В диссертационный совет Д 003.005.02
Федеральное государственное бюджетное учреждение
науки «Институт автоматики и электрометрии Сибирского
отделения Российской академии наук»

О Т З Ы В

официального оппонента

на диссертационную работу Свитова Давида Вячеславовича «Оптимизация производительности свёрточных нейронных сетей в системе распознавания лиц», представленную на соискание учёной степени кандидата технических наук по специальности 1.2.2 – Математическое моделирование, численные методы и комплексы программ

1. Актуальность избранной темы исследования

В современном мире сложно переоценить значимость технологий искусственного интеллекта. Одним из бурно развивающихся направлений исследований в этой области является компьютерное зрение. И хотя с момента победы сверточной сети AlexNet в Large Scale Visual Recognition Challenge прошло уже больше 10 лет, в научной и прикладной сферах наблюдается широкое применение сверточных архитектур нейронных сетей для решения задач компьютерного зрения. А по данным на конец десятых готов двадцать первого века современные модели распознавания лиц, базирующиеся на глубоком обучении, превосходят человека в точности верификации.

Однако современные тенденции показывают, что прирост эффективности во многом достигается за счет увеличения обучаемых параметров искусственных нейронных сетей. Очевидно, платой за прирост в точности становится вычислительная мощность. Вместе с тем, реализация подобного рода моделей на различных мобильных или встраиваемых платформах

сопряжена с рядом сложностей, на решение которых и направлена диссертационная работа Свитова Давида Вячеславовича.

Тем не менее, часто требуется реализация алгоритмов компьютерного зрения на вычислителях небольшой мощности. Характерным примером являются системы биометрической идентификации или контроля и управления доступом. Решая довольно узкую задачу определения личности, они, как правило, должны выполнять ряд этапов. Во-первых, обнаружение человека и лица на кадре видеопоследовательности. Во-вторых, нормализация лица для последующего анализа. В-третьих, непосредственно идентификация личности. Действительно, чем сложнее и точнее модель, тем больше времени на выполнение прогноза требуется нейронной сети.

При решении увеличения производительности альтернативным увеличению вычислительной мощности вариантом является оптимизация существующих моделей глубокого обучения. Например, компания Intel разработала собственный набор инструментов для уменьшения времени инференса. Данный инструмент получил название OpenVINO и за счёт оптимизации моделей способен сокращать время обработки одного сэмпла на порядки.

Исследователи сегодня также пользуются такими подходами, как прунинг, квантизация и дистилляция знаний. Однако их рассмотрение в узкой задаче построения биометрических векторов на сегодняшний день в литературе освещено недостаточно.

Темой диссертационного исследования соискателя и являются алгоритмы и способы оптимизации моделей распознавания лиц. Учитывая вышесказанное, можно сделать вывод, что представленная тема соответствует современному тренду научных исследований, а также является актуальной, поскольку направлена на решение существующей проблемы низкой скорости работы больших моделей глубокого обучения. Разработанные Свитовым Д.В. решения по оптимизации сверточных нейронных сетей с применением технологий дистилляции весов вносят важный вклад в развитие систем распознавания лиц

и обработки биометрической информации, и при сохранении точности становятся приоритетными для их использования в реальных системах компьютерного зрения.

2. Мнение о научной работе соискателя в целом

Структура диссертации соответствует предъявляемым требованиям. Диссертационная работа состоит из списка сокращений и определений, введения, четырёх глав основного текста, заключения, списка литературы, списка рисунков и списка таблиц. Также имеется приложение Акта о внедрении результатов диссертационного исследования в ООО «Экспасофт».

Во **Введении** автор дает краткое обоснование применения именно архитектур сверточных нейронных сетей, историческую справку, а также обращает внимание на существующую проблему низкой производительности на CPU или встраиваемых устройствах (домофоны, камеры наружного наблюдения и т.д.). Цель сформулирована научным языком, а по большому счёту, заключается в повышение производительности маломощных и микропроцессорных устройств при реализации моделей распознавания лиц без существенных потерь в качестве. На основе цели сформулированы задачи исследования, направленные на разработку новых методов оптимизации детекторов лиц и построителей биометрических векторов.

Отмечена научная новизна, связанная с модификацией функции Софтмакс, а также с разработкой метода ранней остановки детекции на основе видеопотока.

Описана теоретическая и практическая значимость, заключающиеся в подходе передачи знаний при дистилляции и реализации исследуемых моделей в программном комплексе соответственно.

Кроме того, перечислены положения, выносимые соискателем на защиту и обладающие научной новизной. Следует отметить, что на защиту выносится 8 положений. Представленные далее разделы «Достоверность», «Апробация работы», «Личный вклад», «Публикации», «Объем и структура работы»

подтверждают вовлеченность автора в диссертационное исследование, а также удовлетворяют требованиям, предъявляемым к диссертационным работам.

В целом, во **Введении** присутствуют все необходимые и достаточные разделы. Опционально можно было бы добавить раздел «Методы исследований».

В **первой главе** Свитов Д.В. проводит аналитический обзор методов оптимизации производительности, а также исследует применимость существующих технологий оптимизации к системам распознавания лиц. Автор уделяет особое внимание современным популярным подходам, а именно прунингу, квантованию, дистилляции и нейросетевому архитектурному поиску. Достаточно подробно описаны идеи, преимущества и недостатки каждого из подходов. Автором отмечена оригинальная работа с предложением прунинга, но также отмечены и современные исследования этого способа оптимизации, включая «гипотезу о лотерейных билетах» и алгоритм SynFlow. В конце раздела приводится классификация методов прунинга по различным критериям. В разделе про дистилляцию также отмечена оригинальная работа. Также отмечен метод преобразования функции Софтмакс с температурой для сглаживания расхождений вероятностей. Рассмотрены и современные подходы к дистилляции. Более того, автором сделано важное замечание о том, что для задач распознавания лиц лучшие метрики получаются при применении функции Софтмакс с отступом. Представлены разные методы дистилляции для такого случая, в том числе разработанный автором. Кратко отмечены недостатки и преимущества различных подходов к дистилляции. Рассмотрен также раздел про квантование. Представляется математическая база и идея методов квантования. Далее приводится классификация методов квантования. Автором отмечаются фреймворки, в которых реализованы те или иные из рассмотренных подходов. Наконец, еще один раздел посвящен методом нейросетевого архитектурного поиска, которые напрямую не связаны с оптимизацией производительности, но позволяют подбирать подходящие для решаемой задачи архитектуры в автоматическом режиме. Делается

классификация таких методов, рассматриваются актуальные статьи по поиску архитектур. Однако подобного рода процессы поиска сами занимают достаточно много времени и требуют оптимизации. В конце **первой главы** на основе проведенного анализа предлагается алгоритм получения оптимизированной модели для распознавания лиц, который может быть обобщен на другие случаи применения глубокого обучения. При этом автор обосновывает удаление этапа прунинга в своем исследовании, поскольку будут использоваться специализированные под задачу распознавания лиц модели. Наконец, делается упор на развитие метода дистилляции и повышение метрики точности (precision) за счет уменьшения доли ложных срабатываний.

Вторая глава посвящена оптимизации работы моделей, связанных с обнаружением лиц на изображениях, как этапа предшествующего идентификации и требующего значительных вычислительных ресурсов. Соискателем предлагается алгоритм для отсечения статичных кадров, что позволяет повысить производительность детектора и снизить нагрузку на вычислитель. В качестве модели строится модель фона изображения, но в предложенном алгоритме этапы построения модели фона и детекции лиц объединены, что также позволяет говорить об оптимизации. Во втором разделе приводится краткое описание существующих моделей детектирования, а также методов определения движения на последовательности изображений. Рассмотрена задача объединения в одном алгоритме обнаружения объектов и движения. Предлагается новый подход к решению задачи одновременного обнаружения объектов и движения – AmphibianDetector. Данный метод использует карту признаков после первых m слоев обработки для принятия решения о необходимости дальнейшей детекции. Полученные карты признаков используются и для фильтрации ложных срабатываний на статичные объекты. Так, для остановки детектора необходимо определить косинусное расстояние между моделью текущего кадра и моделью фона, а затем сравнить его с порогом. При этом модель фона постоянно обновляется в случае, если порог не превышен. Кроме того, во **второй главе** описаны и выбраны метрики для

последующего сравнения различных подходов. В качестве основной метрики для детектирования предложена классическая метрика средней точности mAP. Предложенный алгоритм детектирования движения сравнивается с попиксельным алгоритмом. Для моделирования несовершенства регистрирующего устройства предложена модель изменения изображения на основе гауссового шума. Проведенные эксперименты для датасета CDNet2014 pedestrian показали преимущество использования предложенного подхода. В частности, в среднем наблюдаются выигрыши, как по скорости обработки, так и по точности. Сравнение с известными архитектурами при оптимальных параметрах числа слоев и порога показало выигрыш по mAP порядка 20%. Также следует отметить, что разработанный детектор AmphibianDetector размещен в открытом доступе на GitHub, куда приводится ссылка.

Результаты исследования предложенных во *второй главе* алгоритмов оптимизации детекторов объектов на изображений развиваются в **третьей главе** применительно к задаче извлечения признаков таких объектов в виде вектора. В частности, задача оптимизации рассмотрена для построения биологических векторов на основе обнаруживаемых лиц. Свитов Д.В. уделяет особое внимание функции активации Софтмакс, поскольку в данной главе предложено использование её менее распространённой версии, а именно Софтмакс с отступами. Отмечена разница между традиционными задачами классификации и задачей распознавания лиц. Важно отметить метод дистилляции, который использует из сети-учителя центры имеющихся классов, и основан на обучении сети-ученика воспроизводить такие вектора, углы от которых до центров кластеров близки к углам из сети-учителя. В данной главе также продемонстрировано преимущество дистиллированных моделей по производительности по отношению к большим моделям. Все реализованные методы дистилляции, для которых осуществляется сравнение представлены в открытом доступе. Результаты проведенных экспериментов показали, что предложенный метод дистилляции позволяет обеспечить наименьшую просадку точности по сравнению с известными методами. При этом сама

просадка колеблется от 0,15% до 1,7% для разных датасетов. В случае датасета с большим количеством негативных примеров точность падает на 7%, однако предложенный подход остается подходом с максимальной точностью среди прочих методов дистилляции. Прирост в точности 0,9% удалось обеспечить при копировании центров классов в сеть-ученика.

Четвертая глава описывает программный модуль, который был использован для обучения, дистилляции и тестирования различных моделей. Представлены три этапа разработанной системы распознавания лиц – обучение, конвертация моделей и их реализация на конечном устройстве. В качестве языка программирования на первом и втором этапе использовался Python, при этом обучение выполнялось на фреймворке PyTorch, а конвертация происходила в формат TensorFlow с помощью ONNX. Третий этап использовал модели TensorFlow, но на языке программирования C++ для маломощных устройств и на языке Python для серверных устройств. Для каждой подзадачи происходило обучение отдельной модели нейронной сети. В главе отмечено, что после реализации программного кода было выполнено тестирование программы Американским национальным институтом стандартов (NIST). Результаты тестирования показали удовлетворительный результат на датасетах Visa и Wild. Сравнительный анализ предложенного подхода и решений от российских вендоров показал, что для датасета Wild при незначительной просадке по точности распознавания, выигрыш по скорости составляет более 20 раз. Соискателем было показано, что разработанные методы и предложенные подходы могут быть применены в реальных устройствах. Так, например, было проведено тестирование на изображениях, полученных с умных домофонов. В данных условиях было выявлено снижение качества работы алгоритмов при сильном сжатии изображений на 8,2%, но предложенный метод дистилляции позволяет повысить качество на 3-4% и превосходить сеть-учителя на сжатых данных. Приемлемые результаты были продемонстрированы и для изображений в ночное время суток при использовании камер с ИК-подсветкой.

В **Заключении** автором представлены основные полученные результаты исследования, в числе которых предложенный метод инициализации весов с малым числом параметров, обеспечивающий повышение точности распознавания; метод дистилляции, приводящий к повышению точности малых сверточных нейронных сетей с использованием Софтмакс с отступами; предложенный детектор объектов и движения, обеспечивающий повышение производительности обнаружения. Также в качестве результата диссертации в рамках поставленных задач следует выделить разработанный программный комплекс, автоматизирующий процесс обучения, дистилляции, инференса, тестирования моделей распознавания лиц. По приведенным результатам можно сделать вывод, что поставленные в диссертационном исследовании задачи были успешно выполнены Свитовым Д.В.

Порядка 45% источников в **Списке литературы** были опубликованы в последние 5 лет, что также говорит о высокой актуальности тематики и высоком уровне проработки автором современного состояния исследований. В списке литературы также отражены ряд работ, в которых соискатель является соавтором, среди которых имеются и работы в высокорейтинговых журналах.

Результаты работы используются в технологиях ООО «Экспасофт», которые применяются в ООО «Новотелеком», ООО «Рубитек РУС», ООО «Открытая мобильная платформа» (подтверждается актом от 24.01.22), что говорит об их теоретической и практической значимости.

В рамках работы над исследованием Свитовым Д.В. опубликованы 5 работ в рецензируемых журналах, в том числе 2 в изданиях из перечня ВАК, 2 в журналах Scopus первой квартили (Q1).

Диссертация написана хорошим научным (научно-техническим) языком, в работе присутствует ряд графиков, отражающих полученные результаты. Материалы также сопровождаются множеством таблиц со сравнением различных алгоритмов, что позволяет читателю получить положительное представление о представляемых решениях. Автором соблюдается цельность и последовательность изложения материала, достаточное единство терминологии

и обозначений, что позволяет оценить степень проработки выдвигаемых на защиту научных положений и результатов.

Автореферат в большей степени отражает содержание диссертации и является её сокращенной версией.

3. Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

При рассмотрении работы установлено, что все теоретические выводы и практические зависимости, представленные в диссертации, в достаточной мере обоснованы и имеют завершённый вид.

В процессе исследования и решения научных задач применяется единый подход, методы, выбранные для исследований адекватны, используемые в диссертации аппарат нейросетевых технологий (во многом математический) и популярные фреймворки для работы с такими технологиями позволяют судить о высокой степени обоснованности представленных результатов.

Поставленные задачи исследования хорошо детализированы и приводят к достижению поставленной цели. Все задачи были решены соискателем на высоком уровне.

4. Новизна исследований и достоверность полученных результатов

Научная новизна диссертации заключается в разработке методов, применяемых для дистилляции знаний от тяжеловесных моделей с учетом распределения центров классов, а также методы повышения производительности свёрточных нейронных сетей в задачах обнаружения движущихся объектов за счет условной детекции только для кадров с движением. Кроме того, научная новизна состоит в том, что автором:

- предложен и реализован новый алгоритм дистилляции для моделей, обученных с функцией Софтмакс с отступом, для задачи построения биометрического вектора по изображению лица;

- предложен подход, позволяющий эффективнее разносить на гиперсфере биометрические вектора, полученные нейронной сетью с малым числом параметров;

- разработан и впервые применён метод ранней остановки исполнения нейросетевого детектора объектов на основе значения признаков промежуточных слоёв сети;

- предложен и реализован устойчивый к шуму алгоритм локализации движения в видеопотоке.

Достоверность результатов подтверждается применением адекватных решаемым проблемам теоретических методов, алгоритмов и специального программного обеспечения. Кроме того, тестирование NIST и показатели в условиях обработки реальных данных также позволяют судить о работоспособности разрабатываемых систем компьютерного зрения, а именно распознавания лиц.

Исследования выполнены с применением известных наборов изображений с лицами и использованием популярного в задачах глубокого и машинного обучения языка программирования Python.

Кроме того, полученные результаты были опубликованы в рецензируемых источниках и представлены научному сообществу на конференциях различного уровня, включая Conference on Computer Vision and Pattern Recognition (CVPR) 2021.

5. Соответствие работы установленным требованиям

Диссертационное исследование Свитова Давида Вячеславовича «Оптимизация производительности свёрточных нейронных сетей в системе распознавания лиц» соответствует требованиям п.9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24.09.2013 №842, содержание диссертации соответствует паспорту специальности 1.2.2 – Математическое моделирование, численные методы и комплексы программ.

Диссертация представляет собой выполненную автором научно-квалификационную работу, содержащую решение актуальной задачи повышения производительности свёрточных нейронных сетей в системах распознавания лиц на основе оптимизации детектора и дистилляции знаний предобученных моделей.

6. Замечания по работе

Несмотря на сформированное положительное впечатление о работе, к ней имеется ряд замечаний:

1. В качестве детекторов лиц также интересно было бы рассмотреть в обзорной части модель MTCNN, которая использовалась в работе, а также сравнить скорость детекции предложенного подхода с методом Виола-Джонса.

2. Не до конца понятно, на чем основан выбор параметров гауссовского шума, и какова модель наблюдаемых изображений. То есть он складывается с нормализованными значениями яркости изображения или с фактическими?

3 Тестирование NIST для датасета Border показало, что FNMR = 0,9996. С чем могут быть связаны такие плохие значения метрики?

4. В акте внедрения название темы: «Оптимизация производительности сверточных нейронных сетей без потери точности» не совпадает с темой диссертации.

5. Интересно было бы провести исследование оптимизируемых моделей на FPGA.

6. Для кандидатской диссертации выносится довольно большое число положений на защиту (восемь штук). На мой взгляд, можно было бы сократить их количество до 4-5.

7. На рис. 4.5 числа накладываются друг на друга, что делает чтение графика неразборчивым.

8. На с. 5 для термина «квантование», судя по всему, автор хотел передать понятие термина «дистилляция». А сам список рекомендуется упорядочить по алфавиту.

9. На с. 24 размещена формула без номера и описания ее составляющих. В целом и далее нет нумерации формул. На с. 63 идет ссылка на формулу 1, но нумерации для формул нет.

10. В работе имеются опечатки, пунктуационные и грамматические ошибки. Например, на с. 6 к задачам относятся … генеративные модели. Лучше было бы сказать – генерация изображений. На с. 10 в пункте 3 присутствует опечатка – «Продемонстрировано повышение средний скорости…» вместо «Продемонстрировано повышение средней скорости…» и другие.

11. В автореферате сначала идет ссылка на таблицу 2, а потом на таблицу 1. Целесообразно представлять ссылки в порядке появления.

Надеюсь, некоторые замечания будут полезны диссертанту в его дальнейших научных исследованиях, а некоторые будут сняты во время дискуссии.

7. Выводы

Несмотря на указанные замечания, диссертационная работа Свитова Д.В. заслуживает положительной оценки, поскольку в ней решены актуальные задачи в области оптимизации искусственных нейронных сетей применительно к системам распознавания лиц. Решение задач происходит на основе численных методов, таких как предложенный новый алгоритм дистилляции; а также с применением моделей зашумленных изображений для оптимизации детектора лиц. Адекватность разрабатываемых решений проверяется в программном модуле для работы с нейронными сетями.

Автором проведена достаточная апробация результатов диссертационного исследования в рамках участия на конференциях различного уровня и путём опубликования результатов исследования в рецензируемых научных журналах, включая журналы перечня ВАК.

Работа «Оптимизация производительности свёрточных нейронных сетей в системе распознавания лиц» соответствует требованиям Положения о присуждении ученых степеней, утвержденного постановлением Правительства

Российской Федерации от 24.09.2013 №842, а ее автор, Свитов Давид Вячеславович, заслуживает присуждения ему ученой степени кандидата технических наук по специальности 1.2.2 – Математическое моделирование, численные методы и комплексы программ.

Доцент Департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных Финансового университета, кандидат технических наук Андриянов Никита Андреевич

Контактная информация организации:

федеральное государственное образовательное бюджетное учреждение высшего образования «Финансовый университет при Правительстве Российской Федерации»

Адрес: Российская Федерация, 125167, г. Москва, пр-кт Ленинградский, д. 49/2

Сайт: http://www.fa.ru/
Телефон: +7 (499) 943-98-29
E-mail: naandriyanov@fa.ru

Подпись Андриянова Н.А.
ЗАВЕРЯЮ
секретарь Ученого совета
Финансового университета
В.В. Звягинцева
19 апреля 2023 г.