УТВЕРЖДАЮ
Директор ФГУ ФНЦ НИИСИ РАН
Доктор технических наук С.Е. Власов
«26» января 2023 г.

ОТЗЫВ

ведущей организации 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. Практически все современные методы распознавания и обработки изображений тем или иным способом задействуют сверточные нейронные сети. Часто они являются глубокими, то есть включают в себя большое количество сверточных слоев. Эти слои являются крайне вычислительно- и энергоемкими, что затрудняет их использование на мобильных устройствах и во встраиваемых системах. Тем не менее, именно использование нейронных сетей на конечных устройствах позволяет решить проблемы с безопасностью пользовательских данных и повысить автономность работы приложений. Поэтому уменьшение вычислительной нагрузки глубоких сверточных нейронных сетей на сегодняшний день крайне актуально. Целевыми вычислителями при этом являются мобильные центральные процессоры и специализированные или программируемые логические устройства. Они обладают ограниченными вычислительными ресурсами, а также от них ожидается ограниченное потребление энергии, связанное с работой от аккумулятора. Именно этим проблемам посвящена работа Лимоновой Е. Е.

В своей диссертационной работе Елена Евгеньевна рассматривает аппроксимацию базового элемента современных нейронных сетей – классического математического

нейрона – и показывает, как она может применяться в сверточных моделях. Эта аппроксимация упрощает вычисления в слое, обеспечивая более высокую скорость работы за счет использования операций максимума и сложения вместо операций сложения и умножения. Приведенные оценки показывают, что предложенная аппроксимация может быть использована для программируемых или специализированных логических интегральных схем. Подобная подход (замена классического поля вещественных чисел с операциями сложения и умножения полуполем вещественных чисел с операциями сложение и максимум) уже использовался и принес интересные результаты в теоретической физике и классической математике (идемпотентная математика или тропическая математика). В диссертационной работе продемонстрировано, что такой подход оказывается плодотворным и в информатике и приближенных вычислениях.

Целью работы Лимоновой Е.Е. являлись разработка и исследование вычислительно-эффективных аппроксимаций нейросетевых моделей, методов их обучения и оптимизации их вычисления на существующих и перспективных вычислителях

Задачи. В диссертации были решены следующие задачи:

1. Разработан метод аппроксимации вычислительно-интенсивных частей нейросетевых моделей, а именно, биполярный морфологический нейрон, исследованы его вычислительная эффективность и точность.
2. Выполнена оценка вычислительной эффективность моделей, использующих биполярный морфологический нейрон на различных платформах.
3. Разработан методы обучения моделей с биполярными морфологическими нейронами в сверточных или полносвязных слоях.
4. Проведена экспериментальная оценка точности предложенного метода обучения для различных нейросетевых архитектур.
5. Разработан комплекс программ, позволяющий моделировать биполярные морфологические нейросетевые модели, их обучение и проверку результирующего качества работы.

Научная новизна. В проделанной работе Лимоновой Елены Евгеньевны **впервые:**

1. Предложена новая аппроксимация классического нейрона нейроном с морфологической структурой, позволяющая создавать глубокие нейронные сети с морфологическими слоями и обеспечивающая высокую точность распознавания.

2. Предложен новый метод обучения произвольных, в том числе биполярных морфологических и целочисленных, аппроксимаций классических нейросетевых моделей путем послойного преобразования и дообучения, позволяющий повысить их качество.
3. Впервые показано, что для предложенной аппроксимации метод послойного преобразования и дообучения позволяет добиться более высокого качества работы нейросетевой модели, чем прямое обучение с помощью метода обратного распространения ошибки и градиентных методов оптимизации.
4. Проведено оригинальное исследование точностных характеристик нейросетевых моделей LeNet- и ResNet-подобных архитектур, использующих предложенную морфологическую аппроксимацию.
5. Впервые теоретически показано, что нейросетевая модель с достаточным числом нейронов биполярного морфологического вида может приблизить произвольную непрерывную на компакте функцию с любой заранее заданной точностью.

Общая характеристика диссертационной работы. Диссертация Лимоновой Е. Е. состоит из введения, трех глав, заключения, списка литературы и двух приложений. В конце каждой главы приводятся выводы. Общий объем диссертации составляет 138 страниц, список литературы содержит 135 наименований.

Во введении обоснована актуальность работы, ее новизна, практическая значимость, приводятся положения, выдвигаемые на защиту.

В главе 1 описываются модели нейронов, такие как классический математический нейрон, спайковый нейрон и морфологические нейроны с дендритами и без. Эти модели обладают различной вычислительной эффективностью, однако массово применяется лишь классический математический нейрон, основанный на нейроне Маккаллока-Питтса. Он стал основной структурной единицей современных глубоких нейросетевых моделей, так как позволяет добиться высокого качества при использовании существующих методов обучения. Далее Елена Евгеньевна кратко рассматривает устройство современных нейросетевых архитектур и анализирует существующие методы повышения их вычислительной производительности. Эти методы разнообразны и включают в себя методы уменьшения числа арифметических операций в моделях (использование более простых моделей, удаление наименее значимых весовых коэффициентов, декомпозицию сверток и

т.д.), методы квантования, то есть преобразования коэффициентов в целые числа, менее требовательные к вычислительным ресурсам, а также методы модификации базовых вычислений в слое или отдельном нейроне. Методы последней категории представляют значительный научный интерес, поскольку часто предполагают совместную модификацию вычислителя и нейросетевой модели. При этом также необходимо уделить внимание разработке метода обучения, чтобы разработанная модель обеспечивала не только высокую скорость, но и качество работы. Именно такая комплексная задача и решается в диссертационной работе.

В главе 2 Елена Евгеньевна предлагает биполярную морфологическую аппроксимацию классического математического нейрона. Суть этой аппроксимации заключается в логарифмировании и последующем потенцировании знакопостоянных входных значений и весовых коэффициентов. В результате нейрон использует взятие максимума от сумм преобразованных входных значений и весов, а не сумму произведений, как классический нейрон. Стоит отметить, что такая аппроксимация явным образом выделяет возбуждающие и тормозящие сигналы, как это происходит в биологических биполярных нейронах, что и дало название аппроксимации.

Для нейронной сети, состоящей из биполярных морфологических нейронов, Елена Евгеньевна доказывает теорему о том, что такая сеть может равномерно приблизить произвольную непрерывную на компакте функцию. В диссертационной работе приведено полное доказательство для одномерного случая.

Далее для сверточных и полносвязных слоев, использующих биполярные морфологические нейроны, проводятся оценки числа арифметических операций, а также латентности и транзисторной сложности. Поскольку операции логарифмирования и возведения в степень применяются лишь для входных и выходных данных каждого слоя, биполярные морфологические слои продемонстрировали заметное снижение латентности. Однако, за счет наличия двух или четырех вычислительных блоков, число транзисторов, необходимое для реализации такого слоя становится несколько больше, чем у классического слоя. Кроме того, в случае двух вычислительных блоков необходимо дополнительное время для расчетов, если вход слоя был знакопеременный, что несколько снижает диапазон эффективно реализуемых моделей.

Для улучшения этих оценок были рассмотрены кусочно-линейные аппроксимации функций двоичного логарифмирования и возведения в степень двойки. Благодаря стандартному представлению чисел с плавающей точкой, эти аппроксимации просты для вычисления и реализации в виде логических схем. Их использование действительно позволило дополнительно упростить предложенную модель. Приведенные оценки и вычислительные эксперименты показывают, что биполярные морфологические сверточные слои в широком диапазоне параметров обладают меньшей латентностью, чем классические сверточные слои и практически не уступают им по транзисторному бюджету.

В главе 3 исследуется обучение моделей с биполярными морфологическими нейронами. Для этого в диссертации рассматриваются модели типов LeNet-5, ResNet и UNet. Они содержат последовательность сверточных слоев и применяются для задач классификации и сегментации изображений. Для обучения биполярных морфологических моделей Елена Евгеньевна предложила послойный метод, который последовательно преобразует и аппроксимирует слои классической модели. Далее для каждого преобразованного слоя выполняется обучение всей сети, позволяющее восстановить качество модели с биполярными морфологическими слоями.

Эти модели тестируются в задаче классификации рукописных цифр на выборке MNIST, задаче классификации объектов выборки CIFAR10, а также для семантической сегментации сканированных изображений исторических документов. Проведенные эксперименты показали, что снижение качества распознавания зависит от задачи и используемой модели, оставляя их пригодными для применения. При этом в зависимости от допустимого снижения качества автор предлагает выбрать число преобразуемых слоев и использовать частично биполярную морфологическую модель, сочетающую вычислительную эффективность и высокое качество.

В заключении приводятся основные результаты диссертации. В приложениях приводятся описания зарегистрированных программ для ЭВМ и акты о внедрении результатов диссертационной работы.

Достоверность результатов диссертации подтверждается математическими выкладками. Научные результаты обосновываются вычислительным экспериментом и находятся в соответствии с результатами, полученными другими авторами.

Диссертация обладает внутренним единством, написана хорошим языком, содержит подробный анализ существующих работ и богатую библиографию, включающую актуальные публикации последних лет. Приведенные результаты представляют значительный теоретический и практический интерес, так как позволяют упростить нейросетевые модели на специализированных устройствах и ПЛИС при теоретическом сохранении выразительной способности модели и незначительном контролируемом снижении качества распознавания на практике. Устройства этих типов широко применяются в промышленности и быту, все чаще включаются в конструкции массового применения, например, многие современные смартфоны и планшеты оснащаются нейроморфными сопроцессорами. В частности, методы, развитые в диссертации, потенциально применимы для улучшения функциональных характеристик нейросетевых модулей цифровой образовательной среды ПиктоМир разработки ФГУ ФНЦ НИИСИ РАН, эксплатируемой на планшетах низшего ценового диапазона.

Основные научные результаты по теме диссертации опубликованы в 10 работах, среди них: 1 статья в журнале, рекомендованном ВАК РФ (перечень ВАК 2022), 3 статьи в журналах, индексируемых международными базами Web of Science и Scopus, 6 публикаций в трудах российских и международных конференций. Зарегистрировано 2 программы для ЭВМ.

Замечания. В диссертации может быть отмечены следующий недостаток: в работе не хватает детального описания параметров реального ПЛИС-устройства, на котором можно получить оценки сложности/латентности, идентичные приведенным в работе. Такое описание позволило бы уточнить класс устройств, на которых результаты диссертационной работы могут успешно применяться на практике.

Указанный недостаток не является существенным и не влияют на общую положительную оценку диссертационной работы. Результаты Лимоновой Е.Е. представляют значительный интерес и с теоретической и с практической точек зрения.

Заключение. Диссертационная работа Лимоновой Е.Е. «Биполярная морфологическая аппроксимация нейрона для уменьшения вычислительной сложности глубоких сверточных нейронных сетей» является законченной научно-исследовательской работой и соответствует критериям, установленным пп. 9-14 Положения о присуждении ученых степеней, утвержденного Постановлением Правительства РФ от 24.09.2013 № 842

(ред. от 26.09.2022). Работа соответствует паспорту специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ» для технических наук, в частности, пункту 2 (Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий), пункту 3 (Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента), пункту 7 (Качественные или аналитические методы исследования математических моделей) и пункту 9 (Постановка и проведение численных экспериментов, статистический анализ их результатов, в том числе с применением современных компьютерных технологий).

Таким образом, Лимонова Елена Евгеньевна заслуживает присуждения ученой степени кандидата технических наук по специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ».

Отзыв обсужден и утвержден на семинаре отдела учебной информатики Федерального научного центра Научно-исследовательский институт системных исследований Российской академии наук с участием сотрудников отдела ЦОНТ ФГУ ФНЦ НИИСИ РАН, протокол № 1 от 20 января 2023 г.

Зав. отделом учебной информатики
ФГУ ФНЦ НИИСИ РАН, _____ / Кушниренко А. Г.
кандидат физико-математических наук « 26» января 2023 г.

Адрес ведущей организации:
17218, Москва, Нахимовский просп., 36, к.1.
Федеральное государственное учреждение
«Федеральный научный центр
Научно-исследовательский институт
системных исследований
Российской академии наук»
e-mail: niisi@niisi.msk.ru

Подпись руки Кушниренко А.Г.
заверяю
инспектор отдела кадров
(М.В. Гуреева)

ОТДЕЛ КАДРОВ

26.01.2023