ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА 24.1.028.01 (Д 003.005.02) НА БАЗЕ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО БЮДЖЕТНОГО УЧРЕЖДЕНИЯ НАУКИ ИНСТИТУТА АВТОМАТИКИ И ЭЛЕКТРОМЕТРИИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК ПО ДИССЕРТАЦИИ НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ КАНДИДАТА НАУК

аттестационное дело № _____

решение диссертационного совета от «14» мая 2024 г. № 3

О присуждении Гончаренко Александру Игоревичу, гражданину Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами» по специальности 1.2.2. «Математическое моделирование, численные методы и комплексы программ» принята к защите «_8_» _февраля_ 2024 г. протокол № 1 диссертационным советом 24.1.028.01 (Д 003.005.02) на базе Федерального государственного бюджетного учреждения науки Института автоматики и электрометрии Сибирского отделения Российской академии наук (ИАиЭ СО РАН), 630090, г. Новосибирск, проспект Академика Коптюга, д. 1, приказ Минобрнауки России 255/нк от 28 марта 2020 года.

Соискатель Гончаренко Александр Игоревич 30.04.1992 года рождения,

в 2015 году окончил Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ),

в 2020 году освоил программу подготовки научно-педагогических кадров в аспирантуре Федерального государственного автономного образовательного

учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ),

работает в должности руководителя бизнес-единицы в ООО Экспасофт.

Диссертация выполнена на Кафедре систем информатики Факультета информационных технологий Федерального государственного автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ).

Научный руководитель – доктор технических наук

Нежевенко Евгений Семёнович, ведущий научный сотрудник Лаборатории волоконной оптики (15) Федерального государственного бюджетного учреждения науки Института автоматики и электрометрии Сибирского отделения Российской академии наук (ИАиЭ СО РАН), г. Новосибирск.

Официальные оппоненты:

Оселедец Иван Валерьевич, д.ф.-м.н., профессор, директор АНО «Институт искусственного интеллекта», г. Москва.

Куликов Виктор Александрович, к.т.н., должность Senior ML Scientist 2, Picsart Al Research (Picsart центр исследований ИИ), г. Ереван, Армения.

дали положительные отзывы о диссертации.

Ведущая организация Федеральное государственное бюджетное образовательное учреждение высшего образования «Новосибирский государственный технический университет» (НГТУ), г. Новосибирск,
в своем положительном заключении, подписанном

- Спектор Александр Аншелевич, д.т.н., профессор, профессор Кафедры теоретических основ радиотехники НГТУ, г. Новосибирск

заверенном

- Проректор по научной работе и инновациям НГТУ - Отто Артур

Исаакович, к.т.н.

указала, что, 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 предъявляемым к кандидатским диссертациям требованиям п.9-11, 13, 14 «Положения о присуждении учёных степеней», утверждённого Постановлением Правительства РФ от 24.09.2013 г. № 842, 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».

Соискатель имеет 12 опубликованных работ, в том числе по теме диссертации 4 научные работы, из которых **4** в рецензируемых научных журналах и изданиях, **3** акта о внедрении и **1** свидетельство о регистрации программы для ЭВМ:

1. 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.
2. 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.
3. 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.
4. 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.

На автореферат поступили следующие положительные отзывы:

- отзыв Муллядженова Рустама Илхамовича (д.ф.-м.н., заведующий Лабораторией прикладных цифровых технологий ММЦ ММФ, Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ), г. Новосибирск)

и Бондаренко Ивана Юрьевича (научный сотрудник Лаборатории прикладных цифровых технологий ММЦ ММФ, Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет» (НГУ), г. Новосибирск)

существенных замечаний к материалам диссертационной работы нет. Возникли 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?»

- отзыв Окунева Алексея Григорьевича (к.х.н., и.о.директора Института интеллектуальной робототехники НГУ, г. Новосибирск),

содержит замечания:

« 1. Низкое качество изображения под номером 2 в автореферате.

2. При описании раздела 2.7 в автореферате следовало бы повторить конкретные выводы, которые были сделаны в рамках предыдущих глав. Это облегчило бы понимание работы при чтении.

3. Некоторое количество опечаток. »

Выбор официальных оппонентов и ведущей организации обосновывается их высокой научной квалификацией в области математического моделирования, компьютерного зрения и глубоких нейронных сетей.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

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;

Предложен и реализован алгоритм перемасштабирования весовых коэффициентов для процедуры скалярного квантования для ограниченной функции активации ReLU6, наиболее распространенной в мобильных архитектурах нейронных сетей;

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.

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.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

разработанный подход был **реализован** 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 ИВА ТЕХНОЛОДЖИС”. Разработанные подходы внедрены в программно-аппаратный комплекс “Микропроцессор IVA TPU”
3. “ООО Экспасофт”. Разработанные методы легли в основу программного комплекса ускорения нейронных сетей “Expasoft Embedded NeuralNetwork Technology” или “EENNT”, имеющего свидетельство о государственной регистрации программы для ЭВМ и включенного в реестр отечественного ПО с номером реестровой записи 8738;

Оценка достоверности результатов исследования выявила:

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 непротиворечивые результаты, согласующиеся с теоретическими выкладками.

Личный вклад соискателя состоит в:

разработке и реализации подходов к оптимизации производительности

нейронных сетей на основе процедуры квантования и вычисления в типах данных сокращенной разрядности.

Разработанная автором процедура дообучаемых порогов **позволила** снизить требования к разметке и вычислительным мощностям, на которых выполняется тонкая настройка квантованной сети при незначительном падении её точности.

Разработанная автором процедура поиска оптимальной разрядности мантиссы и порядка **позволила** сократить количество необходимых бит для вычисления сверточных и реккурентных архитектур с 32 до 11 без процедуры тонкой настройки.

Также автором проведен **теоретический анализ** разработанной процедуры дообучаемых порогов на основе алгоритма обратного распространения ошибки.

В ходе защиты диссертации были высказаны следующие критические замечания:

Было высказано сомнение в применимости результатов диссертации к проблеме управления процессами с помощью нейросетей в реальном времени. Диссертант сформулировал своё видение перспектив в этой области.

Была высказана критика использования в докладе «антропоморфных» терминов применительно к нейросетям. Диссертант согласился с замечанием.

На заседании 14 мая 2024 года диссертационный совет постановил: за новые научно обоснованные технические, технологические решения в области разработки методов для ускорения вычисления нейронных сетей на мобильных платформах, имеющие существенное значение для развития страны, присудить Гончаренко Александру Игоревичу ученую степень кандидата технических наук по специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ».

При проведении тайного голосования диссертационный совет в количестве 22 человека, из них 6 членов диссертационного совета по специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ» -

технические науки, участвовавших в заседании, из 30 человек, входящих в состав совета, дополнительно введены на разовую защиту 0 человек, проголосовали: за 22 , против 0 , недействительных бюллетеней 0 .

Председатель диссертационного совета
академик РАН Шалагин Анатолий Михайлович

Ученый секретарь диссертационного совета
д. ф.-м. н. Ильичев Леонид Вениаминович

«15» мая 2024 г.