данных и скорости тонкой настройки (все обучение было выполнено на 10% от исходного датасета и проводилось не более 10 эпох).

В разделе **2.5 Анализ алгоритма быстрых дообучаемых порогов с точки зрения обратного распространения ошибки** проанализирован метод быстрых дообучаемых порогов с теоретической точки зрения.

[1] https://github.com/agoncharenko1992/FAT-fast-adjustable-threshold

Было показано, что использование обучаемых порогов и процедуры симуляции квантования действительно позволяет находить баланс между уменьшением выбросов и схожестью между оригинальной версией весов и квантованной.

В разделе **2.6 Описание процедуры перемасштабирования каналов для скалярного квантования** описано возможное улучшение процедуры, позволяющее оптимизировать разброс между выходными каналами сверточных слоев, что дает прирост качества при процедуре скалярного квантования. Процедура перемасштабирования основана на свойствах эквивариантности функций активации ReLU (см. рис. 1) и выравнивания распределения активаций после процедуры свертки, а также была адаптирована для функции ReLU6.

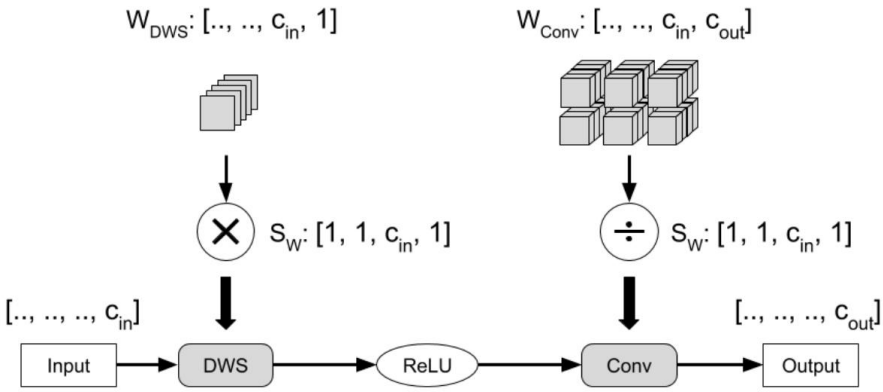

Рис. 1 — Процедура перемасштабирования выходов DWS слоя.

В разделе **2.7 Заключение по второй главе** подытоживаются полученные в данной главе результаты.

В **третьей главе** проанализированы особенности проектирования современных аппаратных архитектур для исполнения нейронных сетей. Приведен анализ вычислительной сложности основных слоев нейронных сетей, как прямого распространения, так и рекуррентных. Также приведено описание основных принципов работы ускорителей нейронных сетей. В

результате анализа было показано, что тип данных, с которым оперирует систолический массив, ключевой блок ускорителя, имеет ключевое значение для итоговой производительности аппаратной архитектуры.

Раздел **3.1 Анализ распространенных слоев нейронных сетей на возможность представления в унифицированном виде** сначала описывает принцип работы рекуррентной ячейки типа LSTM (подраздел **3.1.1**), а также анализирует вычислительную сложность основной операции, матричного умножения по сравнению с операциями другого типа в ячейке. Далее, в подразделе **3.1.2** приведен анализ доминирования операции свертки по сравнению с другим типом операций в современных архитектурах компьютерного зрения с позиции вычислительной сложности. В завершении раздела (в подразделе **3.1.3**) представлен способ вычисления операции свертки методом матричного умножения, что унифицирует большинство архитектур нейронных сетей и показывает, что именно эта операция и требует оптимизации как самая распространенная.

В разделе **3.2 Описание принципов работы аппаратных ускорителей нейронных сетей на основе систолического массива** приведена общая схема работы аппаратного ускорителя с систолическим массивом на основе уже существующего ускорителя от компании Google. Показано, что оптимизация элементарных вычислительных модулей приведет в оптимизации площади на кристалле данной составляющей ускорителя, что влечет за собой более эффективное распределение площади между остальными компонентами.

В **четвертой главе** приведена и экспериментально апробирована процедура нахождения оптимальной разрядности порядка и мантиссы. Апробация была выполнена на сверточных и рекуррентных архитектурах.

Раздел **4.1 Описание метода нахождения оптимальной разрядности порядка и мантиссы** приводит указанную процедуру для сверточных и полносвязных слоев на основе процедуры конвертации. Процесс конвертации заключается в отбрасывании избыточных бит мантиссы с округлением к ближайшему, при этом выполняется проверка переполнения порядка числа. Далее происходит конвертация в тип данных половинной точности и дальнейшие матричные вычисления происходят в нем (рис. 2).

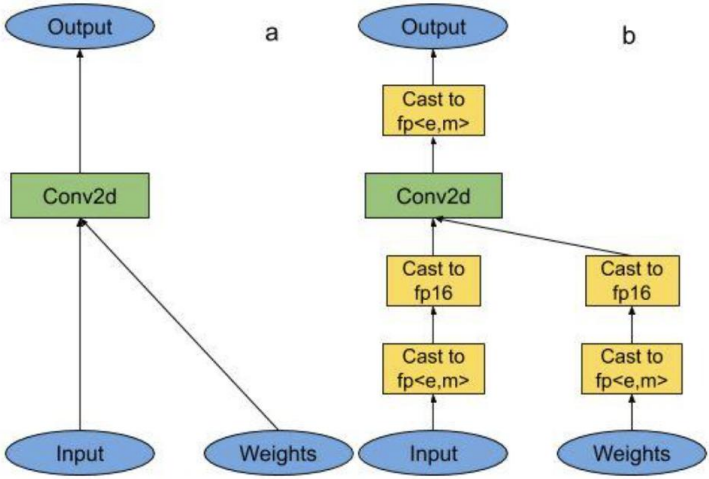

Рис. 2 — Визуализация процедуры конвертации. Пример исходного (a) и модифицированного (b) сверточного слоя.

Путем перебора всех возможных разрядностей порядка и мантиссы, начиная с разрядностей, соответствующих числам половинной точности, находится оптимальные разрядности для нестандартного типа данных.

В разделе **4.2 Нахождение оптимальной разрядности порядка и мантиссы для сверточных нейронных сетей** приведена экспериментальная апробация данной методики для архитектур ResNet-50, MobileNet-v2, GoogleNet. Было выяснено, что оптимальная разрядность, в которой могут работать все 3 сети соответствует 5 битам порядка и мантиссы. Результаты падения точности представлены в таблице:

Таблица 2 — Точности оригинальных и minifloat версий моделей

Architecture	minifloat<5,5> %	Original %
MobileNet v2	71.82	72.83
Resnet-50	73.9	74.04
GoogleNet	70.66	70.99

В разделе **4.3 Нахождение оптимальной разрядности порядка и мантиссы для рекуррентных нейронных сетей** приведена апробация процедуры на архитектуре DeepSpeech-v1 как на наиболее ярком представителе глубоких рекуррентных архитектур. Для данной архитектуры оптимальным оказался minifloat<4,4>: целевая метрика деградировала не более чем на 1% относительно float32 аналога.

В разделе **4.4 Заключение по четвертой главе** подытоживаются полученные результаты, проводится анализ, на основе которого делается вывод, что вычисления в типе данных с суммарной разрядностью меньше 9 приемлемы для избыточных сетей, однако для мобильных сетей данная разрядность не является подходящей.

В **заключении** приведены основные результаты диссертационной работы.

1. Был разработан алгоритм квантования с дообучаемыми порогами, который позволил оптимизировать современные мобильные архитектуры нейронных сетей с незначительной потерей (<1%) в точности;
2. Разработанный программный код, позволяющий воспроизвести результаты экспериментов, а также получить версии квантованных архитектур, был опубликован в открытом доступе в репозитории GitHub;
3. Предложенный алгоритм квантования позволил ускорить нейронные сети с архитектурами MobileNet-v2 и MNasNet-1.0 в 1.44 и 1.4 раз соответственно на процессоре типа ARM от компании Qualcomm на наборе данных ImageNet;
4. Разработанная методика нахождения разрядности порядка и мантиссы позволила сократить разрядность числового типа данных с 32 бит до 10 бит для сверточных нейронных сетей с архитектурами GoogleNet и ResNet-50 с падением точности менее 1% на наборе данных ImageNet и до 11 бит для архитектуры MobileNet-v2;

5. Разработанная методика была апробирована на рекуррентной нейронной сети с архитектурой DeepSpeech - v1. Для данной нейронной сети аналогичным образом удалось сократить разрядность с 32 бит до 9 при увеличении метрики WER на 0,0047;
6. Предложенные в данной работе методы реализованы и внедрены в виде ключевых модулей в компаниях ООО “Диалоговые системы”, ООО “Экспасофт” и ее клиентов, о чём имеются соответствующие акты о внедрении, прикреплённые к данной диссертации.

Список литературы

1. URL: https://lpcv.ai/competitions/2018. — (Дата обр. 25.06.2023).
2. *Гончаренко, А. И.* ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ НИЗКОРАЗРЯДНЫХ ПРЕДСТАВЛЕНИЙ ЧИСЕЛ С ПЛАВАЮЩЕЙ ЗАПЯТОЙ ДЛЯ ЭФФЕКТИВНЫХ ВЫЧИСЛЕНИЙ В НЕЙРОННЫХ СЕТЯХ [Текст] / А. И. Гончаренко, А. Ю. Кондратьев // Автометрия. — 2020. — Т. 56, № 1. — С. 93—99.
3. On practical approach to uniform quantization of non-redundant neural networks [Текст] / A. Goncharenko, A. Denisov, S. Alyamkin, E. Terentev // Lecture Notes in Computer Science. — 2019. — Т. 11728. — С. 349—360.
4. Trainable thresholds to uniform quantization of non-redundant neural networks [Текст] / A. Goncharenko, A. Denisov, S. Alyamkin, E. Terentev // Lecture Notes in Computer Science. — 2019. — Т. 11507. — С. 302—312.
5. Low-power computer vision: Status, challenges, and opportunities [Текст] / S. Alyamkin, ..., A. Goncharenko, G. Xuyang [и др.] // IEEE Journal on Emerging and Selected Topics in Circuits and Systems. — 2019. — Т. 9, № 2. — С. 411—421.