

## ОТЗЫВ

на автореферат диссертации Гончаренко Александра Игоревича «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами», представленной на соискание учёной степени кандидата технических наук по специальности 1.2.2 - Математическое моделирование, численные методы и комплексы программ

### Актуальность темы

Современные нейронные сети являются одним из наиболее эффективных методов распознавания образов, анализа и синтеза речи и текстов. Именно глубина, т. е. существенная многослойность, позволяет нейронной сети в ходе своего обучения решению некоторой задачи выстраивать эффективную иерархию обучаемых представлений, которую можно потом переиспользовать и при решении других задач того же класса. Но глубина несёт не только новые возможности, но и новые проблемы, связанные, прежде всего, с резко возрастающим количеством синаптических весов и, как следствие, сильным усложнением вычислений при обучении и при эксплуатации уже обученной нейросети (особенно при эксплуатации, потому что обучение происходит один раз на мощном сервере, а эксплуатация обученной модели выполняется неоднократно и, возможно, на маломощных мобильных устройствах). Одним из вариантов решения такого рода проблем является техника квантования нейронной сети, когда мантисса и экспонента синаптических весов представляется на сокращённой разрядной сетке. Но при выполнении квантования и других оптимизационных техник до конца не решены и потому **актуальны** проблемы определения оптимальной разрядной сетки и схемы квантования, и позволяющей сохранить точность работы нейронной сети при достижении максимального коэффициента

сжатия. Таким образом, поставленная в диссертационной работе цель и решённые задачи представляются **актуальными**.

### **Общая характеристика диссертации**

Диссертация состоит из введения, четырёх глав и заключения. Работа изложена на 98 страницах, включая 37 рисунков и 6 таблиц. Список цитируемой литературы содержит 94 наименований.

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

**В первой главе** приводится обзор предыдущих исследований других авторов области оптимизации скорости работы нейросетевых моделей. Рассматриваются принципы работы и основные характеристики прунинга, дистилляции, квантования, матричных и тензорных разложений, а также нейросетевого архитектурного поиска (как автоматического, так и «ручного», включая работы по созданию более эффективных нейросетевых архитектур).

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

**В третьей главе** проанализированы особенности проектирования современных аппаратных архитектур для исполнения нейронных сетей. Приведён 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.

**Четвертая глава** содержит описание разработки и экспериментальной апробации процедуры нахождения оптимальной разрядности порядка и мантиссы.

Апробация была выполнена на свёрточных и рекуррентных архитектурах для решения некоторых задач компьютерного зрения и распознавания речи.

**В заключении** приведены основные результаты работы.

Соискателем получен ряд оригинальных результатов, определяющих **научную новизну** диссертации:

1. Разработан новый метод квантования для нейросетей свёрточного и рекуррентного типов, основанный на точной настройке масштабирующих коэффициентов для пороговых значений квантования. По сравнению с некоторыми другими методами этот подход позволяет существенно сократить время, необходимое для тонкой настройки сети (до 10 раз) при минимальном снижении точности (в пределах 1%) относительно оригинальной (неквантованной) модели;

2. Создана и применена методика изменения масштаба весовых коэффициентов для процесса скалярного квантования, применяемая к ограниченной функции активации ReLU6, которая часто используется в нейронных сетях на мобильных устройствах;

3. Разработан и применён алгоритм определения разрядности для специализированных типов данных. Предложенный алгоритм не нуждается в дополнительном обучении нейросети, что облегчает процесс внедрения нейросетевых алгоритмов в сложные программно-аппаратные системы;

4. Предложенный алгоритм определения разрядности был впервые применён к глубокой рекуррентной сети для повышения вычислительной эффективности процесса распознавания речи.

### **Теоретическая и практическая значимость работы**

Научная значимость работы связана с более глубоким пониманием процессов дискретизации при квантовании глубоких свёрточных и рекуррентных нейронных сетей.

Практическая значимость работы заключается в реализации предложенного подхода в виде специализированного программного комплекса, позволяющего с меньшей потерей точности преобразовывать исходную модель свёрточной нейронной сети в её более оптимизированную (по памяти и по

скорости) версию, способную работать в реальном масштабе времени даже на маломощном ARM-процессоре мобильного устройства.

**Достоверность полученных результатов** обеспечивается корректным проведением экспериментов на реальных данных, проведённых автором, а также правильным применением математического аппарата. Для замера точности системы использовался ряд объективных метрик, значения которых в целом подтвердили теоретические выводы автора.

**Научные положения**, выносимые на защиту, подкреплены экспериментальными данными и теоретическими выкладками.

Существенных замечаний к материалам диссертационной работы нет. По содержанию диссертации возникли следующие вопросы:

1. Какие преимущества нового алгоритма квантования, предложенного автором, перед алгоритмами адаптивного квантования на основе гессиана или иных способов анализа чувствительности в глубоких нейронных сетях?

2. Почему не рассматривались наиболее популярные для многих классов задач нейросетевые модели трансформерного типа с механизмом внимания? Возможно, в механизме внимания есть какие-либо особенности, ограничивающие применение предложенного подхода нахождения разрядности для специализированных типов данных?

## **Заключение**

Диссертация Гончаренко Александра Игоревича является законченной научно-квалификационной работой, в которой последовательно изучены сложные проблемы эффективной оптимизации глубокой свёрточной нейронной сети по скорости её работы и объёму занимаемой памяти без существенной потери в качестве распознавания. Основные результаты работы опубликованы в трудах международных конференций «Artificial Neural Networks and Machine Learning (ICANN)» (Ираклион, 2023) и «15th International Work-Conference on Artificial Neural Networks (IWANN)» (Коста-Мелонерас, 2019).

Автореферат полностью отражает содержание диссертации.

Исходя из актуальности, новизны, научной и практической значимости представленной работы, можно сделать заключение, что диссертация


«Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами», представленная на соискание учёной степени кандидата технических наук по специальности 1.2.2 - Математическое моделирование, численные методы и комплексы программ, выполнена на высоком научном уровне. Она отвечает требованиям п.9-11, 13, 14 «Положения о присуждении учёных степеней», утверждённого Постановлением Правительства РФ от 24.09.2013 г. № 842, предъявляемых к кандидатским диссертациям, а её автор Гончаренко Александр Игоревич **заслуживает** присуждения учёной степени кандидата технических наук по специальности 1.2.2 - Математическое моделирование, численные методы и комплексы программ.

Отзыв составили:

Заведующий лабораторией прикладных цифровых технологий ММЦ ММФ Новосибирского государственного университета, доктор физико-математических наук

Научный сотрудник лаборатории прикладных цифровых технологий ММЦ ММФ Новосибирского государственного университета



  
Мулляджанов Рустам  
Илхамович

  
Бондаренко Иван