УТВЕРЖДАЮ

Проректор по научной работе и инновациям Федерального государственного бюджетного образовательного учреждения высшего образования «Новосибирский государственный технический университет»

_____ А. И. Отто

«09» февраля 2024 г.

ОТЗЫВ

ведущей организации - Федерального государственного бюджетного учреждения высшего образования «Новосибирский государственный технический университет» - на диссертацию 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

За последние 10 лет нейронные сети достигли значительного прогресса в своем развитии и используются практически повсеместно в задачах обработки речи, распознавания изображений и обработке естественного языка. Однако,

поскольку нейронные сети принципиально создавались как вычислительно-избыточные математические конструкции, их применение может быть затруднено в тех областях, где отсутствуют высокие вычислительные мощности. В таких приложениях, как поиск пропавших людей с дронов, детектирование в режиме реального времени пожаров на скважинах или в лесу, обнаружение незаконных мест складирования отходов, отсутствует доступ к широкополосному интернету, что ведет к невозможности пересылки данных с камеры и их обработки на сервере. Таким образом, возникает необходимость работы нейронных сетей на устройстве с низкими вычислительными ресурсами, такими как процессоры архитектуры ARM, что в свою очередь, остро ставит вопрос снижения вычислительной сложности нейронных сетей. При этом, на сегодняшний день уже существует ряд методов, применение которых может ускорить нейронную сеть, - это квантование, прунинг и дистилляция. Однако, эти методы не лишены следующих недостатков:

- Для применения каждого из методов требуется наличие качественной обучающей выборки;
- Применение каждого из методов может потребовать значительных вычислительных ресурсов;
- Применение каждого из методов может занимать до двух недель, что значительно замедляет их внедрение в реальные приложения.

Альтернативным подходом может являться создание собственных архитектурных решений на основе программируемых логических интегральных схем с типами данных, несоответствующими стадарту IEEE-754. Однако, данная область была слабо исследована в связи с ее специфичностью.

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 наименования.

Во введении обоснована актуальность работы, сформулированы цели и задачи исследований, научная новизна работы, представлены научные положения, выносимые на защиту, приведена научная и практическая значимость работы и дано обоснование их достоверности.

В первой главе приводится анализ современных методов оптимизации производительности нейронных сетей на основе литературных источников. Рассмотрены такие категории подходов, как дистилляция, квантование, прунинг и нейросетевой архитектурный поиск. Рассмотрены сильные стороны и недостатки существующих подходов. В результате были выявлены проблемы и узкие места в применимости имеющихся подходов к поставленной задаче.

Во второй главе описывается, предложенный в рамках данной диссертации, метод обучаемых порогов квантования. Описанный метод имеет две особенности, повышающие его практическую значимость. Во-первых, он позволяет найти баланс между сохранением формы распределения и влиянием выбросов, тем самым повышая точность квантованной сети после его применения. Во-вторых, в методе используется механизм дистилляции (в данном случае сетью-учеником служит квантованная нейронная сеть, а сетью-учителем, является оригинальная нейронная сеть), что позволяет применять его в задачах, где отсутствует качественная разметка.

В третьей главе проведен анализ математических операций, присутствующих в нейронных сетях, с целью отбора наиболее вычислительно сложных для возможной аппаратной реализации. В ней также показано, каким образом снижение типа данных, в котором происходит вычисление данных операций приводит к возможному ускорению работы нейронной сети.

В четвертой главе демонстрируется разработанная кандидатом процедура конвертации из стандартного типа данных в сокращенный.

Приведены экспериментальные результаты по подбору разрядности для сверточных и рекуррентных архитектур.

В заключении приведены основные результаты работы.

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

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 и практическая значимость работы

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.

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.

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

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

1. Почему автором диссертации не были рассмотрены альтернативные функции потерь, такие как перекрестная энтропия, для задачи дистилляции знаний?

2. Насколько применима предложенная процедура перемасштабирования весовых коэффициентов к альтернативным функциям активациии, в частности, к гиперболическому тангенсу?

3. В диссертации отсутствуют пояснения используемых в ней терминов «сплавление» и «избыточность» сети.

4. В диссертации приведены некоторые формулы без пояснения используемых в них обозначений.

5. Из диссертации неясно, почему определяющая роль в оптимизации сети отводится требованию к скорости обучения, а быстродействию нейронной сети в режиме реального времени уделяется меньшее внимание.

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) и 15th International Work-Conference on Artificial Neural Networks (IWANN).

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.

Настоящий отзыв обсуждён и одобрен на научном семинаре кафедры теоретических основ радиотехники Новосибирского государственного технического университета 9 февраля 2024 г., протокол № **2**. На заседании присутствовало **8** человек, в том числе **7** кандидатов и докторов наук по профилю диссертации.

Отзыв составил
профессор кафедры теоретических основ радиотехники Новосибирского государственного технического университета,
доктор технических наук,
профессор  Спектор Александр Аншелевич