В диссертационный совет Д 003.005.02

Федеральное государственное бюджетное учреждение науки «Институт автоматики и электрометрии Сибирского отделения Российской академии наук»

ОТЗЫВ

официального оппонента Ивана Валерьевича Оселедца

на диссертационную работу Гончаренко Александра Игоревича «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

Нейронные сети получили широкое распространение в мире благодаря своим возможностям к решению задач, связанных с распознаванием образов. При этом, год от года нейронные сети потребляют все большие объемы вычислительные ресурсов, что обусловлено возможностью крупных и сильно избыточных сетей обобщать входной сигнал и находить в нем определенные закономерности, которые человек не способен формализовать. Вышеупомянутая тенденция в свою очередь вступает в противоречие с тенденцией к увеличению количества носимых устройств, поскольку последним недостаёт вычислительных ресурсов для выполнения ПО с «большими» нейронными сетями. Именно поэтому оптимизация нейронных сетей на сегодняшний день является актуальной областью прикладных и фундаментальных технических и математических наук. Существующие алгоритмы ускорения, в свою очередь, могут требовать процедуры дообучения и качественно размеченных данных, что увеличивает время до получения результата и сдерживает развитие научного прогресса. Кроме того, на сегодняшний день имеется тренд на снижение выработки CO2-выбросов при использовании и обучении нейронных сетей. Поэтому разработка процедуры ускорения нейронных сетей, которая сама по себе не является ресурсоемкой, является актуальной задачей. Создание такой процедуры является одним из направлений исследований, выполненных Гончаренко Александром Игоревичем.

Другим направлением является поиск меньшей разрядности числа с плавающей запятой, при которой падение качества разнообразных архитектур нейронных сетей будет незначительным. Мотивированно данное исследование тем, что при создании

специализированного модуля для вычисления нейронных сетей уменьшение разрядности элементарных регистров позволяет более гибко проектировать остальную аппаратную инфраструктуру.

2. Соответствие требованиям Положения ВАК РФ по специальности

Тема исследования диссертационной работы соответствует следующим пунктам паспорта специальности 1.2.2 - «Математическое моделирование, численные методы и комплексы программ»:

Пункт 2. «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»;

Пункт 3. «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента»;

Пункт 8. «Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента»;

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

3. Структура диссертации

Диссертация состоит из введения, четырех глав и заключения. Общее число страниц в диссертации равняется 98. В диссертации присутствует 37 рисунков и 6 таблиц, а в списке литературы указаны 94 источника.

Во *введении* показана актуальность темы, сформулирована цель и задачи диссертационного исследования и представлены основные положения выносимые на защиту.

В *первой главе* был проведен обзор существующих способов оптимизации производительности нейронных сетей, в частности описаны такие методы как ручное конструирование вычислительно-эффективных моделей, нейросетевой поиск архитектур, прунинг, матричные разложения, а также процедуры квантования, дистилляции и использования низкоразрядных вычислений. В конце главы выявлены недостатки приведенных выше процедур, не позволяющие применять данные методы напрямую для мобильных архитектур нейронных сетей.

Во *второй главе* описан новый, предложенный Гончаренко А.И. алгоритм квантования. Данный алгоритм суть комбинация дистилляции и модифицированной процедуры равномерного квантования. В модифицированной процедуре к порогам квантования были применены методы стохастической оптимизации, что позволило найти их оптимальные значения с точки метрик качества нейронной сети. Благодаря использованию дистилляции, автору диссертации удалось получить улучшения показателей нейронной сети без использования размеченных данных. Помимо метода

дообучаемых порогов, во второй главе представлена новая процедура перемасштабирования фильтров сепарабельной свертки, основанная на свойствах репараметризации нейронной сети. Также в этой главе были приведены экспериментальные результаты, подтверждающие эффективность предложенных методик.

В *третьей главе* приведен анализ особенностей проектирования специализированных аппаратных устройств для вычисления нейронных сетей. В данной главе соискатель проанализировал различные архитектуры нейронных сетей, в том числе и рекуррентные, и в результате анализа показал, что самой вычислительно-емкой процедурой в нейронных сетях является матричное умножение и свертка, представимая в аналогичном виде. После этого, автор проанализировал основные блоки для аппаратных ускорителей с архитектурами типа систолический массив и показал, что изменения площади матричного умножителя способно добавить гибкости при проектировании устройств подобного типа.

В *четвертой главе* автор диссертационной работы описывает процедуру поиска оптимальной разрядности порядка и мантиссы, а также приводит графическую иллюстрацию модифицированных операций матричного умножения и свертки, позволяющих работать с числами с плавающей запятой меньшей разрядности. Суть процедуры заключается в поэтапном изменении порядка и мантиссы с округлением к ближайшему. В главе также приведены экспериментальные данные изменения показателей метрик качества архитектур для задач компьютерного зрения и обработки речи при различных значениях экспоненты и мантиссы. После получения экспериментальных результатов автор выбирает оптимальные значения разрядности экспоненты и мантиссы для исследованных нейросетевых архитектур.

В *заключении* приведены основные результаты, полученные в диссертационной работе.

4. Основные результаты, их новизна и ценность для науки и практики

В диссертационной работе Александра Гончаренко предложены два новых алгоритма для оптимизации нейронных сетей, связанные с понижением разрядности вычислений. Алгоритм квантования на основе обучаемых порогов может быть использован с произвольной архитектурой нейронной сети, поскольку в качестве функции потерь используется среднеквадратичная ошибка между сетью учителем и учеником, и не требует размеченных данных. При этом, используя схему дистилляции, удалось не только снизить количество данных, требуемых для тонкой настройки, что привело сокращению времени на дообучение сети, но и избавиться от требования к качеству разметки.

Алгоритм подбора порядка и мантиссы, приведенный автором в четвертой главе, может быть полезен при разработке собственных архитектур аппаратных ускорителей нейронных сетей. Данный алгоритм был впервые применен к глубокой рекуррентной нейронной сети на основе LSTM ячейки. Предложенный Гончаренко А.И. метод не требует дообучения модели, что опять же сокращает время проведения экспериментов.

Оба подхода, разработанные автором диссертации, были применены к сравнительно легким мобильным архитектурам (например, MobileNet-v2, MNasNet и GoogleNet), что зачастую игнорируется научным сообществом, но крайне ценно с практической точки зрения.

Практическая ценность диссертации также подтверждается победой на конкурсе, докладами на конференциях и актами о внедрении, приложенными к работе. В актах о внедрении указаны нейронные сети, различные с точки зрения архитектуры, и решающие задачи, связанные не только с обработкой изображений.

5. Достоверность и обоснованность выводов и результатов диссертации

Достоверность результатов подтверждается проведением тестирования на публичных и открытых данных, а также наличием открытого исходного кода у предложенных алгоритмов. Полученные результаты не противоречат теоретическому анализу, приведенному в работе. Все выводы, сделанные в работе соискателя, имеют твердое и логичное обоснование.

6. Основные замечания и вопросы по диссертации

К диссертации соискателя имеются следующие замечания и вопросы:

- Результаты в главе 2 приведены для функции потерь RMSE, которая не всегда согласована с настоящей функцией потерь, используемых при обучении нейронной сети. Как изменятся выводы при использование других функций потерь?
- В последний год появилось несколько архитектур нейросетей, использующих экстремальной квантование (например, BitLinear). Могут ли результаты диссертации как-то помочь в таких исследованиях?
- Таблица 4 на стр. 76: неясно, почему при увеличении e, m точность модели падает.
- Во многих таблицах приведены значения точности с двумя знаками, однако известно, что при разных начальных приближениях/настройках оптимизатора могут получаться разные цифры; было бы более познавательно иметь значения качества моделей вместо со стандартным отклонением.
- Глава 3 стоит особняком, так как не содержит численных экспериментов. Более того, утверждение в выводах о том, что "операции являются частными случаями матричного умножения" является достаточно очевидным, так как любая линейная операция может быть реализована в виде матричного умножения. Вопрос состоит в том, насколько структурированной является такая матрица. Не

очень ясно, в чем состоит преимущество использования систолических массивов.

7. 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 соответствует всем требованиям Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24.09.2013 №842, а ее автор, Гончаренко Александр Игоревич, заслуживает присуждения ему ученой степени кандидата технических наук по специальности 1.2.2 — Математическое моделирование, численные методы и комплексы программ.

Иван Валерьевич Оселедец,

доктор физико-математических наук, профессор РАН,

директор АНО «Институт искусственного интеллекта»

Россия, 121170, г. Москва, Кутузовский пр-т, 32, к.1.

Телефон: 89154309949

Электронная почта: Oseledets@airi.net

Подпись _____ И.В. Оселедец

Подпись И.В. Оселедца подтверждаю
Директор [illegible] Кузнецов

АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОРГАНИЗАЦИЯ * ИНСТИТУТ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА * РОССИЙСКАЯ ФЕДЕРАЦИЯ ГОРОД МОСКВА