УТВЕРЖДАЮ
Проректор по
научно-исследовательской
деятельности НГУ
д. ф.-м. н.
_____ Чуркин Д.В.
12.05 2023 г.

ОТЗЫВ

ведущей организации - Федерального государственного автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет» - на диссертацию Свитова Давида Вячеславовича «Оптимизация производительности свёрточных нейронных сетей в системе распознавания лиц», представленной на соискание учёной степени кандидата технических наук по специальности 1.2.2 - Математическое моделирование, численные методы и комплексы программ

Актуальность темы

За последние десять лет глубокие нейронные сети стали основным инструментом решения задач распознавания образов. Это обуславливается, во-первых, их способностью к переносу обучения за счёт переиспользования иерархии обучаемых представлений, а во-вторых, парадоксальной на первый взгляд формой разрешения дилеммы смещения-разброса, стоящей перед любой моделью машинного обучения. Обычно при увеличении ёмкости модели снижается её ошибка смещения, но растёт ошибка разброса, в то время как в

ряде современных работ по глубокому обучению показано, что для нейронных сетей, обучаемых стохастическим градиентным спуском, с ростом их ёмкости ошибка разброса ведёт себя нетипично и тоже уменьшается. Это подводит теоретическую базу под бурным ростом современных нейронных сетей как «вглубь», так и «вширь». Тем не менее, практика применения нейросетевых алгоритмов в промышленности диктует другие требования - не увеличивать, а наоборот, уменьшать размеры нейронной сети, чтобы снизить вычислительную мощность и энергопотребление устройства, на котором выполняется нейросетевой алгоритм, а также углеродный след этого устройства. Устоявшаяся методология такого упрощения заключается в применении к исходной «большой» модели различных оптимизационных техник, таких как квантизация весов, дистилляция знаний (или просто дистилляция) и прунинг (прореживание). Особенно эффективными для реализации на вычислительных устройствах общего назначения являются квантизация и дистилляция, при этом дистилляция - обучение малой нейросети-«ученика» минимизировать невязку моделирования поведения большой нейросети-«учителя» - может приводить к более радикальному снижению размеров модели и упрощению вычислений, чем квантизация. Но при выполнении дистилляции и других оптимизационных техник до конца не решены и потому **актуальны** проблемы разработки таких алгоритмов оптимизации нейросети по её размеру, которые позволяют сохранить обобщающую способность и связанное с ней качество распознавания, достигнутые исходной «большой» моделью. 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 109 страницах, включая 23 рисунка и 18 таблиц. Список цитируемой литературы содержит 106 наименований.

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, а также возможность их использования в задаче распознавания лиц. 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 для оптимизации производительности детектора лиц. Описанный метод имеет две особенности, повышающие его практическую значимость. Во-первых, он позволяет снизить количество ложных срабатываний детектора, тем самым повышая его надежность. Во-вторых, метод основывается на модификации уже обученной нейронной сети и может быть применен для улучшения точности существующей системы путём небольших изменений в ней.

В третьей главе рассматривается разработка и тестирование метода оптимизации вычисления биометрического вектора. В моделях, обученных с функцией Софтмакс с отступами, для достижения наибольшей точности, предлагается новый метод дистилляции моделей. Описанный метод позволяет получить наибольшую точность на открытых наборах данных LFW, AgeDB-30 и MegaFace. Основная идея метода заключается в использование центров кластеров сети-«учителя» в модели ученика и добавление адаптивного отступа в функции Софтмакс вместо фиксированного.

Четвертая глава содержит описание разработки и реализации системы распознавания лиц, которая состоит из набора нейронных сетей и программного модуля, который используется на маломощных встраиваемых устройствах. В главе приводятся результаты тестирования системы на данных с домофона в различных условиях. А также результаты тестирования системы независимым институтом стандартов.

В заключении приведены основные результаты работы.

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

1. Предложен и реализован новый алгоритм дистилляции для моделей, обученных с функцией Софтмакс с отступом, для задачи построения биометрического вектора по изображению лица. Софтмакс с отступом — это модификация функции Софтмакс с добавлением константы для большего разнесения векторов в пространстве. Предложенный алгоритм впервые использует адаптивное вычисление отступов в функции Софтмакс на основе расстояния до центра кластера;

2. Предложен подход, позволяющий эффективнее разносить на гиперсфере биометрические вектора, полученные нейронной сетью с малым числом параметров. Предложенный подход впервые использует модель с большим числом параметров для нахождения центров кластеров векторов для повышения эффективности обучения малой нейронной сети;

3. Разработан и впервые применён метод ранней остановки исполнения нейросетевого детектора объектов на основе значения признаков промежуточных слоёв сети. Продемонстрировано повышение средний скорости обработки кадров за счёт использования предложенного метода;

4. Предложен и реализован устойчивый к шуму алгоритм локализации движения в видеопотоке. Впервые для решения этой задачи применены

глубокие признаки предобученного детектора объектов, что не влечёт дополнительных вычислительных затрат.

Теоретическая и практическая значимость работы

Научная значимость работы связана с более глубоким пониманием процессов оптимизации глубоких свёрточных нейронных сетей и связанным с этим расширением пространства возможных каналов передаче знаний между нейросетью-«учителем» и нейросетью-«учеником» при дистилляции знаний.

Практическая значимость работы заключается в реализации предложенного подхода в виде специализированного программного комплекса, позволяющего с меньшей потерей точности преобразовывать исходную модель свёрточной нейронной сети в её более оптимизированную (по памяти и по скорости) версию, способную работать в реальном масштабе времени даже на маломощном ARM-процессоре мобильного устройства.

Достоверность полученных результатов обеспечивается корректным проведением большого числа тестов на реальных данных, как проведённых самим автором, так и организованных независимым Американским национальным институтом стандартов (National Institute of Standards and Technology, или сокращённо NIST). Для замера точности системы использовался ряд объективных метрик, значения которых в целом подтвердили теоретические выводы автора.

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

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

1. Почему при дистилляции не применялась иерархическая многозадачная функция ошибки сети-«ученика», которая может более эффективно моделировать иерархию обучаемых представлений сети-«учителя» и за счёт этого снижать невязку дистилляции знаний?

2. Если целью введения функции Софтмакс с отступом является большее разнесение векторов в пространстве, то каковы её принципиальные преимущества перед процедурой разнесения векторов в пространстве, реализованной в алгоритме SupCon (Supervised Contrastive Learning, или сопоставительного обучения с учителем), который был предложен Prannay Khosla в 2020 году?

Заключение

Диссертация Свитова Давида Вячеславовича 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 «ICDLFR 2020: 22nd International Conference on Deep Learning and Face Recognition» (Амстердам, 2020) и «ММРО 2021: Математические методы распознавания образов» (Москва, 2021).

Автореферат полностью отражает содержание диссертации.

Исходя из актуальности, новизны, научной и практической значимости представленной работы, можно сделать заключение, что диссертация «Оптимизация производительности свёрточных нейронных сетей в системе распознавания лиц», 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 Свитов Давид Вячеславович **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 "Интеллектуальные системы и системное программирование" кафедры программирования механико-математического факультета Новосибирского государственного университета 11 мая 2023 г., протокол № 10. На заседании присутствовало 25 человек, в том числе 8 кандидатов и докторов наук по профилю диссертации.

Отзыв составил:
Заведующий лабораторией прикладных цифровых технологий ММЦ ММФ Новосибирского государственного университета,
доктор физико-математических наук

Муллюджанов Рустам Ильхамович