УТВЕРЖДАЮ:
Ректор Новосибирского
национального исследовательского
университета
академик РАН
М.П. Федорук

«20» октября 2023 г.

# ЗАКЛЮЧЕНИЕ

Новосибирского национального исследовательского университета
(НГУ)

Диссертация «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами» выполнена на кафедре Систем Информатики факультета информационных технологий Новосибирского государственного университета (ФИТ НГУ).

В период подготовки диссертации соискатель Гончаренко Александр Игоревич проходил обучение в аспирантуре ФИТ НГУ в период с 2015 - 2020 гг.

В 2015 г. Гончаренко А. И. окончил магистратуру физического факультета Новосибирского государственного университета по направлению подготовки 03.04.02 – «Физика». В 2020 г. окончил аспирантуру НГУ по направлению 09.06.01 – «Информатика и вычислительная техника»

Справка о сдаче кандидатских экзаменов № 2023/71 выдана 04 октября 2023 г. НГУ.

Научный руководитель - доктор технических наук Нежевенко Евгений Семёнович, ведущий научный сотрудник лаборатории информационной оптики ИАиЭ СО РАН.

Диссертация «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами» была рассмотрена на семинаре кафедры Систем информатики ФИТ НГУ 19 октября 2023 г.

На семинаре присутствовали:

1. Лаврентьев М.М. д.ф.-м.н., профессор
2. Юркевич В.Д., д.т.н., профессор
3. Васкевич В.Л., д.ф.-м.н., профессор
4. Терсенов А.С., д.ф.-м.н., профессор
5. Власов А.А., к.т.н., доцент
6. Загорулько Ю.А, к.т.н., доцент
7. Власов В.Н., к.ф.-м.н., доцент
8. Лемешко С.Б., к.т.н., доцент
9. Ануреев И.С., к.ф.-м.н., доцент
10. Апанович З.В., к.ф.-м.н., доцент
11. Котов К.Ю. к.т.н., старший преподаватель
12. Держо М.А., старший преподаватель

В обсуждении работы приняли участие д.т.н Юркевич В.Д., к.т.н Лемешко С.Б, д.ф.-м.н Лаврентьев М.М., к.ф.-м.н. Ануреев И.С. и другие. Заведующий кафедрой профессор М.М. Лаврентьев отметил высокую практическую значимость изложенных в диссертационной работе результатов и самостоятельность Гончаренко А.И. при их получении.

**Актуальность темы диссертационного исследования**

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.

Основная цель диссертационной работы Гончаренко А.И. состояла в разработке 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.

**Личное участие соискателя.**

В ходе выполнения работ Гончаренко А. И. проявил высокую самостоятельность, принимал активное участие в постановки задач, проведении экспериментов и анализе и обсуждении результатов. Гончаренко А.И. разработал и реализовал подходы к оптимизации производительности нейронных сетей. Автором проведен теоретический анализ разработанной процедуры быстрых дообучаемых порогов на основе алгоритма обратного распространения ошибки.

**Новизна**

В диссертации получены следующие новые научные результаты:

1. Предложен и реализован новый алгоритм квантования для моделей произвольного типа. 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;

2. Разработан и применен алгоритм перемасштабирования весовых коэффициентов для процедуры скаляного 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.

**Степень достоверности результатов.**

Достоверность полученных результатов обеспечивается корректным проведением большого числа тестов на реальных данных, в том числе независимым институтом стандартов. Для замера точности системы использовались различные объективные метрики, продемонстрировавшие непротиворечивость друг-другу результатов и теоретических выкладок.

**Практическая значимость.**

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.

Технология распознавания лиц, основывающаяся на разработанном подходе внедрена в продуктах компаний: ООО Новотелеком, ООО Рубитек РУС, ООО Экспасофт, ООО ИВА ТЕКНОЛОДЖИС.

**Соответствие специальности.**

Диссертационная работа соответствует специальности 1.2.2 – «Математическое моделирование, численные методы и комплексы программ».

Результаты диссертации соответствуют четырём пунктам паспорта специальности 1.2.2 – «Математическое моделирование, численные методы и комплексы программ»:

1. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий;
2. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента;

3. Разработка систем компьютерного и имитационного моделирования, алгоритмов и методов имитационного моделирования на основе анализа математических моделей;
4. Постановка и проведение численных экспериментов, статистический анализ их результатов, в том числе с применением современных компьютерных технологий.

**Полнота изложения материалов диссертации в работах, опубликованных соискателем.**

Результаты работы докладывались соискателем лично на следующих конференциях: Artificial Neural Networks and Machine Learning (ICANN 2019) и 15th International Work-Conference on Artificial Neural Networks (IWANN 2019). Используя предложенный алгоритм квантования, автор занял первые места в двух из трех номинациях на конкурсе LPIRC-2018.

Результаты диссертационной работы достаточно подробно отражены в пяти публикациях в рецензируемых научных журналах, индексируемых в российской и международных базах данных:

1. 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.
2. Low-power computer vision: Status, challenges, and opportunities [Текст] / S. Alyamkin, ..., D. Svitov, G. K. Thiruvathukal, B. Zhang, J. Zhang, X. Zhang, S. Zhuo [и др.] // IEEE Journal on Emerging and Selected Topics in Circuits and Systems. — 2019. — Т. 9, № 2. — С. 411—421.
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..

Диссертация «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами» Гончаренко Александра Игоревича рекомендуется к защите на соискание учёной степени кандидата технических наук по специальности 1.2.2 – «Математическое моделирование, численные методы и комплексы программ»

Руководитель семинара
д.ф.-м. н., профессор Лаврентьев М.М.

Секретарь семинара Держо М.А.