ОТЗЫВ

официального оппонента Куликова Виктор Александровича на диссертацию Гончаренко Александра Игоревича на тему: «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами», представленную на соискание ученой степени кандидата технических наук по специальности 1.2.2 - "Математическое моделирование, численные методы и комплексы программ"

1. Актуальность темы исследования и соответствие требованиям Положения ВАК РФ по специальности

Мобильные устройства с низкими вычислительными ресурсами получили широкое распространение и используется в повседневной жизни миллиардов людей. Алгоритмы, основанные на искусственных нейронных сетях, позволяют автоматизировать ряд процессов в том числе на мобильных устройствах. В основе искусственных нейронных сетей лежит большой объем матричных умножений которые требуют повышенных вычислительных ресурсов в виде графических ускорителей.

Оптимизация искусственных нейронных сетей для работы на мобильных устройствах является актуальной и востребованной проблемой. В диссертации Гончаренко А.И. рассмотрены методы оптимизации основанные на квантизации и понижении разрядности вычислений. Предложенные методы квантования проверены проведены в задачах компьютерного зрения и анализа речи, что позволяет говорить о практической значимости и актуальности диссертационной работы.

Содержание диссертационной работы соответствует следующим пунктам направлений исследование по специальности 1.2.2 - "Математическое моделирование, численные методы и комплексы программ":

- 2. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий.
- 3. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента.
- 8. Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента.

- 9. Постановка и проведение численных экспериментов, статистический анализ их результатов, в том числе с применением современных компьютерных технологий (технические науки)

2. Структура диссертации

Диссертация включает в себя введение, четыре главы и заключение. Всего в ней 98 страниц, на которых содержится 37 рисунков и 6 таблиц. В списке литературы указаны 94 источника.

Во введении показана актуальность и практическая значимость работы, приведены цели и задачи исследование, а также основные положения, выносимые на защиту.

В первой главе проведен обзор методов ускорения и оптимизации нейронных сетей для мобильных устройств. Проведен анализ литературы, выявлены проблемы, связанные с использованием существующих решений.

Во второй главе описывается метод обучаемых порогов, который эффективно ускоряет нейронные сети, не требуя больших вычислительных ресурсов для тонкой настройки. Предложенный метод предлагает добавление обучаемого через метод обратного распространения ошибки масштабирующего коэффициента для порога квантования. Предложена схема обучения через процедуру дистилляции, который применим на неразмеченных наборах данных. Показаны результаты на задаче классификации изображений.

В третьей главе проведен анализ современных архитектур нейронных сетей, в результате которого выявлены наиболее трудоемкие операции. Оказалось, что эти операции сводятся к матричному умножению, для которого наиболее эффективна аппаратная архитектура систолического массива.

Показано, что при разработке новых нейронных процессорных устройств (NPU) можно улучшить производительность с использование нестандартных типов данных.

В четвертой главе показывается, как на основе симуляции типа данных с плавающей запятой с произвольной разрядностью можно моделировать потерю точности при конвертации. Сделаны выводы что для нейронных сетей наибольшую важность имеет диапазон, в котором производится вычисление, нежели точность самих вычислений. Данные факты подтверждены экспериментально.

В заключении приведены основные результаты работы, которые соответсвуют поставленным целя диссертационной работы.

3. Основные научные результаты, их новизна и ценность для науки и практики

В диссертационной работе Гончаренко А.И. предложены два подхода к оптимизации работы искусственных нейронных сетей на устройствах с малым количеством вычислительных ресурсов.

Предложен подход обучения параметров квантования весов нейронной сети, который может встраиваться в существующие схемы квантования, улучшая точность. Метод обучаемого порога квантования заключается в добавлении дифференцируемого масштабирующего коэффициента, который может быть настроен через метод обратного распространения ошибки. Для применения предложенного метода не требуется доступ к размеченному набору данных, достаточно иметь доступ к исходной модели и не размеченному набору данных, так называемая процедура дистилляции модели.

Второй вклад в диссертационную работу является процедура нахождения оптимальной разрядности порядка и мантиссы для программируемых микропроцессоров.

Подходы, предложенные в работе Гончаренко А.И., являются новыми и имеют практическую ценность, что подтверждается актом внедрения "ООО Диалоговые системы", "ООО Экспасофт" и "ООО ИВА ТЕХНОЛОДЖИС".

4. Достоверность и обоснованность выводов и результатов диссертационной работы

Достоверность научных результатов подтверждена комплексным исследованием предложенных подходов на стандартных архитектурах искусственных нейронных сетей с использованием доступных тестовых наборов данных, что позволяет сравнивать предложенный подход с существующими решениями. Полученные результаты не противоречат друг-другу, теоретическим выкладкам и результатам работ других авторов. Дополнительную достоверность полученных результатов обеспечивает открытый исходный код предложенных методов.

5. Основные замечания по диссертации

5.1 В главе 2 описана процедура дистилляции минимизация функции потерь между студентом (квантованной) и учителем (исходной) нейросетью описана следующим образом: "функции потерь использовался корень из среднеквадратичной ошибки (RMSE) между оригинальной и квантованной сетями" верно ли что функция потерь считается между всеми активациями или только между выходами сетей?

5.2. В таблицах 4 и 5 у модели с нестандартным типом данных точность получается выше, чем у модели во float32 как вы можете объяснить данное явление? В таблица 5 опечатка в названии MobileNet.

5.3. В главе 4 утверждается, что порядок более важен, чем точность, и возникает следующий вопрос: Функции нормализации приводят данные к стандартному диапазону значений (вычитая скользящее среднее и деля на скользящую дисперсию). Как наличие функции нормализации влияет на оптимальные параметры порядка и мантиссы?

5.4. В диссертационной работе рассмотрены сверточные сети (ResNet, MobileNet) и LSTM (DeepSpeech). Насколько предложенные методы подходят для архитектур, основанных на механизме внимания (Attention)? Возможно ли обобщить данные методы для трансформеров?

5.5. Из текста неясно, как понижение разрядности (использование нестандартны типов данных) влияет на скорость вывода нейронной сети или уменьшение потребления оперативной памяти устройства. Хотелось бы иметь сравнительную таблицу по использованию ресурсов при изменении точности вычисления.

6. Публикация и апробация результатов диссертации

Материалы диссертации опубликованы в 4 печатных работах в рецензируемых журналах индексируемых Scopus и ВАК РФ. Основные положения и результаты диссертационной работы докладывались на международных конференциях.

7. Содержание автореферата

Автореферат диссертации отражает основные проблемы, идеи, результаты, выводы и положения диссертационной работы.

8. Оценка диссертации в целом

Диссертационная работа Гончаренко А.И. «Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами» представляет собой законченную научно-квалификационную работу. Работа удовлетворяет требованиям пункта 9 “Положения о присуждении ученых степеней” ВАК при Минобрнауки РФ, а ее автор Гончаренко Александр Игоревич заслуживает присуждения ученой степени кандидата технических наук по

специальности 1.2.2 - “Математическое моделирование, численные методы и комплексы программ”.

Куликов Виктор Александрович,
кандидат технических наук,
Senior ML Scientist 2,
Picsart AI Research,
Армения, 0038, г. Ереван, ул. Алабяна, 16
Телефон: +37477328823
Эл. почта: kulikov.victor@gmail.com

Подпись  В.А. Куликов

15.04.2024

Отчёт о проверке электронной подписи

Отчет сформирован 23 апреля 2024 г., 15:31

Электронная подпись документа действительна

Проверяемые файлы

Исходный документ отзыв.pdf
Файл подписи отзыв.pdf.sig
Создан 23 апреля 2024 г., 13:08
Размер 3.12 KB

Поставлено 1 подпись

Статус подписи 1: Действительна

Стандарт CAdES-BES
Время подписи 19 апреля 2024 г., 0:53
Математическая корректность верна
Цепочка сертификатов действительна
Владелец сертификата CN=Куликов Виктор Александрович, ИНН=540862273528, СНИЛС=14036028414, SN=Куликов, G=Виктор Александрович, C=RU
Издатель сертификата CN=АО "ИИТ", ИНН ЮЛ=7743020560, ОГРН=1027739113049, O=-Акционерное Общество "ИнфоТеКС Интернет Траст", L=Москва, S=77 г. Москва, C=RU, STREET=-ул. Мишина, д. 56, стр. 2, эт. 2, пом. IX, ком. 11, E=SupportIIT@infotecs.ru
Действителен с 19 апреля 2024 г., 0:40 по 19 апреля 2025 г., 0:40
Серийный номер 01DA91B78535F70000178E2410EF0001
Отпечаток SHA1 08bb9f4ba51db037531490c33bc540ece74205f5