МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РЕАЛИЗАЦИЯ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ОПТИМИЗАЦИИ ГРАДИЕНТНОГО СПУСКА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование
информационных систем
факультета компьютерных наук и информационных технологий
Солостовского Василия Сергеевича

Научный руководитель:
к.ф.-м.н., доцент _____ Огнева М. В.

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М.В.

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Одним из ключевых этапов обучения моделей машинного обучения, особенно в контексте глубоких нейронных сетей, является оптимизация их параметров. Наиболее распространённым и эффективным подходом к решению этой задачи остаётся градиентный спуск и его многочисленные модификации. Классический стохастический градиентный спуск (SGD) достаточно прост в реализации и показывает хорошие результаты, однако он не лишён недостатков — в частности, нестабильности при выборе шага обучения и медленной сходимости в сложных пространствах параметров. В связи с этим, в последние годы активно разрабатываются адаптивные методы оптимизации, которые призваны устранить эти ограничения и повысить эффективность обучения моделей.

Наиболее известным и широко используемым из таких адаптивных методов является алгоритм Adam [1, 2, 3], предложенный Д. Кингмой и Дж. Ба. Он так же не лишен недостатков, поэтому на его основе возникло множество новых алгоритмов, которые стремятся устранить, например, чрезмерное накопление моментов или отсутствие гарантированной сходимости. В числе таких улучшенных методов можно выделить Nadam, AMSGrad, AdaBelief [4], SuperAdam [5] и AdEMAMix [6]. Каждый из них вносит те или иные изменения в процедуру адаптивного обновления параметров модели, опираясь на различные эвристики и теоретические основания.

Актуальность настоящего исследования обусловлена необходимостью глубокого анализа и сравнения современных методов оптимизации, применяемых в задачах машинного обучения. Несмотря на большое количество предлагаемых модификаций градиентного спуска, в практике разработки часто возникают вопросы выбора наиболее подходящего метода с учётом конкретной задачи, объёма данных и структуры модели. Проведение сравнительного анализа этих методов на различных датасетах позволит

выявить их сильные и слабые стороны и определить области наилучшего применения.

Цель бакалаврской работы – реализация и сравнительный анализ современных адаптивных методов оптимизации градиентного спуска: Adam, Nadam, AMSGrad, AdaBelief, SuperAdam и AdEMAMix.

Поставленная цель определила **следующие задачи**:

1. изучить теоретические основы метода градиентного спуска и его адаптивных модификаций;
2. провести обзор существующих методов и проанализировать их математическую модель;
3. реализовать каждый из выбранных методов с применением современных библиотек машинного обучения;
4. провести экспериментальное исследование работы методов на нескольких открытых датасетах;
5. выполнить сравнительный анализ по критериям показаний метрик и скорости обучения;
6. сформулировать рекомендации по выбору оптимизатора для разных типов данных.

Методологические основы методов оптимизации градиентного спуска представлены в работах Kingma D.P., Ba J. Adam, Dozat T., Reddi S.J., Kale S., Kumar S., Zhuang J., Huang F., Li J., Huang H., Pagliardini M., Ablin P., Grangier D.

Практическая значимость бакалаврской работы. Среди множества существующих методов оптимизации градиентного спуска данный анализ помогает выбрать, какой из методов использовать при различных входных условиях, рассказывает про их достоинства и недостатки.

Структура и объём работы. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и трёх приложений. Общий объем работы – 94 страницы, из них 59 страниц –

основное содержание, включая 12 рисунков и 9 таблиц, бумажный носитель в качестве приложения, список использованных источников информации – 33 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы обучения нейронных сетей» посвящен ознакомлению с понятием градиентного спуска, принципом его работы. Рассматривается стохастическая (SGD) и адаптивные модификации градиентного спуска, такие как AdaGrad, Adam, Nadam, AMSGrad, AdaBelief, SuperAdam, AdEMAMix. Сравнительный анализ методов оптимизации реализован в сопоставлении их характеристик, выявлении ключевых различий в механизмах и преимуществ методов в различных контекстах обучения и ограничения, которые необходимо учитывать при выборе оптимизатора для конкретной задачи. Сравнительный анализ показал, что каждый из методов демонстрирует свою силу в определённых условиях. Если важна строгость — выбирается AMSGrad. Если скорость и гибкость — Nadam или AdEMAMix. Если важно качество обобщения — AdaBelief. Если требуется контроль и эксперимент — SuperAdam. Итоговый выбор зависит от характера данных, архитектуры модели, доступных вычислительных ресурсов и требований к результату.

Второй раздел «Сравнение методов оптимизации» посвящен сравнению и анализу методов SGD, Adam, NAdam, AMSGrad, AdaBelief, SuperAdam, AdEMAMix на разных типах данных (изображения, числовые данные, текстовые данные), разных архитектурах нейронных сетей, разных значениях скорости обучения и количества эпох. В работе использовались библиотеки pytorch (позволяет заменять оптимизирующую функцию проще всего среди остальных библиотек подобного вида), pytorch.ignite (высокоуровневая библиотека с реализацией подсчета различных метрик), numpy (библиотека для работы с матрицами), pandas (библиотека для работы

с табличными данными), matplotlib (библиотека для построения двумерных и трехмерных графиков), sklearn (применяется на этапе препроцессинга данных), adabelief_pytorch (библиотека с реализацией метода AdaBelief), natasha (библиотека для препроцессинга текста на русском языке).

Для подготовки к сравнительному анализу методов изображения переведены в числовые массивы, числовые данные нормализованы, текст векторизован, составлены архитектуры нейронных сетей различной сложности, реализованы функции обучения и тестирования моделей с подсчетом метрик. Каждый из методов (SGD, Adam, NAdam, AMSGrad, AdaBelief, SuperAdam, AdEMAMix) обучил модели на разных комбинациях архитектуры нейронной сети, типов данных и гиперпараметров. Для каждого случая построены таблицы со значениями метрик и графики функции потерь, на основании которых проведен сравнительный анализ.

В пункте **2.1** проводится исследования на наборе данных Fashion MNIST – изображения товаров Zalando. Он представляет набор собой набор из полутоновых черно-белых картинок размером 28 на 28 пикселей, цвет указывается значением от 0 до 255. У каждого изображения есть своя метка, соответствующая одному из десяти типов: футболка/топ, брюки, пуловер, платье, пальто, сандалия, рубашка, кроссовок, сумка, полуботинок. Исследование проведено при скорости обучения 10^{-3} и 10^{-4}, количестве эпох 5 и 20 соответственно. При большей скорости обучения все методы, кроме SGD, переобучились. Метод AdEMAMix обучал модель дольше всех. При меньшей скорости методы SGD и SuperAdam показали худшие результаты по сравнению с остальными методами. Метод AdaBelief чувствителен к гиперпараметру ε, но его довольно просто подобрать. У метода SuperAdam несколько гиперпараметров, которые очень сложно подбирать, поэтому он не смог показать хороший результат. Затем проведено исслудование на усложненной по структуре нейронной сети. Метод AdEMAMix показал лучшую точность, затратив наибольшее время на обучения.

В пункте **2.2** проводится исследования на наборе данных, представляющих собой измерения метеостанциями температуры воздуха и других параметров в разных точках местности – абсолютная высота рельефа в данной точке местности, топографичекий индекс влажности, экспозиция склона, светотеневая отмывка, степень расчлененности рельефа, угол наклона склона, температура поверхности Земли, измеренная при давлении 1000 гектопаскалей, температура воздуха на высоте 2-х метров, дата, время, географическая широта и долгота. Исследование проведено при скорости обучения 10^{-3} и количестве эпох 300. Такое сочетание является самым наглядным. В результате метод AdEMAMix достиг наилучшей точности, затратив наибольшее время на обучение, AMSGrad и Adam оказались оптимальными методами – они выигрывают по скорости и не отстают по точности. Метод AdaBelief плохо показал себя на числовых данных. AdeMAMix имеет краткосрочную и долгосрочную память, из-за чего позволяет себе сделать шаг больше в попытке найти лучший результат. Из-за этих шагов его график выглядит колеблющимся. При усложнении структуры нейронной сети результат всех моделей ухудшился, а графики колеблются из-за сложной структуры функции потерь. Метод AdEMAMix обучил модель быстрее всех, поэтому в конце произошло переобучение.

В пункте **2.3** проводится исследования на данных, представляющих собой набор русскоязычных научных статей. Для каждой стать указаны название, список ключевых слов, категорию, которую нужно предсказать по остальным данным, журнал, в котором была опубликована статья, год публикации, количество просмотров статьи, количество загрузок статьи, краткое описание статьи. Данные для обучения подготовлены при помощи библиотеки natasha. Каждое описание научной статьи приведено к нижнему регистру, разделено на слова и лемматизировано, то есть каждое слово приведено к начальной форме. Затем при помощи библиотеки sklearn описания векторизованы (каждому описанию соответствует набор чисел) и

разделены на тренировочный и тестовый.

Сначала проведено исследование на первом варианте нейронной сети, скорости обучения 10^{-5} и количестве эпох 200. Метод AdEMAMix достиг наилучшего результата, после чего переобучился, из-за чего показал худший результат. В предыдущих исследованиях метод SGD показывал наихудшую точность, а в данном случае он достиг хорошего результата наравне с остальными моделями. Лучшие результаты показали методы Adam и Nadam. При скорости обучения 10^{-4} и количестве эпох 30 все модели показали хорошую точность, лучше, чем при большем количестве эпох и меньшей скорости обучения. Для метода SuperAdam удалось найти хорошие, но не идеальные гиперпараметры. Метод Nadam имеет наименьшее значение функции потерь, но по метрикам лучшим является AdaBelief.

Затем исследовалось поведение методов на более сложной по структуре нейронной сети. При скорости 10^{-5} и количестве эпох 200 все модели переобучаются. Меньше всех подвержен проблеме переобучения метод AMSGrad. Метод AdEMAMix больше всего подвержен переобучению, но лучший результат точности у него близок к результату AMSGrad. При скорости 10^{-6} и количестве эпох 50 метод AdEMAMix справился с задачей лучше остальных методов, методы Adam, AMSGrad, SGD показали хорошую точность и быстрое время обучения, метод NAdam показал неудовлетворительный результат наравне с SuperAdam.

Проведя исследования на трёх видах данных, очевидно, что методы AMSGrad и Adam являются оптимальными, поскольку они выдают хорошую точность и являются самыми быстрыми. Благодаря своей стабильности и осторожным шагам они достигают почти лучшего результата метрик.

Метод AdeMAMix отличается лучшими результатами во всех трёх исследованиях, однако у него есть недостаток – затрачивается больше всего времени на обучение. Это объясняется сложностью алгоритма и количеством данных, которые он хранит.

Из рассмотренных методов SuperAdam показывает самый плохой результат по метрикам. Он имеет больше всего гиперпараметров и очень чувствителен к ним, поэтому его результат зависит от тщательности подбора гиперпараметров.

Метод Nadam сравним по времени с AMSGrad благодаря своей простоте, но сильно проигрывает ему в метриках из-за слишком простого алгоритма.

На данных с изображениями и числовых данных SGD показал плохую точность, но на данных с текстом он получает удовлетворительный результат.

AdaBelief показал лучший результат на первом наборе данных с изображениями, но на табличных данных он считается довольно долго и имеет плохую точность (относительно победителей), поскольку этот метод был создан преимущественно для данных с изображениями.

ЗАКЛЮЧЕНИЕ

В ходе проведенной работы было проведено сравнение методов оптимизации градиентного спуска SGD, Adam, NAdam, AMSGrad, AdaBelief, SuperAdam, AdEMAMix. Для этого были изучены теоретические основы метода градиентного спуска и его адаптивных модификаций, реализован каждый из выбранных методов с применением современных библиотек машинного обучения, проведено экспериментальное исследование работы методов на трёх открытых различных датасетах – изображения (Fashion MNIST), количественные числовые измерения (измерения параметров местности для прогноза температуры) и текстовые данные (набор русскоязычных статей) с помощью спроектированных и обученных нейронных сетей, выполнен сравнительный анализ по критериям точности и скорости обучения.

Анализ показал, что для работы с изображениями стоит использовать метод AdaBelief, он продемонстрировал наилучший результат на данных с изображениями, однако на числовых и текстовых данных его результаты хуже, чем методы AMSGrad и Adam, которые показали себя самыми быстрыми и достаточно результативными на всех трех наборах данных.

Метод AdEMAMix хотя и достигает наилучших результатов по точности, но работает дольше остальных алгоритмов, поэтому его надо использовать в задачах, где очень важна точность и не важно время.

SuperAdam затруднён в использовании из-за того, что он очень чувствителен к своим гиперпараметрам, которые надо подбирать тщательно и осторожно.

Методы NAdam и SGD достаточно быстрые, но в большинстве случаев дают неконкурентные результаты.

Основные источники информации:

1. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization [Электронный ресурс]. – URL: https://arxiv.org/abs/1412.6980 (дата обращения: 16.05.2025).

2. Dozat T. Incorporating Nesterov Momentum into Adam [Электронный ресурс]. – URL: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ (дата обращения: 16.05.2025).

3. Reddi S.J., Kale S., Kumar S. On the Convergence of Adam and Beyond [Электронный ресурс]. – URL: https://arxiv.org/abs/1904.09237 (дата обращения: 16.05.2025).

4. Zhuang J., et al. AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients [Электронный ресурс]. – URL: https://arxiv.org/abs/2010.07468 (дата обращения: 17.05.2025).

5. Huang F., Li J., Huang H. SUPER-ADAM: Faster and Universal Framework of Adaptive Gradients [Электронный ресурс]. – URL: https://arxiv.org/abs/2106.08208 (дата обращения: 17.05.2025).

6. Pagliardini M., Ablin P., Grangier D. The AdEMAMix Optimizer: Better, Faster, Older [Электронный ресурс]. – URL: https://arxiv.org/abs/2409.03137 (дата обращения: 18.05.2025).