ОТЗЫВ

на автореферат диссертации Гончаренко Александра Игоревича, представленной на соискание ученой степени кандидата технических наук, на тему “Высокопроизводительные нейронные сети глубокого обучения для устройств с низкими вычислительными ресурсами” по специальности 1.2.2 - “Математическое моделирование, численные методы и комплексы программ”

С развитием нейронных сетей растет и их вычислительная сложность, что подтверждает появление современных языковых моделей, мультимодальных моделей и обобщенных моделей компьютерного зрения. При этом, с развитием робототехники и беспилотных летательных аппаратов, возникают естественные ограничения на вычислительную сложность моделей, работающих “на борту” таких систем. Работа Гончаренко Александра Игоревича затрагивает актуальный пласт исследований, связанных с уменьшением вычислительной сложности моделей. Сами по себе процедуры ускорения нейронных сетей - это отдельный пласт методов в области искусственного интеллекта, который требует тщательной подготовки данных и творческого подхода в процессе дообучения сети. Все это значительно снижает скорость научно-технического прогресса и внедрения нейросетевых методов в устройства интернета вещей

В главе 1 приводится обзор методов, позволяющих ускорять нейронные сети, объясняется выбор в пользу низкоразрядных вычислений (то есть квантования и вычислений с плавающей запятой с сокращенной разрядностью) в качестве основных методов, пригодных к дальнейшим улучшениям.

В главе 2 рассказываются две процедуры, разработанные соискателем: квантование с дообучаемыми порогами и процедура перемасштабирования весов. Автором сделано предположение, что выбросы, которые возникают в процессе работы нейронной сети могут существенно снизить качество ее работы при наивном квантовании. Гончаренко А.И. предлагает рассмотреть пороги квантования в качестве параметров, участвующих непосредственно в обучении нейронной сети и использовать для их оптимизации метод стохастического градиентного спуска. Соискателем также была представлена процедура, которая улучшает скалярное квантование нейронной сети MobileNet-v2. Процедура основана на идее выравнивания выходных распределений каналов при использовании функции активации ReLU6.

В главе 3 соискатель рассматривает проблему проектирования специализированных микропроцессоров для вычисления нейронных сетей. Им был проведен анализ вычислительной сложности основных слоев современных архитектур нейронных сетей. В рамках анализа было выяснено, что наибольшую вычислительную сложность представляют операции, связанные с матричным умножением. После этого был проведен обзор типичной архитектуры аппаратных ускорителей нейронных сетей на основе систолического массива. Было показано, что оптимизация площади матричного умножителя является актуальной задачей в проектировании микропроцессоров, поскольку влечет за собой более эффективное распределение площади между остальными компонентами, что в свою очередь повышает вычислительные возможности проектируемого устройства.

В главе 4 приведена процедура поиска порядка и мантиссы для типов данных с плавающей запятой сокращенной разрядности, проведены эксперименты, в результате которых была найдена оптимальная разрядность для таких архитектур как MobileNet-

v2, ResNet-50, GoogleNet и глубокая рекуррентная архитектура для распознавания речи DeepSpeech.

В заключении были приведены основные результаты диссертационной работы. Результаты, входящие в диссертацию, были опубликованы в журналах из списка ВАК, а также докладывались на международных конференциях.

К недостаткам автореферата можно отнести:

1. Низкое качество изображения под номером 2, приведенное в автореферат.
2. При описании раздела 2.7 в автореферате соискателю следовало бы повторить конкретные выводы, которые были сделаны в рамках предыдущих глав. Это облегчило бы понимание работы при чтении.
3. Некоторое количество опечаток.

Указанные замечания не влияют на общую положительную оценку работы А.И. Гончаренко. Представленная кандидатская диссертация является законченной научно-исследовательской работой, соответствующей требованиям ВАК, а ее автор, Александр Игоревич Гончаренко, **заслуживает** присуждения ему степени кандидата технических наук.

к.х.н,
и.о. директора Института Интеллектуальной Робототехники.
адрес: 630090, г. Новосибирск, ул. Пирогова, 1.
e-mail: okunev@nsu.ru
тел: +79139275749

02.05.2024 Окунев Алексей Григорьевич

Подпись Окунева А.Г. заверяю: