ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию Лимоновой Елены Евгеньевны «Биполярная морфологическая аппроксимация нейрона для уменьшения вычислительной сложности глубоких сверточных нейронных сетей», представленную на соискание ученой степени кандидата технических наук по специальности 1.2.2 – «Математическое моделирование, численные методы и комплексы программ»

Актуальность исследования

В настоящее время многие задачи компьютерного зрения и распознавания образов решаются методами обучения нейросетевых моделей со сверточным слоями. При этом растет запрос на повышение быстродействия моделей в целях использования на малопроизводительных устройствах, таких как мобильные телефоны и встраиваемые системы, в том числе в режиме реального времени. Особенно актуально увеличение скорости работы сверточных слоёв, на которые приходится значительная часть вычислений.

Для решения этой проблемы на сегодняшний день существует множество подходов. Одним из них является преобразование базовых операций внутри слоев или даже отдельных нейронов модели. В качестве примера такого преобразования можно привести аддитивную сеть AdderNet, в которой в операции свертки используется L_1 расстояние между входным вектором и фильтрами. На специализированных вычислителях такие модели являются более энергоэффективными и обеспечивают большую скорость работы, чем классические модели. Другим широко используемым подходом является квантизация весов. В этом подходе числа с плавающей точкой преобразуются в целочисленный вид, над которыми выполняются куда более простые операции, например, целочисленное умножение или даже просто доступ по индексу. Все эти методы активно исследуются, однако их применение чаще всего сопряжено со снижением качества работы преобразуемой модели и рассматривается индивидуально для каждой нейросетевой модели. Поэтому дальнейшие исследования в этой области актуальны и востребованы.

В диссертационном исследовании Лимоновой Е.Е. предложена новая схема вычислений в нейросетевых моделях, призванная ускорить их работу при сохранении высокого качества распознавания. Такие модели автор называет биполярными

морфологическими. Основной идеей их построения является использование вычислительно-простых морфологических операций (в данном случае – максимума и сложения) при вычислении сверток и матричных произведений в нейросетевых моделях. Для этого автор предлагает: аппроксимацию, позволяющую перейти к такому виду модели, методы обучения для нее, а также выполняет теоретическое и экспериментальное исследование скорости и точности полученных моделей.

Содержание работы

Диссертация состоит из введения, трех основных глав, заключения и приложений. Общий объем работы составляет 138 страниц, она содержит 29 рисунков, 19 таблиц, и 135 источников содержится в списке литературы.

Во введении раскрывается актуальность темы исследования, формулируются его цели и ставятся задачи диссертационной работы. Целью исследования Лимоновой Е.Е. являлись разработка и исследование вычислительно-эффективных аппроксимаций нейросетевых моделей, методов их обучения и оптимизации их вычисления на существующих и перспективных вычислителях. Для этого автором были поставлены и решены следующие задачи:

1. Разработка метода аппроксимации вычислительно-интенсивных частей нейросетевых моделей, а именно, биполярного морфологического нейрона, исследование его вычислительной эффективности и точности.
2. Оценка вычислительной эффективность моделей, использующих биполярный морфологический нейрон на различных платформах.
3. Разработка методов обучения моделей с биполярными морфологическими нейронами в сверточных или полносвязных слоях.
4. Экспериментальная оценка точности предложенного метода обучения для различных нейросетевых архитектур.
5. Разработка комплекса программ, позволяющего моделировать биполярные морфологические нейросетевые модели, их обучение и проверку результирующего качества работы.

Первая глава посвящена анализу низкоуровневого устройства нейронных сетей классической архитектуры LeNet и современной глубокой архитектуры ResNet. Подробно

рассмотрены базовые модели нейронов и их альтернативные варианты, устройство вычислительно-емких слоев и выполнен анализ существующих методов их ускорения. Также, поскольку рассматриваемое направление находится на стыке анализа изображений и проектирования вычислительных устройств, автор рассматривает два способа оценки вычислительной эффективности и аппаратной сложности вычислительного метода. Первый из них позволяет оценить параметры метода на центральных процессорах и учитывает количество арифметико-логических устройств, наличие SIMD-расширений, различную длительность этих операций. Второй метод подходит для программируемых логических интегральных схем или специализированных устройств, предполагающих, что логические вентили будут формировать необходимые пользователю операции. Именно эти методы в дальнейшем используются в работе. Таким образом, показано, что создание эффективных аппроксимированных схем вычисления нейросетевых моделей, ориентированных на конкретные классы вычислителей, является востребованным как среди ученых, так и среди практиков. В главе вводится необходимая терминология, формулируется цель исследования и ставятся научно-технические задачи.

Во второй главе Лимонова Е.Е. предлагает биполярный морфологический нейрон – упрощенную версию базового нейрона. После нелинейного преобразования входного и выходного векторов операции в слое аппроксимируются максимумом от сумм входных данных и весовых коэффициентов. При этом такая аппроксимация выполняется четырежды для входных векторов, содержащих данные разных знаков, и дважды для случая знакопостоянных данных. Автор рассматривает ее с теоретической точки зрения и доказывает, что в одномерном случае с помощью достаточного количества нейронов биполярного морфологического вида, организованных в трехслойную нейронную сеть, возможно равномерно приблизить любую непрерывную на компакте функцию с заранее заданной точностью. То есть, потенциально биполярные морфологические нейронные сети не уступают по выразительной способности классическим нейронным сетям.

Далее соискательница предлагает дальнейшее усовершенствование биполярной морфологической модели: модель с упрощенными нелинейными преобразованиями. Для этого она приближает двоичный логарифм по методу Митчелла, а возведение в степень согласно методу Шраудольфа. В результате при использовании биполярных морфологических нейронов в сверточных слоях оценка выигрыша составила 30-40% по

латентности для ПЛИС и специализированных устройств. При этом наблюдается небольшой рост числа вентилей за счет наличия двух наборов (для знакопостоянного входа) аппроксимаций и вычислителей для них.

Третья глава посвящена экспериментальной оценке точности биполярных морфологических моделей. В ней показано, что такие модели не обучаются стандартными методами, и предложен метод постепенного перехода от классической модели к биполярной морфологической, позволяющий осуществить преобразование модели без существенных потерь качества. Далее Лимонова Е.Е. рассматривает ряд задач усложняющихся задач распознавания: задачи классификации печатных (символы паспорта РФ, символы машиночитаемой зоны паспорта РФ) и рукописных символов (выборка MNIST), классификации объектов (выборка CIFAR10), семантической сегментации (выборка DIBCO 2017) и с использованием все более сложных моделей демонстрирует, что биполярные морфологические сети способны их решить. При этом в случае LeNet-подобных моделей изменение качества находится в пределах погрешности, а в случае более глубоких 22-слойных ResNet и 10-слойных UNet моделей снижение качества становится заметно. Поэтому автор предлагает использовать гибридные модели, в которых лишь часть слоев аппроксимируется. Это хорошо согласуется с оценками вычислительной эффективности, показавшими преимущество для достаточно вычислительно-емких слоев.

В заключении излагаются основные результаты работы:

1. Предложена биполярная морфологическая аппроксимация классической модели математического нейрона, обладающая схожей выразительностью при более простой внутренней структуре;
2. Доказано, что нейросетевая модель с достаточным числом нейронов биполярного морфологического вида может приблизить произвольную непрерывную на компакте функцию с любой заранее заданной точностью.
3. Оценка числа вентилей и латентности ПЛИС-реализации для биполярных морфологических сверточных слоев по сравнению с классическими сверточными слоями показала, что для слоев целевыми параметрами они используют практически столько же вентилей, сколько и классические слои, однако имеют латентность на 30-40% ниже.

4. Предложен оригинальный метод послойного преобразования и дообучения, демонстрирующий лучшее качество аппроксимированных моделей по сравнению с обучением этих же моделей прямыми методами.
5. Проведенные вычислительные эксперименты показали, что предложенная аппроксимация может использоваться в глубоких нейросетевых моделей в задачах классификации изображений и семантической сегментации без снижения качества распознавания для гибридных моделей и ряда полностью преобразованных моделей.
6. Разработан комплекс программ, позволяющий реализовать предложенные методы и проведенные эксперименты.

В приложении А содержится информация о двух зарегистрированных программах для ЭВМ. Это программный комплекс для осуществления экспериментов с биполярными морфологическими нейронными сетями, их обучение и оценку качества распознавания, а также программа для распознавания идентификационных карт личности «SmartID Reader», использующая методы, предложенные в диссертационной работе. Приложении Б содержит информацию о внедрении результатов диссертационной работы в деятельность АО «МЦСТ», а также программные продукты коммерческих организаций ООО «Смарт Энджинс Сервис» и АО «Тинькофф Банк».

Публикации

Основные результаты по теме диссертации Лимонова Елена Евгеньевна опубликовала в 10 печатных работах. Одна из них – работа без соавторства, опубликованная в журнале из перечня ВАК 2022; еще 8 работ опубликованы в изданиях, входящих в индексы цитирования Scopus и/или Web of Science, в том числе одна работа – в журнале Q1 Scopus; одна работа опубликована в сборнике трудов конференции. Эти работы и автореферат полностью отражают содержание диссертации.

Степень обоснованности научных положений, выводов и рекомендаций

Предложенные методы и подходы подробно описаны в тексте диссертационной работы, детально рассмотрены и исследованы теоретически и экспериментально. Практически все эксперименты, представленные в работе, выполнены на открытых

выборках данных с использованием стандартного программного инструментария. Исходный код ряда экспериментов также опубликован, что делает экспериментальные результаты работы воспроизводимыми и верифицируемыми. Полученные результаты и выводы были апробированы на профильных семинарах и международных конференциях ICMV и ICPR. Таким образом, основные положения, выносимые на защиту, в достаточной степени обоснованы.

Достоверность результатов исследования

Достоверность проведенного Лимоновой Е.Е. исследования опирается на методы математического анализа, теории алгоритмов и вычислительной оптимизации. Все экспериментальные результаты получены с помощью вычислительных экспериментов и не противоречат результатам, полученным другими авторами.

Научная новизна исследования

Следующие результаты, полученные соискательницей, являются новыми:

- предложена новая биполярная морфологическая аппроксимация базового нейрона, использующего операции умножения и сложения, нейроном на основе операций суммы и максимума, которая позволяет строить глубокие нейросетевые модели;
- приведено доказательство, что 3-слойные биполярные морфологические нейронные сети могут приблизить произвольную непрерывную на компакте функцию с любой заранее заданной точностью;
- для аппроксимированных нейросетевых моделей предложен метод обучения путем их постепенного преобразования и уточнения коэффициентов, который позволяет добиться более высокого качества распознавания, чем при прямом обучении; метод впервые опробован для биполярной морфологической и 8-битной целочисленной аппроксимации;
- экспериментально показано, что нейросетевые модели типов LeNet, ResNet и UNet демонстрируют высокую точность распознавания при использовании биполярной морфологической аппроксимации.

Теоретическая и практическая значимость работы

В данной работе Лимоновой Е.Е. предложена упрощенная модель нейронной сети, которая представляет интерес как с точки зрения теории, так и с точки зрения практики. Несмотря на доказанную возможность аппроксимации произвольных непрерывных функций на компакте, методы формирования эффективных нейросетевых архитектур, состоящих из биполярных морфологических нейронов, могут быть нетривиальными, что и показала соискательница в своей работе. Процесс обучения подобных моделей сталкивается с некоторыми трудностями и требует применения дополнительных методов и аппроксимаций. Подробное изучение этих методов позволит лучше понять особенности задач распознавания и усовершенствовать существующие алгоритмы обучения нейросетевых моделей.

С точки зрения практики предложенные Лимоновой Е.Е. методы могут применяться в задачах, которые решаются сверточными нейронными сетями и которые требует высокой производительности и качества работы. Класс таких задач крайне широк: это не только рассмотренные в работе визуальная классификация и семантическая сегментация, но также поиск объектов, обработка текстов на естественных языках и многие другие задачи. Результаты диссертации Лимоновой Е.Е. нашли применение на практике:

- в составе программного обеспечения, в том числе предназначенного для мобильных устройств (приложения ООО «Смарт Энджинс Сервис» и АО «Тинькофф Банк»);
- при планировании и проектировании новых аппаратных устройств в АО «МЦСТ».

Замечания

В работе можно отметить следующие недостатки:

1) Поставленные эксперименты показывают, что идея работает. Но, к сожалению, у метода довольно много ограничений, из-за чего есть сомнения в его применимости на практике без дополнительной проработки. Автор пишет, что при реализации на обычных процессорах (на программном уровне), предложенный нейрон не будет иметь преимуществ над традиционным нейроном. Преимущество достигается только на ПЛИС и СБИС. Тренировать напрямую сеть, построенную на таких нейронах сложно. Процесс не всегда сходится, поэтому автор предлагает сначала тренировать классическую

нейронную сеть. Затем классические нейроны заменить на биполярные и далее сеть дополнительно обучить уже с новой структурой. Только в этом случае тренировка завершается успешно, и качество предсказаний остается на хорошем уровне. Для классических нейронных сетей методы ускорения вычислений часто применяются уже после тренировки на этапе инференса. Это позволяет расширить и упростить возможности для их применения. В предложенном подходе просто заменить нейроны после обучения, возможно, не выйдет - упадёт точность.

2) Во введении не хватает упоминания алгоритма MADNESS со схожей идеей, который позволяет заменить векторное умножение на Look Up таблицу [https://arxiv.org/abs/2106.10860, https://github.com/dblalock/bolt]. Как следствие обычный нейрон работает очень быстро.
3) Сравнение производится с сетью реализованной на базе плавающей точки, однако довольно частый подход, при котором веса квантизуют до скажем 8 бит и сложения и умножения становятся целочисленными. Такая нейронная сеть становится крайне эффективной для реализации, как на программном так и на аппаратном уровне. Если сравнивать предложенный метод с квантизованной нейронной сетью, то есть шанс что предложенная нейронная сеть проиграет.
4) К сожалению, программный код, разработанный автором не выложен в открытый доступ и в целом в диссертации плохо документирован. Сейчас в машинном обучении для почти всех знаковых проектов, статей и идей принято выкладывать код на Github или другую площадку.

Отмечу, что эти замечания не влияют на положительную оценку исследования в целом, его практическую и научную значимость.

Общая оценка работы

Диссертация Лимоновой Е.Е. написана грамотным языком и хорошо структурирована. Научные термины и обозначения определяются и используются корректно. Результаты и выводы обоснованны, новы и представляют теоретический и практический интерес. В целом, диссертация хорошо проработана и, несмотря на вышеуказанные замечания, выполнена на высоком научном уровне и соответствует

требованиям Высшей аттестационной комиссии при Министерстве образования и науки Российской Федерации, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а также критериям пунктов 9-14 Положения о присуждении ученых степеней. Работа выполнена по специальности 1.2.2 – «Математическое моделирование, численные методы и комплексы программ» и соответствует паспорту специальности.

Считаю, что Лимонова Елена Евгеньевна заслуживает присуждения ученой степени кандидата технических наук.

Официальный оппонент,
главный научный сотрудник отдела методологии проектирования интегральных схем Федерального государственного бюджетного учреждения науки Института проблем проектирования в микроэлектронике Российской академии наук
доктор технических наук,
чл.-корр. РАН _____ / Соловьев Р. А.

«21» 01 _____ 2023 г.

Сведения об организации:
Федеральное государственное бюджетное учреждение науки Институт проблем проектирования в микроэлектронике Российской академии наук
124365, Москва, Зеленоград, ул. Советская, д. 3
ippm@ippm.ru,
+7 (499) 729-98-90; +7 (499) 729-93-23

Подпись Соловьева Р.А. заверяю
начальник отдела кадров З.А. Горелова