МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Иванов Дмитрий Александрович

Нейроморфные методы оптимизации систем искусственного интеллекта для задач обучения с подкреплением

Специальность 2.3.5
Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2025

Работа выполнена на кафедре суперкомпьютеров и квантовой информатики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

Научный руководитель: **Воеводин Владимир Валентинович**
доктор физико-математических наук, профессор, член-корреспондент РАН

Официальные оппоненты: **Дьяконов Александр Геннадьевич,**
доктор физико-математических наук, доцент, профессор РАН, ООО «ВК», руководитель отдела

Хохлов Николай Игоревич,
доктор физико-математических наук, доцент, кафедра информатики и вычислительной математики Московского физико-технического института, заведующий кафедрой

Петровский Михаил Игоревич,
кандидат физико-математических наук, доцент, кафедра интеллектуальных информационных технологий факультета вычислительной математики и кибернетики МГУ имени М.В.Ломоносова, доцент

Защита состоится «19» декабря 2025 г. в 11 часов 00 минут на заседании диссертационного совета МГУ.012.2 Московского государственного университета имени М.В. Ломоносова по адресу: 119991, Москва, ГСП-1, Ленинские горы, МГУ, д. 1 строение 52, факультет Вычислительной математики и кибернетики, аудитория №238.

E-mail: ds012.2@cs.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский проспект, д.27) и на портале: https://dissovet.msu.ru/dissertation/3656.

Автореферат разослан «___» _____ 2025 года.

Ученый секретарь
диссертационного совета МГУ.012.2,
кандидат физико-математических наук А.С. Антонов

Общая характеристика работы

Актуальность работы. Современные *системы искусственного интеллекта* (системы ИИ), представляющие собой комбинацию алгоритмов ИИ и аппаратного обеспечения, в большинстве случаев построены на основе глубоких нейронных сетей (НС) и компьютеров на архитектуре фон Неймана. При этом, по сравнению с мозгом человека современные системы ИИ значительно менее энергоэффективны: при меньшем количестве нейронов и связей они потребляют значительно больше энергии. Мозг человека, состоящий приблизительно из 80-100 миллиардов нейронов и 1000 триллионов связей (синапсов), потребляет около 20 Вт. Для сравнения, современная высокопроизводительная видеокарта Nvidia H100 с энергопотреблением 700 Вт имеет память объемом 80 ГБ, достаточную для моделирования лишь 80 миллиардов связей. Таким образом, моделирование на четыре порядка меньшего количества связей требует на порядок большего энергопотребления. Как следствие, высокое энергопотребление ограничивает применение нейронных сетей в робототехнике и других встраиваемых системах.

Столь высокое энергопотребление обусловлено значительными энергозатратами на обращение к основной памяти по сравнению с вычислительными операциями, что является следствием разделения памяти и вычислений, характерного для большинства современных вычислительных архитектур. Кроме того, разделение вызывает большие временны́е задержки при обращении в память и ограничивает общую производительность вычислительной системы из-за узкого канала передачи данных между памятью и вычислителем. Данные проблемы часто в совокупности называются *проблемой бутылочного горлышка фон Неймана*. Стоит отметить, что многие алгоритмы, лежащие в основе нейронных сетей, относятся к алгоритмам с низкой вычислительной интенсивностью, что еще больше усиливает проблему бутылочного горлышка фон-неймановской архитектуры в системах ИИ. Особенно остро проблема обращений в память проявляется в ситуациях, когда размер пакета входных данных НС равен единице, что приводит к еще меньшей вычислительной интенсивности. Характерным примером таких ситуаций являются задачи управления, в частности, *задачи обучения с подкреплением*.

Обучение с подкреплением имеет множество приложений в реальном мире, таких как робототехника, управление сложными устройствами (к примеру, удержание плазмы в токамаке в реальном времени), игры, биржевая торговля и другие. Эти приложения часто накладывают требования малого времени отклика и высокой частоты работы нейронных сетей, тренированных с помощью методов обучения с подкреплением. Применение НС для задач удержания плазмы в токамаке требует частоты их работы в 100 кГц, а при автономном управлении высокоманевренным квадрокоптером необходимы частоты их работы не менее 100 Гц. Время отклика НС в десятки наносекунд необходимо при биржевой торговле. При попытках использования в подобных задачах нейронных сетей с

миллиардами связей, их максимальная частота работы существенно снижается (до нескольких Гц), что вынуждает применять либо малые по размерам сети, либо ограничиваться низкой частотой их работы. Таким образом, алгоритмические свойства НС в сочетании с особенностями фон-неймановской архитектуры накладывают серьезные ограничения на применение больших нейронных сетей в задачах обучения с подкреплением.

Для преодоления обозначенных выше ограничений исследователи предлагают как алгоритмические методы оптимизации нейронных сетей, так и принципиально новые аппаратные решения. Существуют различные методы алгоритмической оптимизации нейронных сетей, такие как структурная разреженность (прюнинг), квантование, дистилляция знаний, поиск вычислительно эффективных архитектур НС. К сожалению, эти методы далеко не всегда учитывают взаимосвязь аппаратного обеспечения и алгоритмов, что затрудняет практическое применение данных методов. Одновременно с этим существует крайне малое количество научных работ, посвященных применению методов оптимизации к нейронным сетям, тренированных методами обучения с подкреплением. Также, в последние годы активно развиваются аппаратные методы оптимизации и предлагаются новые вычислительные архитектуры, такие как не фон-неймановские вычислители, направленные на преодоление ограничений современного аппаратного обеспечения для нейронных сетей.

Одним из наиболее перспективных направлений являются нейроморфные (биологически подобные) архитектуры и методы, стремящиеся имитировать алгоритмически или аппаратно некоторые принципы функционирования мозга человека, что может способствовать повышению энергоэффективности, скорости работы и масштабируемости систем искусственного интеллекта. К данному направлению относятся: импульсные нейронные сети, «вычисления в памяти»/«рядом с памятью», асинхронное исполнение нейронных сетей, поддержка разреженных и аналоговых вычислений.

Таким образом, поиск аппаратных и алгоритмических методов для оптимизации инференса систем искусственного интеллекта, основанных на нейронных сетях, тренированных с помощью методов обучения с подкреплением, представляется крайне значимой и перспективной задачей. В данной работе предлагается использовать нейроморфные методы для её решения.

В диссертации делаются акценты на следующих вопросах:

1. Анализ причин низкой эффективности в аспекте энергопотребления и скорости работы существующих систем искусственного интеллекта, сравнение принципов их работы с принципами функционирования мозга человека.
2. Разработка, применение и анализ эффективности нейроморфных методов для оптимизации систем ИИ для задач обучения с подкреплением.

Цели и задачи работы. Целью данной работы является разработка нейроморфных методов, позволяющих существенно ускорить и понизить

энергозатратность систем искусственного интеллекта для задач обучения с подкреплением.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Проанализировать принципы работы современных систем искусственного интеллекта с целью выявления их узких мест и проблем. Исследовать взаимосвязь работы алгоритмов нейронных систем с современными аппаратными платформами. Провести анализ существующих аппаратных и программных методов оптимизации нейронных сетей.
2. Выделить принципы функционирования мозга человека и их связь с системами ИИ. Оценить их преимущества и недостатки для их возможной имплементации в современные системы ИИ.
3. Исследовать особенности работы нейронных сетей, тренированных с помощью методов обучения с подкреплением.
4. Разработать алгоритмы оптимизации нейронных сетей, тренированных методами обучения с подкреплением, на основе нейроморфных методов: с помощью комбинации временно́й и структурной разреженности и на основе комбинации структурной разреженности и квантования.
5. Программно реализовать описанные подходы, проанализировать их эффективность и возможность потенциальной имплементации в аппаратных вычислительных платформах. Провести сравнение с существующими решениями.

Научная новизна:

1. Проведено детальное сравнение принципов работы мозга человека с принципами работы систем ИИ на основе фон-неймановских и на основе не фон-неймановских нейроморфных вычислителей. На основе проведенного сравнения предложена классификация принципов работы мозга человека и проведен анализ на предмет их имплементируемости в современных системах ИИ.
2. Впервые предложен алгоритм на основе комбинации структурной разреженности и квантования для оптимизации нейронных сетей, тренированных методами обучения с подкреплением. Метод на 1 - 2 порядка (вплоть до 400 раз) уменьшает занимаемую нейронными сетями память без потери качества работы, что позволяет размещать нейронные сети в быстрой памяти или снижать число обращений в память.
3. Впервые предложен алгоритм оптимизации нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации двух видов разреженности: структурной и временно́й. Метод уменьшает на 1 - 2 порядка число обращений в память и число необходимых арифметических операций при инференсе нейронных сетей при сохранении качества работы. Данный метод обладает еще одним преимуществом,

позволяя нейронным сетям работать в асинхронном режиме, что повышает их потенциальную масштабируемость.

Теоретическая и практическая значимость.

В диссертационной работе проведено сравнение принципов работы мозга человека и современных систем искусственного интеллекта, построенных на основе как фон-неймановских вычислителей, так и не фон-неймановских нейроморфных архитектур. Показано, что современные системы ИИ на основе фон-неймановских вычислителей не используют ряд ключевых принципов работы мозга человека. Это обуславливает низкую энергоэффективность, масштабируемость и скорость работы современных систем ИИ. Рассмотрены возможности реализации части принципов работы мозга человека в современных системах ИИ для улучшения их энергоэффективности и скорости работы.

Практическая значимость работы связана с повышением эффективности инференса систем ИИ для задач обучения с подкреплением с помощью нейроморфных подходов. Были предложены новые и развиты существующие подходы на основе структурной разреженности, временно́й разреженности и квантования для оптимизации нейронных сетей, тренированных методами обучения с подкреплением. С помощью данных методов была показана возможность уменьшать размеры нейронных сетей на 1 - 2 порядка и уменьшать на 1 - 2 порядка число арифметических операций. Была показана связь этих методов оптимизации с некоторыми принципами работы мозга человека.

Создан программный комплекс, который позволяет проводить эксперименты для изучения и тестирования методов оптимизации нейронных сетей, тренированных методами обучения с подкреплением.

Методология и методы исследования. При получении основных результатов диссертационной работы использовались методы обучения с подкреплением, глубокое машинное обучение, методы системного и сравнительного анализа. Использовались методы программирования на языке Python и C++.

Основные положения, выносимые на защиту:

1. Нейроморфные методы и подходы к оптимизации систем искусственного интеллекта для задач обучения с подкреплением на основе свойств и принципов функционирования мозга человека с целью повышения энергоэффективности, пропускной способности, масштабируемости и скорости работы современных систем ИИ.
2. Метод оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации структурной разреженности и квантования. Метод уменьшает на 1 - 2 порядка (вплоть до 400 раз) размеры нейронных сетей без потери качества работы, что позволяет размещать нейронные сети в быстрой памяти или уменьшать число обращений в память.
3. Метод оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации временно́й и структурной разреженности. Метод уменьшает на 1 - 2 порядка число

обращений в память и число арифметических операций при инференсе нейронных сетей без потери качества работы. Введенная в методе асинхронность нейронов дает возможность обеспечения большей масштабируемости.

Апробация работы. Представленные в работе результаты докладывались на следующих научных конференциях и семинарах:

1. Научная конференция «Тихоновские чтения» 2023, Москва, Россия, 29 октября - 3 ноября 2023.
2. Всероссийская конференция «Ломоносовские чтения» 2022, Москва, Россия, 14 - 22 апреля 2022.
3. Научный семинар кафедры Интеллектуальных информационных технологий ВМК МГУ, 2024.
4. Научный семинар по машинному обучению под руководством проф. А.Г. Дьяконова, Центральный университет, 2024.
5. Научный семинар «Методы машинного обучения в автоматической обработке текстов», НИВЦ МГУ, 2024.

Личный вклад. Все результаты работы получены автором лично под научным руководством д.ф.-м.н., чл.-корр. РАН Воеводина Владимира Валентиновича. В работах, написанных в соавторстве, вклад автора диссертации является определяющим.

В работе [А.1] автором выполнен анализ принципов работы современных вычислительных систем и проведено сравнение с принципами работы мозга человека. На основе этого предложена классификация принципов работы мозга человека и проанализирована их реализация в нейроморфных системах ИИ. Работа опубликована в журнале Frontiers in Neuroscience [А.1].

В работе [А.2] автором предложен метод оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации структурной разреженности и квантования. Все эксперименты выполнены лично автором. Работа опубликована в журнале Scientific Reports [А.2].

В работе [А.3] предложен метод оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации структурной и временно́й разреженности. Все эксперименты выполнены лично автором. Работа опубликована в журнале Scientific Reports [А.3].

Публикации. Основные результаты по теме диссертации изложены в 3 публикациях [А.1—А.3], изданных в рецензируемых научных изданиях, определенных в п. 2.3 Положения о присуждении ученых степеней в Московском государственном университете имени М. В. Ломоносова.

Объем и структура работы. Диссертация состоит из введения, 4 глав, заключения и 1 приложения. Полный объем диссертации составляет 116 страниц, включая 34 рисунка и 10 таблиц. Список литературы содержит 110 наименований.

Содержание работы

Во **введении** обосновывается актуальность работы, ставится её цель и задачи, излагается научная новизна, теоретическая и практическая значимость, сформулированы положения, выносимые на защиту, и личный вклад автора.

Первая глава посвящена описанию современных систем ИИ для задач обучения с подкреплением, их проблемам и недостаткам, и существующим методам решения данных проблем.

Сначала обосновывается необходимость рассмотрения алгоритмов ИИ и аппаратного обеспечения в совокупности, вводится концепция системы ИИ, которая объединяет эти два понятия.

В *разделе 1.1* представлено описание архитектуры фон Неймана, её ключевых компонентов и основных ограничений. Рассматриваются узкие места этой архитектуры, включая ограниченную пропускную способность шины данных, разницу в скорости работы памяти и процессора, а также энергетические ограничения. Приводятся методы смягчения проблем архитектуры фон Неймана такие как кэширование, конвейеризация, спекулятивные вычисления, мультипоточность, использование памяти с высокой пропускной способностью (HBM) и «вычисления в памяти» («in-memory computations») или «рядом с памятью» («near-memory computations»).

В *разделе 1.2* описаны современные нейронные сети и их ключевые компоненты. Рассмотрены основные типы слоев нейронных сетей с точки зрения вычислений, включая полносвязный, сверточный и рекуррентный слои. Описано влияние размера пакета (batch) на вычислительную интенсивность нейронных сетей. Показано, что вычислительную основу большинства слоев составляют матричные операции.

В *разделе 1.3* описана проблема исполнения нейронных сетей на архитектуре фон Неймана. Основное внимание уделено операциям умножения матриц, которые являются ключевыми для нейронных сетей, и связанными с ними ограничениям по доступу к памяти. Представлен анализ эффективности использования весов и входных данных при различных размерах пакета, с акцентом на критичность этой проблемы для задач управления в реальном времени, при которых размер пакета входных данных равен единице.

Раздел 1.4 посвящен описанию принципов работы основных классов аппаратного обеспечения для задач искусственного интеллекта, включая CPU, GPU и TPU. Представлены особенности архитектуры и методы оптимизации каждого типа вычислителей для выполнения операций с нейронными сетями. Указаны недостатки данных архитектур и проблематичность эффективного инференса нейронных сетей в задачах управления.

В *разделе 1.5* описывается задача обучения с подкреплением. Рассматриваются основные понятия, математическая формулировка, концепция марковского

процесса, уравнения Беллмана. Указывается, что для задач обучения с подкреплением при инференсе характерен размер пакета, равный единице. Подчеркивается важность быстрого и энергоэффективного инференса нейронных сетей, тренированных методами обучения с подкреплением, для многих практических задач.

В *разделе 1.6* вводится идея нейроморфного подхода к разработке систем искусственного интеллекта, которая заключается в использовании некоторых принципов организации и функционирования мозга человека.

В *разделе 1.7* подводятся итоги главы, которые заключаются в том, что современные системы ИИ на базе фон-неймановских вычислителей и искусственных нейронных сетей на основе персептрона Розенблатта, имеют узкие места, связанные с низкой вычислительной интенсивностью НС и дорогими по времени и энергии обращениям в память. Подчеркивается, что особенно остро эта проблема проявляется при инференсе нейронных сетей в задачах обучения с подкреплением. В качестве решения предлагается заимствовать принципы работы мозга человека для повышения скорости работы и энергоэффективности систем ИИ.

Вторая глава посвящена анализу принципов функционирования и устройства мозга человека с целью выделения ключевых механизмов, которые могут быть успешно адаптированы и внедрены в системы ИИ, что позволит существенно повысить их энергоэффективность, скорость работы, масштабируемость и оптимизировать размер нейронных сетей.

В начале производится сравнение мозга человека с современными системами искусственного интеллекта. Подчеркивается превосходство мозга человека в энергоэффективности и количестве нейронных связей, и ставится вопрос о возможности создания более эффективных систем ИИ, использующих принципы работы мозга человека.

В *разделе 2.1* описывается структура и функционирование биологического нейрона. Представлены основные компоненты нейрона (тело, дендриты, аксон) и механизм передачи сигналов через синапсы. Особое внимание уделяется объяснению роли мембранного потенциала и потенциала действия в обработке и передаче информации в нервной системе.

В *разделе 2.1.1* представлена модель LIF (Leaky Integrate-and-Fire) нейрона как более биологически правдоподобная альтернатива модели нейрона Розенблатта. Описывается математическая формулировка LIF-модели, включая уравнение динамики мембранного потенциала и условие генерации потенциала действия. Указываются ограничения LIF-модели и обосновывается её популярность в компьютерных симуляциях нейронных сетей.

В *разделе 2.2* представлена классификация вычислительных принципов функционирования мозга человека. Производится их сравнение с механизмами функционирования современных систем ИИ, а также рассматриваются проекты вычислительных систем, основанные на предложенных принципах. Данная классификация была представлена автором в статье [А.1].

В *разделе 2.2.1* описывается концепция коннекционизма, основанная на представлении ментальных феноменов через призму нейронных сетей.

В *разделе 2.2.2* описывается концепция параллелизма в контексте нейронных сетей и обосновывается необходимость использования массивно-параллельных архитектур для эффективного функционирования искусственных нейронных сетей.

Раздел 2.2.3 посвящен импульсному характеру передачи информации в нейронных сетях. Представлена концепция импульсных нейронных сетей (Spiking Neural Networks, SNN), основанных на модели LIF-нейрона, где информация передается в виде элементарных событий - спайков. Рассмотрены преимущества SNN, включая асинхронность передачи данных и потенциальную энергоэффективность, а также недостатки, такие как сложность обучения, вычислительная сложность и более низкое качество работы. Анализируются причины энергоэффективности специализированных нейроморфных чипов, которые заключаются в наличии в них большого объема SRAM-памяти. Рассматривается возможность использования небинарных импульсов для повышения эффективности передачи информации в искусственных нейронных сетях.

В *разделе 2.2.4* рассматривается проблема асинхронности в контексте параллельных вычислений и нейронных сетей. Раскрывается преимущество асинхронного функционирования биологических нейронов, позволяющее полностью задействовать потенциал параллельной работы. Приведены примеры нейроморфных компьютеров, реализующих асинхронную архитектуру, такие как SpiNNaker, Loihi и NeuronFlow. Также обсуждаются ограничения асинхронных архитектур для нейронных сетей на основе персептрона Розенблатта, связанные с необходимостью их послойной синхронизации.

В *разделе 2.2.5* предлагается концепция активационной разреженности в нейронных сетях, которая заключается в активации небольшой части нейронов при «молчании» остальных. Данное явление отсутствует в искусственных нейронных сетях на основе персептрона Розенблатта, в которых активны все нейроны (за исключением нейронов с ReLU активацией), в то время как в мозге человека обычно активны менее 10% нейронов. Активационная разреженность поддерживается практически всеми процессорами, работающими с импульсными нейронными сетями, к примеру Loihi, TrueNorth, Tianjic и NeuronFlow. Для поддержки активационной разреженности в классических нейросетях был предложен экспериментальный чип EIE.

В *разделе 2.2.6* описывается концепция временно́й разреженности, заключающейся в обработке только измененной части входных данных, вместо их обработки «с нуля», характерной для современных нейронных сетей. Данная концепция схожа с идеей алгоритмов сжатия на основе компенсации движения, которые вместо хранения кадра целиком, хранят информацию о движении его частей. Временна́я разреженность может быть одним из источников (но не единственным) активационной разреженности. Концепция временно́й разреженности для нейронных сетей была реализована в алгоритме SpArNet и чипе NeuronFlow.

Раздел 2.2.7 посвящен структурной разреженности в биологических и искусственных нейронных сетях, характеризующейся отсутствием регулярных полносвязных слоев. Рассмотрены исследования по внедрению разреженности в искусственные нейронные сети. Такие чипы как Loihi, TrueNorth, Tianjic, NeuronFlow и EIE обладают способностью поддерживать структурную разреженность в импульсных и классических нейронных сетях.

В *разделе 2.2.8* описывается концепция квантованности в контексте работы мозга человека и нейронных сетей. Рассматриваются применения квантованных нейронных сетей. Представлены результаты исследований, указывающие на дискретную природу обработки информации в мозге человека и преимущества квантования для стабильности и устойчивости к шуму. Квантованность поддерживается практически всеми современными аппаратными платформами для искусственного интеллекта.

В *разделе 2.2.9* рассматривается аналоговая природа вычислений в биологических нейронах. Описываются преимущества аналоговых реализаций нейронов, включая их высокую скорость, энергоэффективность и естественную поддержку параллелизма. Упоминаются существующие проекты по разработке аналоговых чипов для искусственного интеллекта, такие как BrainScaleS и различные мемристорные разработки, однако отмечается отсутствие их практических применений.

Раздел 2.2.10 посвящен концепции вычислений в памяти и её реализации в биологических и искусственных нейронных сетях. Описывается принцип «один нейрон - один вычислитель», характерный для биологических нейронов, противопоставляемый традиционному подходу в цифровых устройствах на основе архитектуры фон Неймана, при котором множество нейронов совместно «используют» один и тот же вычислитель. Представлен гибридный подход «вычисления рядом с памятью», используемый в современных нейроморфных чипах. Данный подход характеризуется расположением весов нейронов в быстрой, но ограниченной по размерам SRAM-памяти, расположенной рядом с вычислителем. Приводятся примеры реализации подхода вычислений рядом с памятью в таких процессорах как EIE, Cerebras, Groq и NorthPole, демонстрирующих повышенную энергоэффективность и скорость работы благодаря использованию большого объема SRAM-памяти.

В *разделе 2.3* представлены выводы об использовании биологически подобных методов для систем искусственного интеллекта. Описываются преимущества квантования и структурной разреженности для оптимизации нейронных сетей, позволяющие значительно сократить их размеры, что дает возможность при наличии подходящего аппаратного обеспечения повысить энергоэффективность, скорость работы и уменьшить время отклика систем ИИ. Концепция вычислений рядом с памятью рассматривается как компромиссное решение проблемы бутылочного горлышка архитектуры фон Неймана, хорошо подходящая для задач, в которых требуется высокая скорость работы и низкое энергопотребление. Комбинация методов сжатия нейронных сетей и их последующее

расположение в SRAM-памяти рядом с вычислителем дает синергетический эффект, позволяя исполнять изначально большие нейронные сети с высокой скоростью и низкими затратами энергии. Описаны примеры комбинации этих подходов в современных аппаратных решениях, таких как NorthPole, EIE и Loihi, демонстрирующие значительное повышение производительности и энергоэффективности систем ИИ.

Третья глава посвящена разработке биологически подобных методов оптимизации систем ИИ для задач обучения с подкреплением на основе предложенных в предыдущей главе принципов работы мозга человека. Доступ к памяти является основной проблемой при инференсе в задачах обучения с подкреплением. Её можно решать либо путем уменьшения числа обращений к памяти, либо путем расположения нейронной сети в быстрой памяти.

В *разделе 3.1* приводится описание двух современных широкораспространенных алгоритмов глубокого обучения с подкреплением: SAC и DQN. Они будут использоваться в качестве алгоритмов обучения в предложенных методах оптимизации.

В *разделе 3.2* рассматриваются методы оптимизации нейронных сетей. В начале раздела описываются два основных классических подхода: квантование и структурная разреженность. Квантование представляет собой отображение непрерывного множества значений в дискретное множество. Применение квантования к нейронной сети позволяет уменьшить размер модели и упростить вычисления. Структурная разреженность достигается путем удаления (прюнинга, обрезания) малозначимых весов сети, что также приводит к уменьшению размера модели. Рассматриваются различные алгоритмы и критерии применения этих методов, а также особенности их применения в задачах обучения с подкреплением. В конце раздела рассматривается дельта-алгоритм, который позволяет использовать концепцию временно́й разреженности для оптимизации инференса нейронных сетей, последовательно обрабатывающих высококоррелированные по времени данные. Идея дельта-алгоритма заключается в распространении по нейронной сети только тех активаций нейронов, которые отличаются от активаций на предыдущем временном шаге не менее чем на заранее заданный порог.

В *разделе 3.3* описывается алгоритм оптимизации инференса нейронных сетей на основе комбинации структурной разреженности и квантования. Целью данного алгоритма является оптимизация обращений к памяти при инференсе нейронных сетей, тренированных методами RL. Оптимизация выполняется путем уменьшения обращений к памяти и/или расположения нейронных сетей в быстрой памяти. Для этого предлагается сократить размеры нейронной сети с помощью структурной разреженности и квантования. Данный подход был исследован автором в статье [А.2].

Алгоритм оптимизации (см. рис. 1, 2) состоит из двух частей: обучения с одновременным прореживанием нейронной сети и последующего квантования сети. Полученный алгоритм позволяет уменьшить число обращений к памяти

и/или расположить нейронную сеть в быстрой памяти благодаря сжатию нейронной сети в десятки, а иногда в сотни раз.

Обрезание сети выполнялось в процессе её обучения на основе модуля значения параметров с помощью градуального прюнинга в соответствии с расписанием (1):

$$s_t = s_f * \left(1 - \left(1 - \frac{t - t_0}{n\Delta t}\right)^3\right) for\ t \in \{t_0, t_0 + \Delta t, ..., t_0 + n\Delta t\} \qquad (1)$$

на шагах обучения, лежащих в интервале $[t_s, t_f]$, где $t_s = 0.2 \cdot T$ и $t_f = 0.8 \cdot T$, а T - число шагов обучения. После завершения основного обучения в течение дополнительных $0.2 \cdot T$ шагов сеть квантуется с донастройкой с помощью алгоритма QAT. Для сверточных слоев выполнялось асимметричное поканальное квантование, для полносвязных – асимметричное тензорное квантование. Полный алгоритм приведен ниже:

Алгоритм 1 Алгоритм оптимизации инференса нейронных сетей на основе комбинации структурной разреженности и квантования для задач обучения с подкреплением.

Входные данные:

$f_\theta(x)$ – необученная сеть с 32-битными параметрами
s_{final} – целевое значение разреженности
T – число шагов обучения
n – число итераций обрезания
L – частота измерения качества

Выходные данные:

$f_{\theta''}(x)$ – обученная квантованная сеть, M – маска весов

1: *// Этап 1: Обрезание*
2: $M \leftarrow \mathbf{1}(|\theta|)$ ▷ Инициализация маски, $|\theta|$ - количество параметров
3: **for** $t = 1$ to $T \cdot 0.2$ **do** ▷ Начальное обучение
4: train_RL_step(f_θ)
5: **end for**
6: **for** $i = 1$ to n **do** ▷ Фаза обрезания
7: $t \leftarrow t_{start} + i \cdot \Delta t$
8: $s_t \leftarrow s_{final} \cdot (1 - (1 - \frac{t - t_{start}}{n \cdot \Delta t})^3)$
9: **active_weights** $\leftarrow \theta[M \neq 0]$ ▷ Необрезанные веса

Рис. 1 — Псевдокод алгоритма оптимизации инференса нейронных сетей на основе комбинации структурной разреженности и квантования для задач обучения с подкреплением. Часть 1.

```
    k ← calculate_pruning_amount(s_t)
    k ← |θ| · (s_t − |active_weights|/|θ|)    ▷ k - количество обрезаемых весов
    threshold ← kabsmin(active_weights, k)    ▷ Поиск k-го минимального по модулю веса
    M[|θ| ≤ threshold] ← 0
    for step = 1 to Δt do
        train_RL_step(f_θ ⊙ M)                ▷ ⊙ - поэлементное умножение
    end for
end for
θ_best ← θ, R_best ← −∞
for t = t_final to T do                       ▷ Финальное обучение и поиск наилучшей сети
    train_RL_step(f_θ ⊙ M)
    if t mod L = 0 then
        R_current ← evaluate_performance(f_θ ⊙ M)
        if R_current > R_best then
            R_best ← R_current
            θ_best ← θ
        end if
    end if
end for
// Этап 2: Квантование
θ_best_quantized ← θ
R_best_quantized ← −∞
for t = t_final to T do                       ▷ Поиск наилучшей квантованной обрезанной сети
    train_RL_step_with_QAT(f_θ ⊙ M)
    if t mod L = 0 then
        R_current ← evaluate_performance(f_θ ⊙ M)
        if R_current > R_best_quantized then
            R_best_quantized ← R_current
            θ_best_quantized ← θ
        end if
    end if
end for
θ_result ← θ_best_quantized
return f_θ_result, M
```

Рис. 2 — Псевдокод алгоритма оптимизации инференса нейронных сетей на основе комбинации структурной разреженности и квантования для задач обучения с подкреплением. Часть 2.

В *разделе 3.4* предлагается алгоритм оптимизации инференса нейронных сетей на основе комбинации структурной и временно́й разреженностей. Целью данного алгоритма является уменьшение числа обращений к памяти и сокращение объема вычислений при инференсе НС, тренированных методами RL. Для этого предлагается обрабатывать только измененные области изображения и активировать только необходимые нейроны, а также сжать размер сети с помощью структурной разреженности. Так как задачи обучения с подкреплением обычно имеют высокую корреляцию между входными данными в соседние моменты времени, то применение временно́й разреженности путем обработки только измененных частей входных данных и активации только необходимых нейронов является логичным шагом. Данный подход был подробно исследован автором в работе [А.3].

Для получения структурной разреженности используется метод на основе Lottery Ticket Hypothesis (LTH, см. рис. 3). Для получения временно́й разреженности используется дельта-алгоритм (см. рис. 4).

Четвертая глава посвящена описанию результатов практических экспериментов по оптимизации нейронных сетей, тренированных методами обучения с подкреплением, с помощью алгоритмов, предложенных в третьей главе.

В *разделе 4.1* представлено описание тестовых сред (окружений) для оценки эффективности алгоритмов обучения с подкреплением. Рассматриваются два основных семейства сред, Atari games и MuJoCo, которые являются стандартными бенчмарками для измерения качества работы алгоритмов обучения с подкреплением. Окружения Atari games характеризуются высокоразмерным пространством наблюдений (изображениями), дискретным пространством действий и системой вознаграждений с задержкой. MuJoCo относится к классу окружений с континуальным управлением, где алгоритм генерирует команды в виде векторов вещественных чисел для контроля биоподобных механизмов. Состояния в MuJoCo также представлены векторами действительных чисел различной размерности.

В *разделе 4.2* представлены результаты работы алгоритма оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации структурной разреженности и квантования. В качестве методов обучения с подкреплением были выбраны два широкораспространенных алгоритма: SAC и DQN.

В качестве архитектур обучаемых сетей для алгоритма SAC был применен многослойный персептрон (MLP), в качестве окружений для обучения использовались следующие MuJoCo среды: Ant, Hopper, Swimmer, HalfCheetah, Humanoid и Walker. В качестве архитектур обучаемых сетей для алгоритма DQN были применены как классическая сверточная сеть из оригинальной работы, так и сеть ResNet. В качестве окружений для обучения были использованы следующией среды из семейства Atari games: Pong, Tutankham, Boxing и CrazyClimber. На рисунках 5, 6, 7 представлена производительность обрезанных и/или квантованных

Алгоритм 2 Этап 1: Структурная оптимизация (LTH)

Входные данные:

$f_\theta(x)$ – необученная нейронная сеть с параметрами θ
n – число итераций обрезания сети
p – доля удаляемых весов на итерации
T – число шагов обучения на итерации
D – пороговое значение для дельта-алгоритма

Выходные данные: Кортеж $\langle (f_{\theta_1}(x), M_1), (f_{\theta_2}(x), M_2), \ldots, (f_{\theta_n}(x), M_n) \rangle$, где $M_i \in \{0,1\}^{|\theta|}$

$M \leftarrow \mathbf{1}^{|\theta|}$ ▷ Инициализация маски единицами
for $i \leftarrow 1$ to n **do**
  $\theta_i \leftarrow \theta$ ▷ Копирование исходных весов
  $f_{\theta_i}(x) \leftarrow$ RL_train$(f_{\theta_i}(x), T, M)$ ▷ Обучение с учетом маски
  $k \leftarrow p \cdot |\theta|$ ▷ Число весов для обрезания
  active_weights $\leftarrow \theta[M \neq 0]$ ▷ Необрезанные веса
  threshold $\leftarrow$ kabsmin(**active_weights**, k) ▷ Поиск k-го минимального по модулю необрезанного веса
  $M[|\theta| \leq$ threshold$] \leftarrow 0$ ▷ Удаление k наименее значимых по модулю весов
  $M_i \leftarrow M$ ▷ Сохранение текущей маски
end for
return $\langle (f_{\theta_1}(x), M_1), (f_{\theta_2}(x), M_2), \ldots, (f_{\theta_n}(x), M_n) \rangle$ ▷ Кортеж обрезанных сетей

Рис. 3 — Псевдокод алгоритма оптимизации инференса нейронных сетей на основе комбинации структурной и временно́й разреженностей для задач обучения с подкреплением. Этап 1. Получение структурной разреженности.

нейронных сетей в различных средах для вышеперечисленных алгоритмов обучения с подкреплением и архитектур нейронных сетей.

На рис. 5 представлены результаты для MLP-сетей, обученных алгоритмом SAC, для сред MuJoCo. Для всех сред (за исключением HalfCheetah) за счет предложенного алгоритма можно обрезать до 98 процентов весов НС и квантовать оставшиеся без потери качества, что приводит к 200-кратному уменьшению размера оптимизированных нейронных сетей: 4-кратное уменьшение за счет квантования и 50-кратное за счет прореживания. Для HalfCheetah можно без потери качества обрезать 80 % весов НС и квантовать оставшиеся, что приводит к 20-кратному уменьшению размера нейронной сети. Для некоторых сред, таких как Hopper и Swimmer, можно обрезать 99 % весов НС и квантовать оставшиеся без потери качества работы, что приводит к 400-кратному уменьшению размера нейронной сети.

Алгоритм 3 Этап 2: Применение дельта-алгоритма для получения временно́й разреженности

// *Выполняем алгоритм для каждого нейрона в сети*
while True **do**
    $x_i^k(t) = Wait(predecessors)$ ▷ Получаем сигнал от предшественника i с предыдущего слоя k
    $o_j^{k+1}(t) \leftarrow o_j^{k+1}(t-1) + W_{ij} \times \Delta x_i^k(t)$ ▷ Пересчет внутреннего состояния нейрона
    $\Delta x_j^{k+1}(t) \leftarrow f(o_j^{k+1}(t)) - x\_prev_j^{k+1}(t)$ ▷ Пересчет изменения активации нейрона
    if $|\Delta x_j^{k+1}(t)| \geq D$ **then** ▷ Проверка превышения активации порога
        $x\_prev_j^{k+1}(t) = f(o_j(t))$ ▷ Обновление значения активации
        $Send(\Delta x_j^{k+1}(t), successors)$ ▷ Посылка изменения активации $\Delta x_j^{k+1}(t)$ из нейрона последующим нейронам
    end if
end while

Рис. 4 — Псевдокод алгоритма оптимизации инференса нейронных сетей на основе комбинации структурной и временно́й разреженностей для задач обучения с подкреплением. Этап 2. Получение временно́й разреженности.

На рис. 6 представлены результаты для классической DQN-сетей на основе CNN для сред Atari. Для всех сред с помощью предложенного алгоритма можно обрезать без потери качества до 80 процентов НС и квантовать оставшиеся веса, что приводит к 20-кратному уменьшению размера оптимизированных нейронных сетей. Для Pong и Tutankham можно обрезать и квантовать до 95 процентов весов НС, что приводит к общему 80-кратному уменьшению размера нейронных сетей.

На рис. 7 представлены результаты применения предложенного алгоритма для DQN-сетей на основе ResNet для сред Atari. Для сред Boxing и CrazyClimber можно обрезать до 90 процентов весов НС и квантовать оставшиеся веса с потерей качества не более чем в три процента, что приводит к 40-кратному уменьшению размера нейронных сетей. Для окружений Pong и Tutankham можно обрезать без потери качества до 98 процентов весов НС и квантовать оставшиеся, что приводит к 200-кратному уменьшению размера нейронных сетей. Стоит отметить, что нейронные сети на основе ResNet гораздо более пригодны для прюнинга и квантования.

В *разделе 4.3* представлены результаты работы алгоритма оптимизации инференса нейронных сетей, тренированных методами обучения с подкреплением, на основе комбинации структурной и временно́й разреженностей.

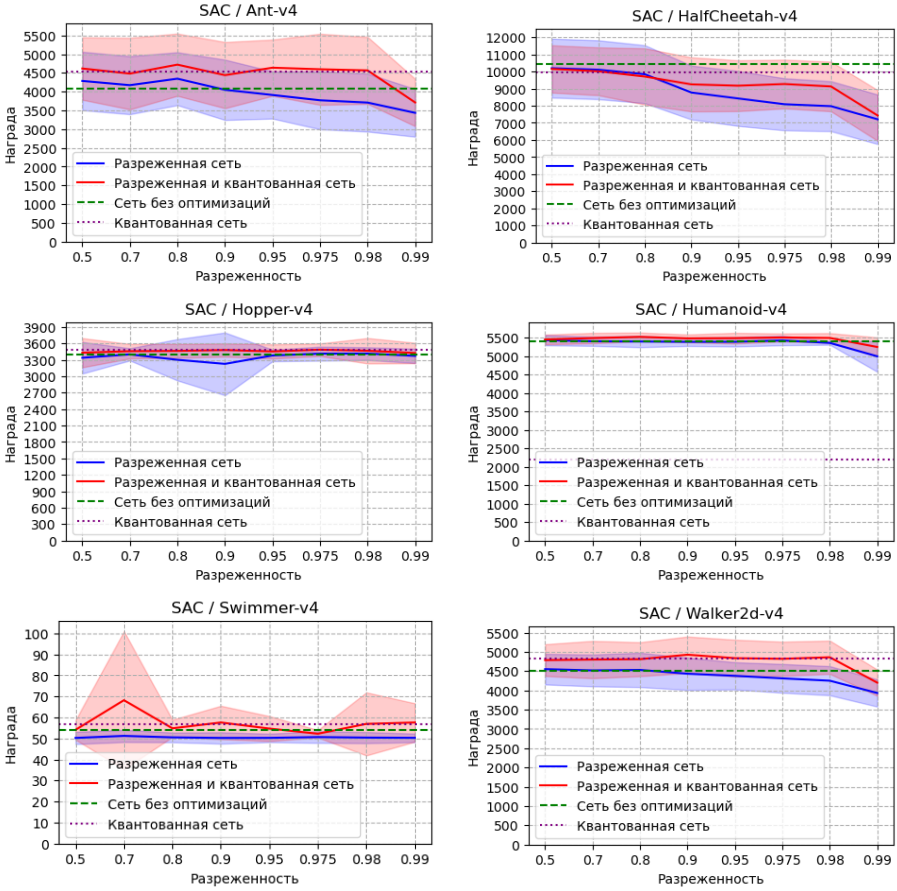

Рис. 5 — Результаты оптимизации для алгоритма SAC применительно к окружениям MuJoCo. Оси x на графиках отображают степень разреженности нейронной сети; оси y обозначают производительность — награду, полученную агентом (чем больше тем лучше). Синяя линия показывает производительность обрезанной сети, красная линия показывает производительность обрезанной и квантованной сети (результат автора) [A.2]. Пунктирная фиолетовая линия показывает производительность только квантованной сети. Зелёная пунктирная линия показывает производительность сети по умолчанию (без оптимизаций).

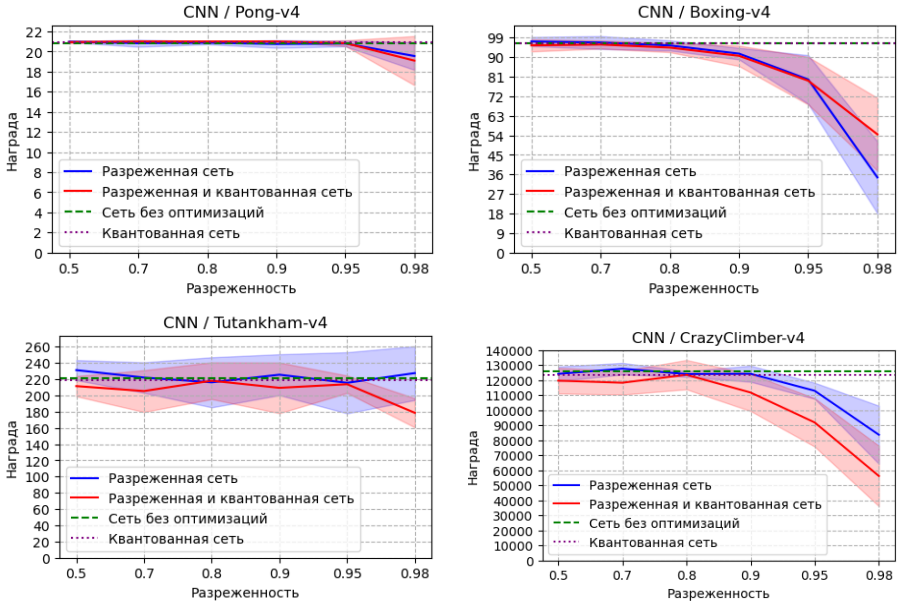

Рис. 6 — Результаты оптимизации для алгоритма DQN и сверточной нейронной сети (CNN), применительно к Atari средам. Оси x на графиках отображают степень разреженности нейронной сети; оси y обозначают производительность — награду, полученную агентом (чем больше тем лучше). Синяя линия показывает производительность обрезанной сети, а красная линия показывает производительность обрезанной и квантованной сети (результат автора) [A.2]. Пурпурная пунктирная линия показывает производительность только квантованной сети. Зеленая пунктирная линия показывает производительность сети по умолчанию (без оптимизаций).

В качестве метода обучения с подкреплением был применен алгоритм DQN, в качестве архитектуры обучаемых сетей - классическая сверточная сеть из оригинальной работы DQN. В качестве модельных среды для обучения были использованы окружения из семейства Atari games: Freeway, Enduro, Krull, Robotank, Breakout, SpaceInvaders. Результаты представлены на рисунке 8.

Показатели вознаграждения (синяя линия на графике) согласуются с результатами, полученными в работах по применению алгоритма LTH в задачах обучения с подкреплением. Оранжевая линия демонстрирует вознаграждение для нейронных сетей с дельта-нейронами (т.е. после применения дельта-алгоритма). Их производительность сопоставима с производительностью сети без дельта-нейронов (синяя линия), что свидетельствует о том, что дельта-алгоритм *не оказывает* существенного влияния на вознаграждение.

В *разделе 4.3.1* проведен анализ количества обращений в память. Усредненные доли числа обращений в память оптимизированного алгоритма от числа

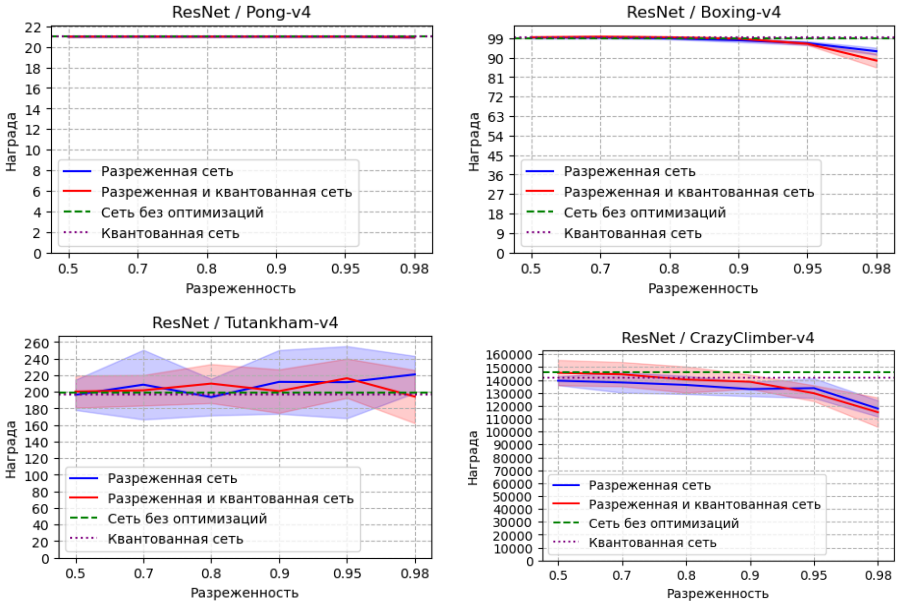

Рис. 7 — Результаты оптимизации для алгоритма DQN и сети ResNet, применительно к Atari средам. Ось x на графиках обозначает степень разреженности нейронной сети; ось y обозначает производительность - награду, полученную агентом (чем больше тем лучше). Синяя линия показывает производительность обрезанной сети, красная линия показывает производительность обрезанной и квантованной сети (результат автора) [A.2]. Пунктирная фиолетовая линия показывает производительность только квантованной сети. Зеленая пунктирная линия показывает производительность сети по умолчанию (без оптимизаций).

обращений в память алгоритма без оптимизаций отображены зеленой пунктирной линией на рисунке 8. Число обращений к памяти зависит от игры, в которую играет агент, и от уровня разреженности сети. Предложенный алгоритм оптимизации обеспечивает различные уровни структурной и временно́й разреженностей в зависимости от выбранного порога, слоя и входных данных (окружения). Дельта-алгоритм (без обрезания весов) приводит к уменьшению количества обращений к памяти от 3.7 раз для Robotank до 14.3 раз для Breakout. Одновременное использование с дельта-алгоритмом структурной разреженности еще больше уменьшает количество обращений к памяти. Например, для Freeway доля обращений к памяти уменьшается при обрезании с 0.33 до 0.026 (уменьшение в 12.6 раз). Однако для Krull она уменьшается с 0.07 лишь, до 0.033 (уменьшение в 2.2 раза). Такая вариабельность объясняется тем, что структурная разреженность весов статически влияет на количество обращений в память, в то время как дельта-алгоритм обеспечивает разные уровни временно́й разреженности в зависимости от выбранного порога, слоя, разреженности и входных данных

(окружения). Таким образом проявляется сложное внутреннее взаимодействие и взаимная интерференция обоих алгоритмов оптимизации.

В *разделе 4.3.2* проведен анализ количества значимых операций умножения (в которых оба операнда не ноль). Усредненные доли ненулевых операций умножения в оптимизированном алгоритме от числа операций умножения алгоритма без оптимизаций отображены на рисунке 8 красной пунктирной линией.

В *разделе 4.3.3* проанализирована зависимость числа значимых операций умножения и качества работы сети от среды и порога D дельта-алгоритма. Проведенные эксперименты показали, что порог равный 0.01 является наиболее подходящим.

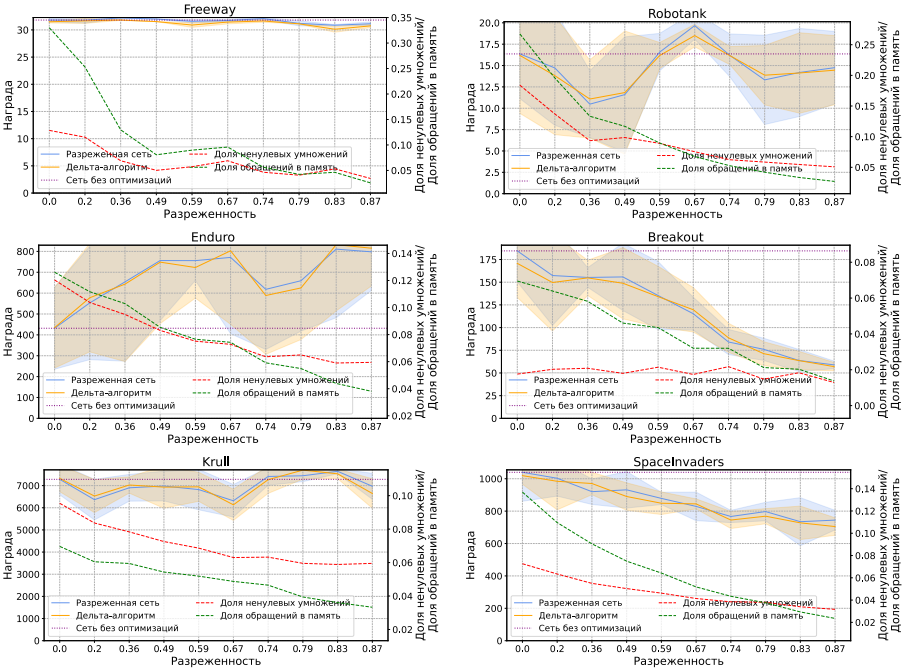

Рис. 8 — Результаты оптимизации для сред Freeway, Robotank, Enduro, Breakout, Krull and SpaceInvaders. На всех графиках ось x отображает степень разреженности нейронной сети. Левая ось y обозначает вознаграждения (качество работы), полученные агентом; правая ось y представляет долю значимых операций умножения и долю обращений к памяти, усредненную по запускам среды. Синяя линия демонстрирует производительность обрезанной сети, а оранжевая - производительность обрезанной сети с дополнительным применением дельта-алгоритма (результат автора) [A.3]. Красная пунктирная линия отображает долю значимых умножений обрезанной нейронной сети, дополненной дельта-алгоритмом от числа значимых умножений неоптимизированной сети (меньшее значение предпочтительнее, результат автора) [A.3]. Зеленая пунктирная линия показывает долю необходимых обращений к памяти обрезанной нейронной сети, дополненной дельта-алгоритмом от числа обращений неоптимизированной сети (меньшее значение предпочтительнее, результат автора) [A.3]. Фиолетовая пунктирная линия иллюстрирует качество работы нейронной сети без какой-либо оптимизации.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. На основе детального анализа и классификации вычислительных принципов работы мозга человека показано, что отсутствие ряда ключевых принципов в современных системах ИИ, построенных на основе вычислителей с фон-неймановской архитектурой, определяет их низкую энергоэффективность, масштабируемость и скорость работы. Указанные ключевые принципы легли в основу предложенных в работе методов, направленных на создание быстрых и энергоэффективных систем ИИ для инференса задач глубокого обучения с подкреплением.
2. Впервые предложен метод оптимизации инференса нейронных сетей, тренированных алгоритмами обучения с подкреплением, на основе комбинации структурной разреженности и квантования. Метод существенно уменьшает размеры нейронных сетей (на 1 – 2 порядка, вплоть до 400 раз) без потери качества работы, что позволяет размещать нейронные сети в быстрой памяти или уменьшать число обращений в память, а также уменьшает на порядок число необходимых арифметических операций для инференса.
3. Впервые предложен метод оптимизации инференса нейронных сетей, тренированных алгоритмами обучения с подкреплением, на основе комбинации временно́й и структурной разреженности. Метод уменьшает на порядок (до 25 раз) число обращений в память и число необходимых арифметических операций при инференсе нейронных сетей без потери качества работы.
4. Предложенные методы оптимизации программно реализованы и прошли апробацию на тестовых окружениях Atari и MuJoCo, являющихся стандартными бенчмарками для задач обучения с подкреплением и хорошо отражающих типовые случаи входных данных в них, подтвердив эффективность предложенных методов.

В заключение автор выражает благодарность и большую признательность научному руководителю Воеводину Владимиру Валентиновичу за помощь, обсуждение результатов и научное руководство. Автор также выражает благодарность своей матери Ивановой Ольге Петровне и своему дедушке Иванову Петру Васильевичу (1930–2025) за воспитание, образование и поддержку. Автор глубоко признателен своей жене Кузнецовой Дарье Владимировне за её неоценимую поддержку, помощь, терпение и понимание.

Публикации автора по теме диссертации

Научные статьи, опубликованные в изданиях, рекомендованных для защиты в диссертационном совете МГУ имени М.В. Ломоносова по специальности и отрасли наук:

A.1. Neuromorphic artificial intelligence systems / D. Ivanov, A. Chezhegov, M. Kiselev, A. Grunin, D. Larionov // Frontiers in Neuroscience. — 2022. — Vol. 16. — P. 959626. — EDN: HZGYDI — (WoS Q2, Импакт-фактор 3.2 (JIF)) [1.25/0.75]
Автором была самостоятельная разработана концепция статьи и предложена классификация, описанная в статье.

A.2. Neural network compression for reinforcement learning tasks / D. A. Ivanov, D. A. Larionov, O. V. Maslennikov, V. V. Voevodin // Scientific Reports. — 2025. — Vol. 15, no. 1. — P. 9718. — EDN: PGLCHZ — (WoS Q1, Импакт-фактор 3.8 (JIF)) [0.68/0.5]
Автором был самостоятельно разработан метод оптимизации нейронных сетей с помощью структурной разреженности и квантования, а также проведено экспериментальное исследование предложенного метода.

A.3. Deep reinforcement learning with significant multiplications inference / D. A. Ivanov, D. A. Larionov, M. V. Kiselev, D. V. Dylov // Scientific Reports. — 2023. — Vol. 13, no. 1. — P. 20865. — EDN: PAYIZL — (WoS Q1, Импакт-фактор 3.8 (JIF)) [0.625/0.45]
Автором был самостоятельно разработан метод оптимизации нейронных сетей с помощью структурной и временной разреженностей, и проведено экспериментальное исследование предложенного метода.

Иванов Дмитрий Александрович

Нейроморфные методы оптимизации систем искусственного интеллекта для задач обучения с подкреплением

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____