На правах рукописи

Козлов Даниил Александрович

ИНТЕГРАЦИЯ ИЕРАРХИЧЕСКИХ АНСАМБЛЕЙ И ТРАНСФОРМЕРНЫХ АРХИТЕКТУР В АЛГОРИТМЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

1.2.1. Искусственный интеллект и машинное обучение

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Самара – 2024

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Самарский национальный исследовательский университет имени академика С. П. Королева» на кафедре геоинформатики и информационной безопасности.

Научный руководитель:
Мясников Владислав Валерьевич, доктор физико-математических наук, профессор.

Официальные оппоненты:

Арлазаров Владимир Викторович, доктор технических наук, Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», заведующий отделом;

Визильтер Юрий Валентинович, доктор физико-математических наук, Федеральное автономное учреждение «Государственный научно-исследовательский институт авиационных систем», начальник подразделения.

Ведущая организация:
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный технический университет», г. Ульяновск.

Защита состоится 24 декабря 2024 г. в 12:00 часов на заседании диссертационного совета 24.2.379.08, созданного на базе федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С. П. Королева», по адресу: 443086, г. Самара, Московское шоссе, 34.

С диссертацией можно ознакомиться в библиотеке федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С. П. Королева» и на официальном сайте по адресу: https://ssau.ru/storage/pages/6528/file_66fc6136932cb4.98509564.pdf

Автореферат разослан «__» _____ 2024 г.

Учёный секретарь
диссертационного совета,
д.ф.-м.н., доцент Дорошин А.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Проблема управления роботами в сложных условиях становится все более актуальной в контексте быстрого развития технологий и увеличения сложности технических систем. Такие подходы к управлению, как обучение с подкреплением, предлагают возможности для значительного повышения эффективности и адаптивности роботов. Эти методы позволяют роботам самостоятельно изучать и оптимизировать свои стратегии поведения в реальном времени, что особенно важно для действий в условиях, где детальное предварительное моделирование среды невозможно или неэффективно.

Применение обучения с подкреплением в робототехнике привело к множеству значимых достижений в улучшении автономности, производительности и адаптивности роботов. Одним из наиболее заметных примеров успешного применения обучения с подкреплением является разработка автономных транспортных средств. Эти системы используют обучение с подкреплением для оптимизации стратегий вождения, позволяя автомобилям самостоятельно принимать решения в сложных дорожных условиях и оптимизировать потоки транспортных средств в городских сетях.

В области промышленного производства обучение с подкреплением применяется для управления роботизированными руками, которые выполняют задачи сборки и манипуляции с объектами. Эти роботы обучаются адаптироваться к изменениям в объектах или их расположении, что позволяет автоматизировать процессы, требующие высокой точности и гибкости.

Роботы, используемые в задачах поиска и спасения, должны работать в условиях высокой неопределенности и динамичных изменений среды. Обучение с подкреплением позволяет этим роботам обучаться на основе взаимодействия с реальной средой, улучшая свои способности к самостоятельному принятию решений в критических ситуациях.

Алгоритмы и методы обучения с подкреплением также могут быть использованы для управления беспилотными летательными аппаратами, шагающими роботами, манипуляционными роботами. Отличительной особенностью алгоритмов обучения с подкреплением является тот факт, что для них не требуется точного моделирования среды, в которой они будут действовать. Вместо этого агент сам изучит среду и обучится принимать оптимальные решения.

О высоком потенциале современных методов обучения с подкреплением свидетельствуют работы таких авторов как *Р.С.Саттон, Д.Сильвер, А.И.Панов, Л. Чен* и другие.

Для объективного анализа актуальности выбранной темы была произведена выборка статей, содержащих заданные ключевые слова: «Reinforcement Learning», «Reinforcement Learning» одновременно с «Robot» и, «Reinforcement Learning» одновременно с «Transformer». Выборка производилась в электронном архиве с открытым доступом для научных статей и рукописей arXiv. Результат анализа представлен на рисунке 1. Нелинейный рост публикаций в данной области подтверждает актуальность выбранной темы.

Рисунок 1 – Распределение статей с заданными ключевыми словами по годам

Цели и задачи исследования

Целью диссертационного исследования является разработка и исследование методов, алгоритмов и способов повышения качественных показателей алгоритмов обучения с подкреплением в рамках класса задач управления роботами, способными к перемещению в трехмерных средах. Для достижения указанной цели в диссертации решались следующие задачи:

1. Анализ лучших современных алгоритмов обучения с подкреплением с целью выявления их ограничений и особенностей использования в рассматриваемом классе задач.
2. Разработка модели интеграции алгоритмов обучения с подкреплением с кодировщиком трансформера, разработка и исследование нового алгоритма обучения с подкреплением, основанного на этой модели интеграции.
3. Разработка метода иерархического ансамблирования алгоритмов обучения с подкреплением, разработка и исследование нового алгоритма обучения с подкреплением, основанного на этом методе.

Методология и методы исследования

При проведении работы использовались методы машинного обучения, машинного обучения с подкреплением, разработки программного обеспечения.

Научная новизна

1. Разработана методика оценки влияния состава набора наблюдений окружающей среды на качество решений, принимаемых агентом, позволяющая упорядочить наблюдения по их полезности.
2. Предложена модель интеграции алгоритмов обучения с подкреплением и кодировщика трансформера для кодирования входных последовательностей состояний с целью повышения качества решения задачи.
3. Разработан алгоритм, интегрирующий кодировщик трансформера и алгоритм обучения с подкреплением Soft Actor-Critic.
4. Предложен метод иерархического ансамблирования алгоритмов обучения с подкреплением, который позволяет объединить несколько алгоритмов в иерархическую структуру для повышения качества обучения без дополнительных обращений к среде.
5. Разработан алгоритм обучения с подкреплением на основе предложенного метода иерархического ансамблирования с использованием алгоритма DQN в качестве управляющего и алгоритмов SAC и REDQ в качестве управляемых.

Практическая значимость

Разработанные решения улучшают качественные показатели обучения агентов, что позволяет их использовать для создания нового поколения роботов. Это расширяет области применения робототехнических систем и повышает их эксплуатационную надежность и эффективность.

На защиту выносятся

1. Метод иерархической интеграции ансамбля алгоритмов обучения с подкреплением, позволяющий объединить несколько алгоритмов в иерархическую структуру для повышения качества обучения без дополнительных обращений к среде. Доказана возможность повышения эффективности обучения за счет использования данной структуры по сравнению с отдельным использованием каждого алгоритма ансамбля.
2. Алгоритм обучения с подкреплением на основе предложенного метода иерархической интеграции, в котором алгоритм DQN используется в качестве управляющего, а алгоритмы SAC и REDQ — в качестве управляемых. Данный подход улучшает показатели качества обучения за счет распределения ролей среди алгоритмов.

3. Модель интеграции алгоритмов обучения с подкреплением и кодировщика трансформера, предназначенная для кодирования входных последовательностей состояний. Предложенная модель улучшает качество решений задач за счет более эффективного представления информации об окружающей среде.
4. Алгоритм обучения с подкреплением, интегрирующий архитектуру трансформера в алгоритм Soft Actor-Critic для кодирования входных последовательностей состояний. Разработанный алгоритм демонстрирует улучшение результатов по сравнению с оригинальным алгоритмом Soft Actor-Critic.

Соответствие специальности

Диссертация соответствует паспорту научной специальности 1.2.1 – «Искусственный интеллект и машинное обучение» и охватывает следующие области исследования, входящие в эту специальность:

- Формализация и постановка задач управления и (поддержки) принятия решений на основе систем искусственного интеллекта и машинного обучения. Разработка систем управления с использованием систем искусственного интеллекта и методов машинного обучения в том числе – управления роботами, автомобилями, БПЛА и т.п.
- Исследования в области многослойных алгоритмических конструкций, в том числе – многослойных нейросетей.

Степень достоверности и апробация результатов

Достоверность научных результатов обеспечена применением методов статистического анализа, сравнением предложенных алгоритмов с существующими решениями и их экспериментальной проверкой на задачах управления роботами в трехмерных средах. Основные результаты научно-квалификационной работы были представлены на четырёх научных конференциях:

1. Международной конференции «Информационные технологии и нанотехнологии» (ИТНТ, Самара, Россия) - 2021 год;
2. Международной конференции «Информационные технологии и нанотехнологии» (ИТНТ, Самара, Россия) - 2022 год;
3. Международной конференции «Информационные технологии и нанотехнологии» (ИТНТ, Самара, Россия) - 2023 год;
4. Международной конференции «Информационные технологии и нанотехнологии» (ИТНТ, Самара, Россия) - 2024 год;

По теме диссертации опубликовано десять работ. Из них одна работа в изданиях, рекомендуемых ВАК, четыре работы опубликованы в изданиях, индексируемых в БД Scopus. Шесть работ выполнены без соавторов. Получено одно свидетельство Роспатента о регистрации программы для ЭВМ.

Результаты диссертационной работы:

1. Внедрены в рамках НИР в ООО «Давтех» в рамках договора №55/08/2023 от 01.08.2023.
2. Использованы в учебном процессе в ФГАОУ ВО «Самарский национальный исследовательский университет имени академика С. П. Королева» в курсе лекций по дисциплине «Машинное обучение и распознавание образов».
3. Использованы в рамках договора 7/2021 от 08.11.2021 (2021–2023) между АО «Самара-Информспутник» и ФГУП «ГосНИИПП».
4. Использованы в ФГАОУ ВО «Самарский национальный исследовательский университет имени академика С. П. Королева» в рамках гранта РНФ №. 21-11-00321, «Методы и алгоритмы совместного и координированного управления сигналами светофоров и подключенными автономными транспортными средствами в транспортной сети».

Структура диссертации

Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы из 94 наименований. Работа содержит 98 страниц текста, включая 4 таблицы и 32 рисунка.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

В первом разделе диссертации изложены ключевые концепции, используемые в области обучения с подкреплением. Вводятся понятия стратегии, функции ценности и марковского процесса. Также приводится обзор существующих методов обучения с подкреплением. Методы обучения с подкреплением выбраны для решения задачи приобретения навыков передвижения в пространстве как наиболее перспективные и адаптивные. Эти методы способны решать некоторую задачу, получая только оценку своих действий в виде значения функции награды (рисунок 2). Таким образом, при должном развитии методов обучения с подкреплением возможно создание автономных систем, адаптирующихся к изменяющимся условиям для достижения целей. Например, робот, использующий для управления своим движением методы обучения с подкреплением, способен адаптироваться даже в случае частичных повреждений и продолжить выполнение поставленной задачи.

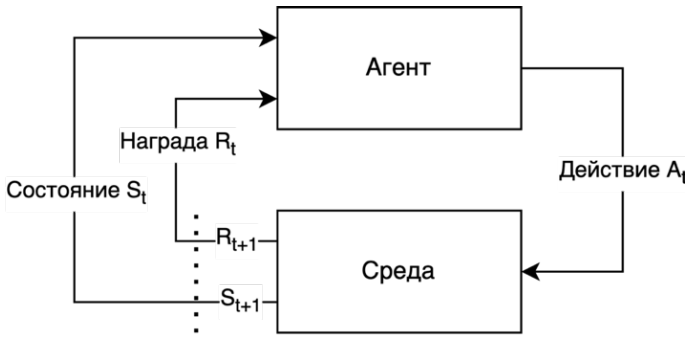

Рисунок 2 – Основная концепция обучения с подкреплением

Приведена классификация существующих методов обучения с подкреплением на модельные и безмодельные методы, а также классификация на основанные на значении и основанные на стратегии методы. Для решения поставленной задачи далее в диссертации будут рассматриваться только безмодельные методы, поскольку они обладают меньшей вычислительной сложностью и не требуют обучения модели окружающего мира.

Приводится краткое описание принципов работы методов Q-learning, Deep Q-Network (DQN), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), Policy Gradients, Actor-Critic, Soft Actor-Critic, Randomized Ensembled Double Q-Learning (REDQ). Для приведенных алгоритмов проанализированы потенциальные преимущества и недостатки. Приводятся примеры успешного применения алгоритмов обучения с подкреплением в реальных задачах, а также информация о проектировании сред и формулировке функции награды. В таблице 1 отображена краткая информация об актуальных алгоритмах обучения с подкреплением.

Таблица 1 – Современные актуальные методы

Метод/Алгоритм	Тип RL	Основная идея/характеристика	Применение	Год
Deep Q-Network (DQN)	Оффлайн Q-обучение	Использует глубокие нейросети для аппроксимации Q-функции ценности.	Игровые среды, такие как Atari	2015
Trust Region Policy Optimization (TRPO)	Градиент стратегии	Ограничивает шаги обновления стратегии с помощью метода доверительных регионов	Робототехника, системы с непрерывными действиями	2015
Deep Deterministic Policy Gradient (DDPG)	Оффлайн, Градиент стратегии	Адаптирует Actor-Critic для задач с непрерывными пространствами действий	Управление роботами, автомобильные симуляции	2016

Метод/Алгоритм	Тип RL	Основная идея/характеристика	Применение	Год
AlphaZero	Моделирование динамики среды	Комбинирует глубокое обучение и алгоритм Монте-Карло для поиска и обучения	Настольные игры, шахматы, го	2017
Proximal Policy Optimization (PPO)	Градиент стратегии	Ограничивает обновления стратегии с помощью KL-дивергенции для повышения стабильности обучения	Робототехника, управление сложными системами	2017
Soft Actor-Critic (SAC)	Оффлайн, энтропийная регуляризация	Максимизирует энтропию стратегии для улучшения стабильности и исследовательской способности	Робототехника, симуляции с непрерывными действиями	2018
MuZero	Моделирование динамики среды	Изучает оптимальную стратегию без знания динамики среды, комбинирует обучение модели и планирование	Игровые среды (Go, Chess, Atari)	2019
Offline Reinforcement Learning (CQL)	Оффлайн	Цель — улучшение эффективности offline RL с помощью консервативного Q-learning	Управление роботами, рекомендательные системы	2020
Behavioral Cloning Transformer (BC-Transformer)	Имитационное обучение	Применяет архитектуру трансформера для имитационного обучения	Игровые среды, задачи с большим объемом данных	2021
Randomized Ensembled Double Q-learning (REDQ)	Off-policy Q-обучение	Использует несколько Q-функций для улучшения стабильности и сходимости	Робототехника, задачи с высокой стохастичностью	2021
Decision Transformer	Моделирование последователь-ностей	Применяет архитектуру трансформера для обучения с подкреплением, формулируя задачу как последовательное моделирование	Игровые среды, управление роботами	2021

Во втором разделе диссертации представлены результаты экспериментальных исследований существующих методов обучения с подкреплением, проводимых для определения сильных и слабых сторон этих методов при решении задачи приобретения навыков передвижения в трехмерном пространстве. Также проведены исследования для определения влияния тех или иных компонент наблюдений среды на эффективность решения задачи.

В первой части второго раздела диссертации приведено описание программного обеспечения, используемого для проведения экспериментальных исследований методов, описанных в диссертации. Описание используемого программного обеспечения необходимо для выполнения условий повторяемости экспериментов и репрезентативности их результатов. Так, для выполнения экспериментальных исследований используется язык программирования Python, совместно с библиотекой для глубокого обучения PyTorch и библиотекой для глубокого обучения с подкреплением TorchRL. В качестве экспериментальных сред рассматриваются среды из пакета ML-Agents для физического движка Unity (рисунок 3), среды из пакета dm-control (рисунок 4) и gymnasium (рисунок 5), использующие физический движок MuJoCo. Такой выбор сред обоснован тем, что среды, основанные на OpenAI gym и MuJoCo, являются общепринятым в области обучения с подкреплением стандартом для тестирования и сравнения алгоритмов. Авторы, многих работ, описывающих лучшие современные алгоритмы (PPO, DDPG, SAC, REDQ, DecisionTransformer и другие) производят тестирование именно в этих средах. Эти среды помимо набора стандартных, ставших классическими, задач предлагают также стандарты взаимодействия сред и алгоритмов.

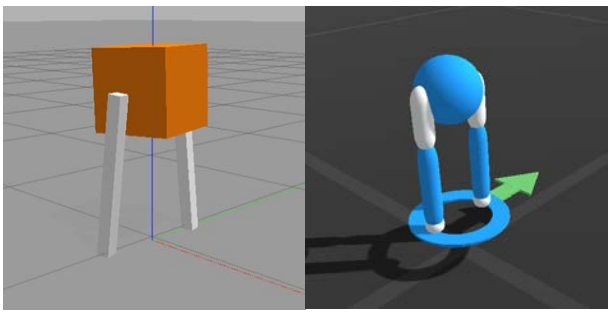

Рисунок 3 – Две реализации среды SimplestBipedal. Первая реализована в симуляторе Gazebo, вторая - в Unity

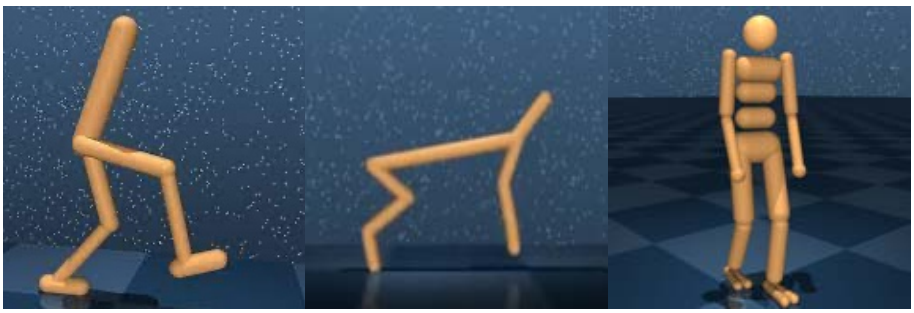

Рисунок 4 – Среды из пакета dm-control. Walker, Cheetah, Humanoid

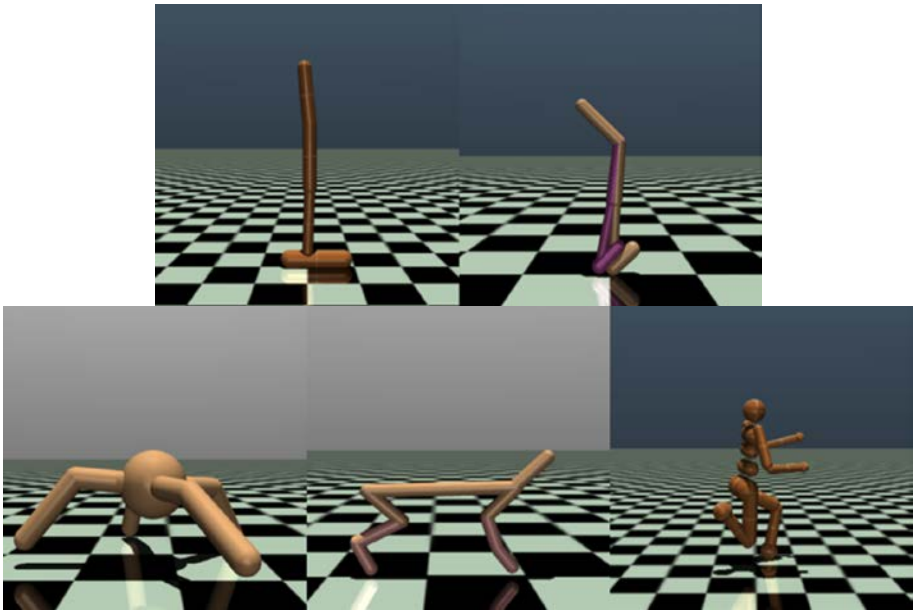

Рисунок 5 – Среды из пакета gymnasium. Hopper, Walker, Ant, Half-Cheetah, Humanoid

Во второй части второго раздела диссертации приведено сравнение реализаций метода DQN в среде симулятора Gazebo. В ходе исследования были сделаны следующие выводы:

- Алгоритмы семейства DQN плохо подходят для сложных задач с непрерывными пространствами действий и большими размерностями.
- Использование симулятора Gazebo позволяет избежать прототипирования реального робота и длительного процесса обучения его в реальном мире, однако этот симулятор не реализует изменения масштаба времени симуляции и параллельного запуска нескольких эпизодов для ускорения процесса обучения. В результате эксперименты занимают куда больше времени, чем, например, в средах Unity, реализованных при помощи пакета ML-Agents.

В третьей части второго раздела диссертации приведено сравнение методов SAC, PPO, REDQ в задаче приобретения навыков передвижения в трехмерном пространстве в средах пакета gymnasium. Это сравнение необходимо для получения четкой отправной точки для других экспериментов. Проблема состоит в том, что многие среды с течением времени претерпевают изменения и обновления, а авторы алгоритмов сравнивают свои алгоритмы только с некоторыми из актуальных.

В качестве сред выбраны среды Hopper, Walker, Ant, Half-Cheetah, Humanoid. Все перечисленные среды подобраны по принципу схожести поставленной задачи. Симулированная среда Ant моделирует трёхмерного робота-муравья с телом и четырьмя ногами, каждая из которых состоит из двух сегментов. Задача включает в себя координацию движений всех четырёх ног для перемещения вперёд, применяя моменты к восьми сочленениям, которые соединяют сегменты ног с туловищем. Среда изображена на рисунке 5 снизу слева. Half-Cheetah представляет собой двумерного робота из девяти сегментов и восьми сочленений. Задача заключается в приложении моментов к сочленениям для максимизации скорости движения вперёд, при этом за движение вперёд начисляются положительные награды, а за движение назад — отрицательные. Среда изображена на рисунке 5 снизу посередине. Среда Humanoid представляет трёхмерного двуногого робота, моделирующего человека. Он состоит из туловища с двумя ногами и руками, где каждая нога разделена на три сегмента, а руки — на два. Основная задача агента заключается в движении вперёд как можно быстрее без падения. Среда изображена на рисунке 5 снизу справа. Hopper — это одноногий двумерный робот, состоящий из четырёх основных частей: торса, бедра, голени и ступни. Задача агента — выполнение прыжков вперёд, применяя силу к трем сочленениям. Среда изображена на рисунке 5 сверху слева. Walker расширяет среду Hopper за счёт добавления второго набора ног, позволяя роботу передвигаться вперёд. Задача заключается в координации движений всех семи частей тела, чтобы продвигаться вперёд, применяя моменты к шести сочленениям. Среда изображена на рисунке 5 сверху справа. Таким образом, все агенты поставлены перед задачей взаимодействия с физическим миром в трехмерном пространстве.

Оценка производительности агентов проводилась по критериям скорости обучения, которая проявляется в виде выборочной эффективности. Выборочная эффективность (англ. sample efficiency) в контексте обучения с подкреплением — это показатель, отражающий способность алгоритма достигать высокого уровня производительности с использованием ограниченного числа обучающих примеров или взаимодействий с окружающей средой.

Сравнительный анализ показал, что SAC демонстрирует наилучшую производительность во многих средах, а особенно в средах, где требуется сложная координация действий. В стабильных условиях PPO и REDQ показывают также хорошие результаты. По результатам исследования становится понятным, что выбор алгоритма может зависеть от конкретной решаемой задачи и соответствующей ей среды, но наиболее адаптивным и универсальным является SAC.

В четвертой части второго раздела диссертации приведено исследование влияния состава набора наблюдений окружающей среды на эффективность решения задачи приобретения навыков передвижения в трехмерном пространстве. Эксперимент проведен таким образом, что двуногому агенту SimplestBipedal (рисунок 3 справа), которому необходимо обучиться передвигаться в трехмерном пространстве, передаётся различный набор наблюдений окружающей среды в каждом эксперименте. Конкретные наборы можно увидеть в таблице 2. График зависимости значения награды агента от шага обучения приведен на рисунке 6 для всех экспериментов из таблицы 2. Также необходимо отметить, что для алгоритмов обучения с подкреплением характерны высокая нестабильность и непредсказуемость процесса обучения, что является неизбежной особенностью данной методологии. Это связано с тем, что агент взаимодействует с динамической средой, где малые изменения в стратегии или стратегии агента могут значительно влиять на получаемые награды, что приводит к отклонениям в процессе оптимизации.

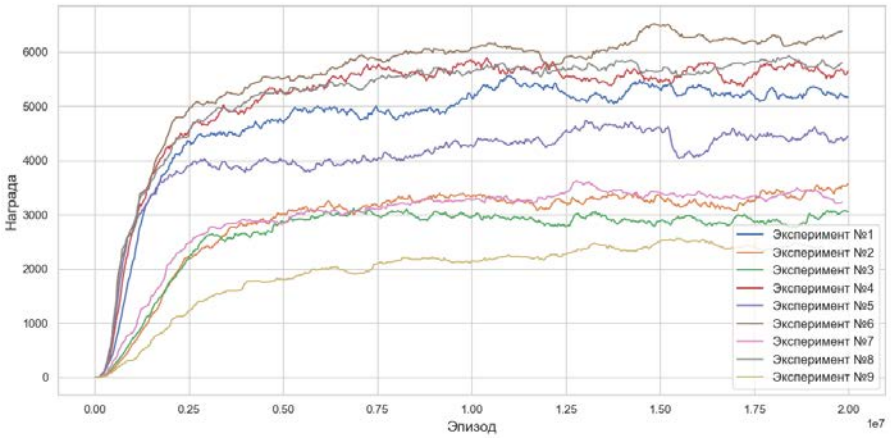

Рисунок 6 – Значение суммарной награды агента в зависимости от шага обучения для каждого эксперимента

Таблица 2 – Конфигурация проведенных экспериментов

Информация, передаваемая агенту	Номер эксперимента								
	0	1	2	3	4	5	6	7	8
Перемещение конечностей (глобальные координаты)	+	+	+	+	+	+		+	+
Перемещение конечностей (локальные координаты)			+	+	+	+	+		
Поворот конечностей (глобальные координаты)	+	+	+	+	+		+	+	+
Поворот конечностей (локальные координаты)	+		+	+	+		+	+	
Скорость конечностей (глобальные координаты)	+			+	+	+	+		
Скорость конечностей (локальные координаты)				+	+			+	
Сила, прилагаемая к суставам					+				+

Исходя из анализа результатов экспериментальных исследований можно заключить, что увеличение объема информации, предоставляемой агенту от среды, не приводит к увеличению качества решения им поставленной задачи. Пятый эксперимент, несмотря на минимальный набор информации, показывает себя наилучшим образом. Результаты исследования подтверждают необходимость предварительного изучения влияния состава наблюдений окружающей среды на решаемую задачу.

На основе результатов исследования предложена *методика оценки влияния состава набора наблюдений окружающей среды на качество решений*, принимаемых агентом.

Рассмотрим множество N доступных признаков состояния среды, где каждый признак представляет собой компонент наблюдения, доступного агенту при принятии решений. Пусть проведено J экспериментов, и каждый эксперимент $e_j (j \in J)$ характеризуется итоговой (суммарной) наградой R_j, полученной агентом, и набором использованных признаков $N_j \subseteq N$. Участие признака $n \in N$ в эксперименте e_j будем обозначать как $n \in e_j$, то есть признак n был использован в эксперименте e_j.

Для каждого признака n определим множество экспериментов, в которых этот признак участвовал, как:

$$E_n = \{e_j | j \in J \text{ и } n \in N_j\}.$$

Для оценки полезности признака n, вычислим его вес w_n, который отражает вклад признака в итоговую награду. Вес рассчитывается по следующей формуле:

$$w_n = \frac{1}{|E_n|} \sum_{e_j \in E_n} \frac{R_{e_j}}{|N_{e_j}|}$$

где R_{e_j} — суммарная награда, полученная в эксперименте e_j, а $|N_{e_j}|$ — количество признаков, использованных в эксперименте e_j.

Таким образом, полезность каждого признака оценивается как среднее значение вкладов признака в итоговую награду во всех экспериментах, в которых он использовался, нормированное на количество признаков в каждом эксперименте. Высокое значение w_n указывает на значительный вклад признака n в успешность обучения агента.

В третьем разделе диссертации предложена модель интеграции алгоритмов обучения с подкреплением и кодировщика трансформера, представляющая собой концептуальный подход к обработке последовательностей состояний. Основная идея модели заключается в том, что вместо использования марковского предположения, как в классических алгоритмах обучения с подкреплением, где текущее состояние полностью определяет будущее поведение системы, предлагается учитывать последовательность предыдущих состояний. Для этого используется кодировщик трансформера, который преобразует последовательности состояний в более репрезентативные латентные представления, содержащие информацию о динамике среды и истории взаимодействий агента.

После того как была разработана указанная модель, был предложен алгоритм обучения с подкреплением, реализующий эту модель на практике. Алгоритм использует архитектуру трансформера для кодирования последовательностей состояний перед их обработкой алгоритмом Soft Actor-Critic (SAC). Веса кодировщика трансформера обновляются в процессе сквозного обучения всего алгоритма. Для обеспечения обучения разработанного алгоритма и его работы в режиме принятия решений был модифицирован стандартный метод воспроизведения опыта. А именно, добавлена возможность выбирать не по одному эпизоду опыта из буфера, а сразу связанные цепочки эпизодов. Блок-схема предложенного алгоритма приведена на рисунке 7.

В реализованном алгоритме в качестве основы для кодирования используется кодировщик трансформера, состоящий из двух слоев и трех «голов» внимания. Рассмотрим пример применения этого кодировщика в среде Humanoid.

В этой среде входное состояние $s \in \mathbb{R}^{376}$ расширяется до размера, кратного числу голов внимания $n_{head} = 3$, путем дополнения нулями. Таким образом, размерность векторного пространства модели d_{model}устанавливается равной 378:

- Размерность головы внимания d_{head} вычисляется как $\frac{d_{model}}{n_{head}} = 126$.

Ключевые параметры трансформера включают следующие матрицы весов:

$$W_K, W_Q, W_V \in \mathbb{R}^{378 \times 126}$$

Далее применяется полносвязная сеть для проекций:

- Первая проекция: преобразование $\mathbb{R}^{378} \to \mathbb{R}^{2048}$.
- Вторая проекция: преобразование $\mathbb{R}^{2048} \to \mathbb{R}^{378}$.

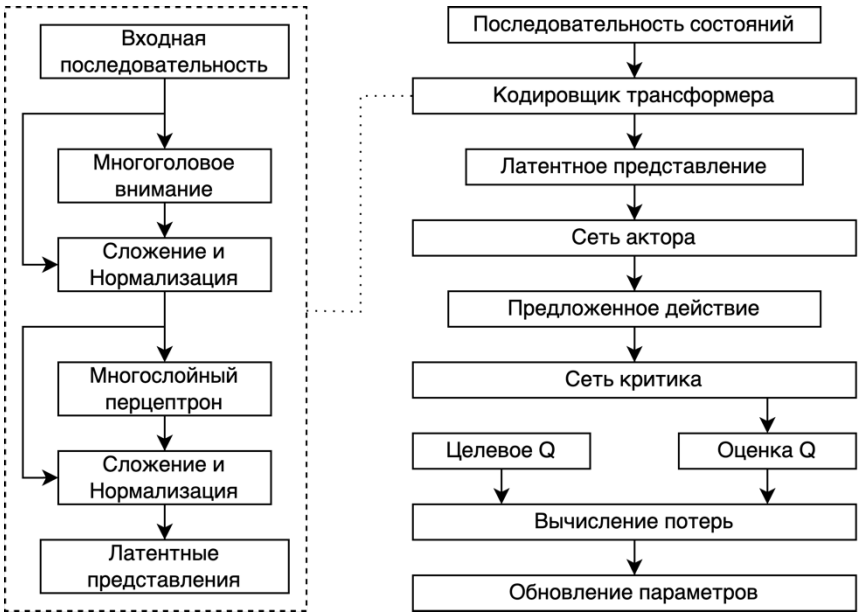

Рисунок 7 – Блок схема предложенного алгоритма

Предложенное решение повышает выборочную эффективность, что ускоряет адаптацию агента к новой среде. Исследование было проведено в средах с использованием физического движка MuJoCo, результаты экспериментов приведены на рисунках 8–12. Видно, что в некоторых средах предложенный алгоритм не приводит к улучшению выборочной эффективности. Однако в большинстве сред виден определенный прирост значения награды агента.

Для оценки изменений в качестве процесса обучения введём величину, характеризующую изменение суммарной награды, полученной алгоритмом A, по сравнению с алгоритмом B:

$$Improvement\ (\%) = \frac{\sum_{i=1}^{10} R_{A,\,argsort(R_A)_i} - \sum_{i=1}^{10} R_{B,\,argsort(R_B)_i}}{\sum_{i=1}^{10} R_{B,\,argsort(R_B)_i}} \times 100,$$

где $R_A = \{R_{A,1}, R_{A,2},\ \ldots R_{A,N}\}$ – массив наград для метода A,

$R_B = \{R_{B,1}, R_{B,2},\ \ldots R_{B,N}\}$ – массив наград для метода B,

$argsort(x)$ – функция, которая возвращает индексы элементов вектора, отсортированных по убыванию, таким образом $\sum_{i=1}^{10} R_{A,\,argsort(R_A)_i}$ – сумма значений 10 наибольших наград алгоритма A.

Таким образом, экспериментальные исследования подтверждают, что использование кодировщика трансформера повышает эффективность обработки последовательностей переходов агента. Разработанный алгоритм обеспечивает улучшение среднего суммарного значения награды на 18,5%. Частота случаев, когда результаты обучения по сравнению с оригинальным алгоритмом SAC улучшаются или остаются на прежнем уровне, составляет 80%. Значения преимущества разработанного алгоритма указаны в таблице 3.

Таблица 3 – Изменение суммарной награды при использовании разработанного алгоритма

Среда	Разница
BipedalWalker	-5,34%
Walker	+103,86%
Ant	+45,07%
Hopper	+38,90%
Humanoid	-90,04%

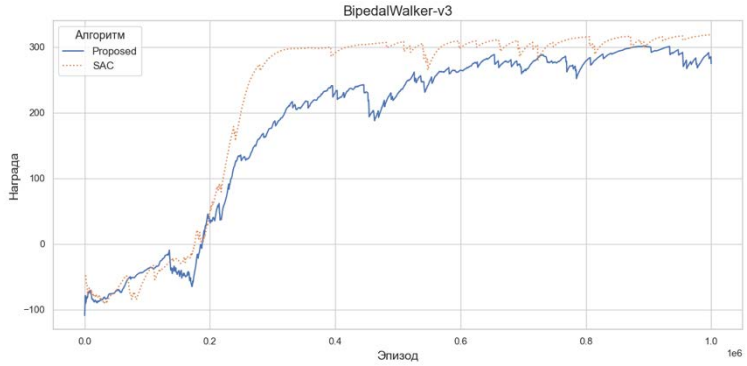

Рисунок 8 – График зависимости значения награды от шага обучения для сред BipedalWalker-v3

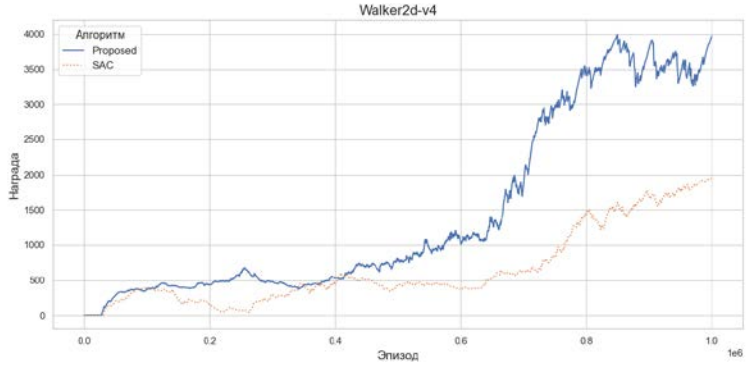

Рисунок 9 – График зависимости значения награды от шага обучения для сред Walker2d-v4

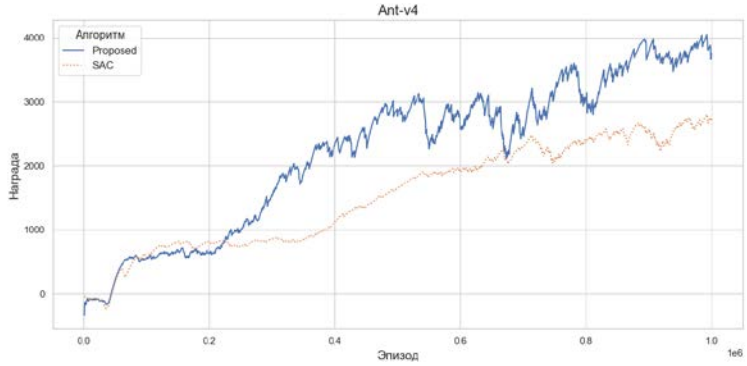

Рисунок 10 – График зависимости значения награды от шага обучения для сред Ant-v4

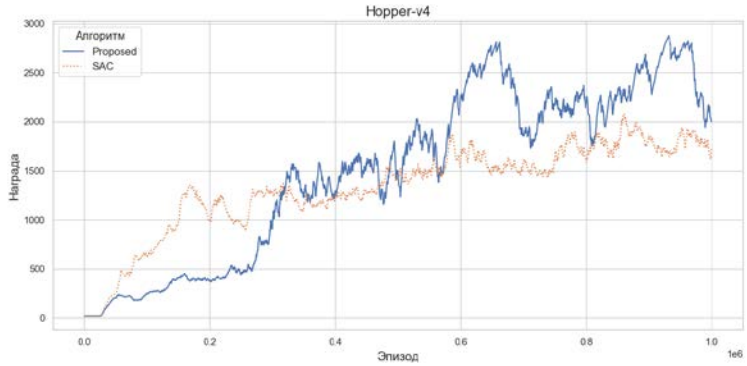

Рисунок 11 – График зависимости значения награды от шага обучения для сред Hopper-v4

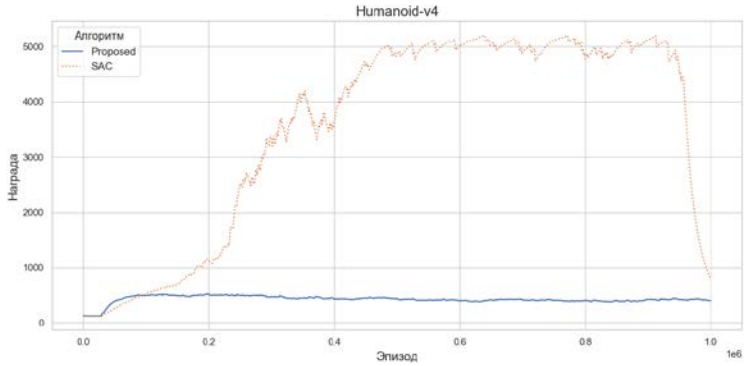

Рисунок 12 – График зависимости значения награды от шага обучения для сред Humanoid-v4

В четвертом разделе диссертации предложен ансамблевый метод обучения с подкреплением на основе иерархичности. Блок-схема предложенного метода приведена на рисунке 13 слева. Его идея заключается в создании иерархической структуры, в которой есть управляющий алгоритм и управляемые, при этом все они обучаются по принципу обучения с подкреплением. При поступлении очередного наблюдения от среды вычисляется значение функции потерь каждого алгоритма в ансамбле, и каждым алгоритмом предсказывается некое действие. Управляемые алгоритмы возвращают действия, а управляющий алгоритм выбирает алгоритм, решение которого и будет выполнять агент. При этом независимо от того, какой из алгоритмов оказывается выбран управляющим, шаг оптимизации параметров будет выполнен для всех алгоритмов ансамбля. Количество взаимодействий со средой при этом не зависит от числа алгоритмов в ансамбле.

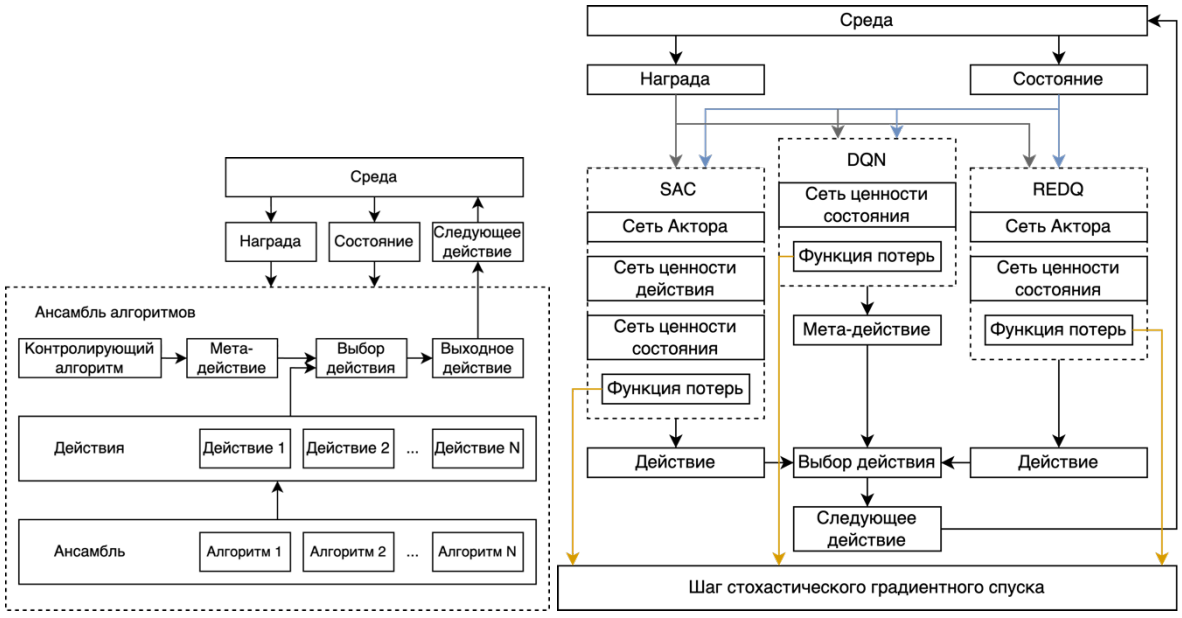

Рисунок 13 – Блок-схема предложенного метода (слева). Блок-схема реализованного алгоритма (справа)

Таким образом, можно объединить любое количество алгоритмов. В диссертационной работе была исследована версия алгоритма, где в качестве управляемых алгоритмов были алгоритмы REDQ и SAC, а управляющим выступил DQN. Такой выбор обуславливается тем, что REDQ и SAC хорошо себя проявляют в средах с непрерывными пространствами наблюдений и действий и способны к быстрой адаптации к задачам, требующим высокой координации. DQN в свою очередь хорошо себя проявляет в средах с дискретными пространствами действий. Блок-схема полученного на базе предложенного метода алгоритма представлена на рисунке 13 справа.

Ключевой особенностью рассматриваемого алгоритма является его способность к распространению опыта между различными алгоритмами в ансамбле. Эта особенность заключается в использовании механизмов обучения с подкреплением вне стратегии, что позволяет эффективно обучать все алгоритмы ансамбля, даже если они не были активированы на текущем этапе.

Каждый управляемый алгоритм в ансамбле может функционировать в двух режимах:

- Режим действия. В данном режиме алгоритм выбирается управляющим алгоритмом и выполняет свои функции в обычном режиме, непосредственно взаимодействуя с окружающей средой.
- Режим наблюдения. В этом режиме алгоритм обучается на основе наблюдаемого опыта, что позволяет ему накапливать знания и адаптироваться к изменениям в среде.

Результаты экспериментальных исследований предложенного алгоритма представлены на рисунках 15–19. На этих графиках отражена зависимость награды, получаемой агентом от шага обучения. Разработанный алгоритм обеспечивает улучшение среднего суммарного значения награды на 2,65%. При этом результаты обучения по сравнению с лучшим алгоритмом ансамбля улучшаются или остаются на прежнем уровне во всех случаях. Значения преимущества разработанного алгоритма указаны в таблице 4.

Таблица 4 – Изменение суммарной награды при использовании разработанного алгоритма

Среда	Разница
Walker walk	-0,39%
Cheetah run	+4,32%
Humanoid run pure state	+4,41%
Humanoid run	+1,40%
Humanoid stand	+3,52%

Рисунок 14 – Сравнение предложенного метода с SAC и REDQ в среде «walker walk»

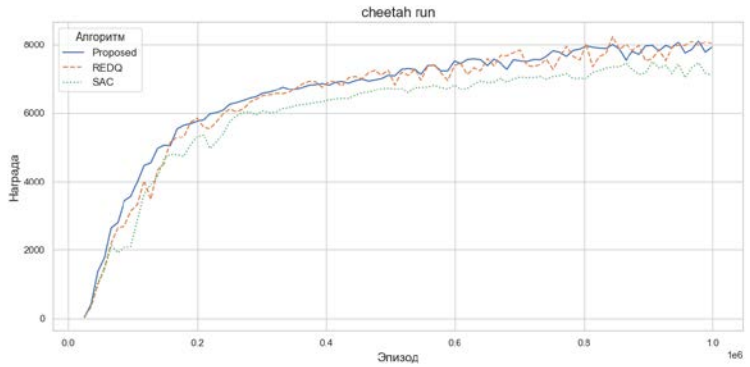

Рисунок 15 – Сравнение предложенного метода с SAC и REDQ в среде «cheetah run»

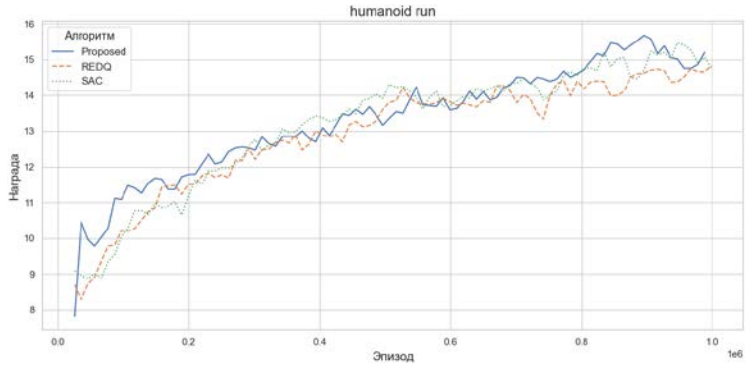

Рисунок 16 – Сравнение предложенного метода с SAC и REDQ в среде «humanoid run»

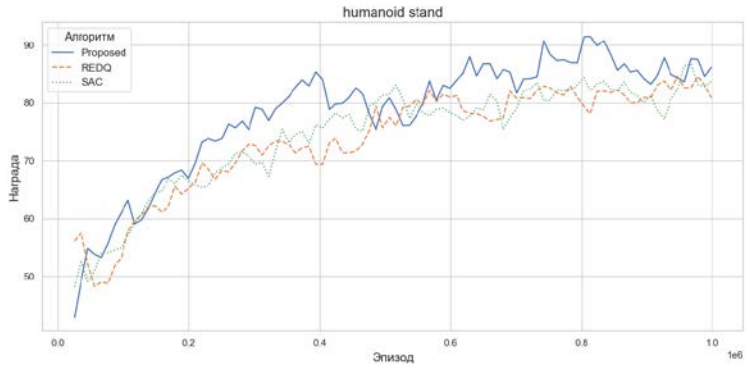

Рисунок 17 – Сравнение предложенного метода с SAC и REDQ в среде «humanoid stand»

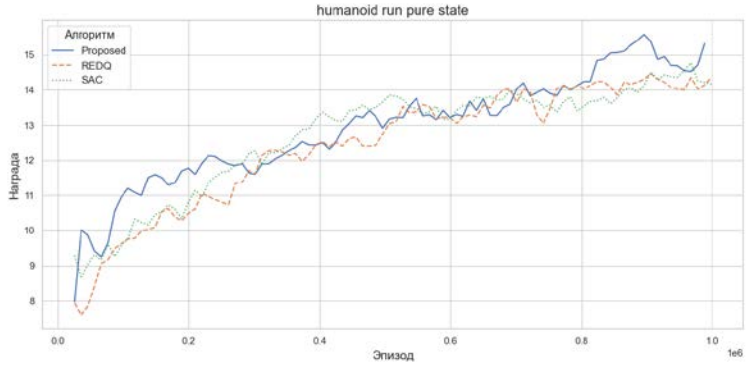

Рисунок 18 – Сравнение предложенного метода с SAC и REDQ в среде «humanoid run pure state»

ЗАКЛЮЧЕНИЕ

В данном диссертационном исследовании разработаны и исследованы методы, алгоритмы и способы повышения качественных показателей алгоритмов обучения с подкреплением в рамках класса задач управления роботами, способными к перемещению в трехмерных средах.

Основными результатами работы являются:

1. Методика оценки влияния состава набора наблюдений окружающей среды на качество решений, принимаемых агентом, позволяющая упорядочить наблюдения по их полезности. Проведенное исследование показало, что избыточная информация о состоянии среды может ухудшать качество решений, принимаемых агентом.
2. Модель интеграции алгоритмов обучения с подкреплением и кодировщика трансформера, которая позволяет учитывать сложную динамику системы и справляться с зашумленными и немарковскими средами. На основе данной модели был разработан алгоритм, интегрирующий кодировщик трансформера с алгоритмом Soft Actor-Critic (SAC). Эксперименты показали, что предложенная

интеграция улучшает среднее суммарное значение награды на 18,5%, а также в 80% случаев результаты превосходят или остаются на уровне оригинального SAC.

3. Метод иерархического ансамблирования алгоритмов обучения с подкреплением, который объединяет несколько алгоритмов в иерархическую структуру, что позволяет повысить качество обучения без дополнительных обращений к среде. Исследование продемонстрировало, что предложенный метод организует взаимодействие между управляющими и управляемыми алгоритмами, улучшая качественные показатели конечного решения.
4. Алгоритм обучения с подкреплением на основе метода иерархического ансамблирования, использующий алгоритм DQN в качестве управляющего и алгоритмы SAC и REDQ в качестве управляемых. Экспериментальные данные показали улучшение среднего суммарного значения награды на 2,65% по сравнению с лучшим из отдельных алгоритмов, а также превосходство или паритет по качеству во всех экспериментах.

Работы, опубликованные автором по теме диссертации

В изданиях, рекомендованных ВАК при Минобрнауки РФ:

1. Kozlov, D. Application of Transformer for Encoding States in Reinforcement Learning / D. Kozlov // Автометрия — 2024. — №5 — С. 60–68. — DOI: 10.15372/AUT20240500

В изданиях, индексируемых реферативными базами данных Web of Science / Scopus:

2. Kozlov, D. Ensemble Method for Reinforcement Learning Algorithms Based on Hierarchy / D. Kozlov, V. Myasnikov // IEEE Xplore 2023 IX International Conference on Information Technology and Nanotechnology — 2023. — С. 1–5. — DOI: 10.1109/ITNT57377.2023.10139122.
3. Kozlov, D. Comparison of Reinforcement Learning Algorithms in Problems of Acquiring Locomotion Skills in 3D Space / D. Kozlov // IEEE Xplore 2022 VIII International Conference on Information Technology and Nanotechnology — 2022. — С. 1–5. — DOI: 10.1109/ITNT55410.2022.9848647.
4. Kozlov, D. The impact of a set of environmental observations in the problem of acquiring movement skills in three-dimensional space using reinforcement learning algorithms / D. Kozlov, V. Myasnikov // IEEE Xplore 2022 VIII International Conference on Information Technology and Nanotechnology — 2022. — С. 1–5. — DOI: 10.1109/ITNT55410.2022.9848598.
5. Kozlov, D. Comparison of Reinforcement Learning Algorithms for Motion Control of an Autonomous Robot in Gazebo Simulator / D. Kozlov // IEEE Xplore 2021 VI International Conference on Information Technology and Nanotechnology — 2021. — С. 1–5. — DOI: 10.1109/ITNT52450.2021.9649145.

В других изданиях:

6. Козлов, Д.А. Сравнение алгоритмов обучения с подкреплением для управления движением автономного робота в симуляторе Gazebo / Д.А. Козлов // Информационные технологии и нанотехнологии. Сборник трудов по материалам VII Международной конференции и молодежной школы. Изд-во Самарского Университета — Самара, 2021. — P. 21442.
7. Козлов, Д.А. Сравнение алгоритмов обучения с подкреплением в задаче приобретения навыков передвижения в трёхмерном пространстве / Д.А. Козлов // Информационные технологии и нанотехнологии. Сборник трудов по материалам VIII Международной конференции и молодежной школы. Изд-во Самарского Университета — Самара, 2022. — P. 41482.
8. Козлов, Д.А. Влияние состава наблюдений окружающей среды в задаче приобретения навыков передвижения в трёхмерном пространстве при использовании алгоритмов обучения с подкреплением / Д.А. Козлов, В.В. Мясников // Информационные технологии и нанотехнологии. Сборник трудов по

материалам VIII Международной конференции и молодежной школы. Изд-во Самарского Университета — Самара, 2022. — P. 41502.

9. Козлов, Д.А. Метод ансамблирования алгоритмов обучения с подкреплением на основе иерархичности / Д.А. Козлов, В.В. Мясников // Информационные технологии и нанотехнологии. Сборник трудов по материалам IX Международной конференции и молодежной школы. Изд-во Самарского Университета — Самара, 2023. — P. 40602.
10. Козлов, Д.А. Применение трансформера для кодирования состояний в обучении с подкреплением / Д.А. Козлов // Информационные технологии и нанотехнологии. Сборник трудов по материалам X Международной конференции и молодежной школы. Изд-во Самарского Университета — Самара, 2024.

Свидетельства:

11. Свидетельство о государственной регистрации программы для ЭВМ № 2023667830. Программный модуль настройки иерархической композиции алгоритмов обучения с подкреплением / Козлова Юлия Ханифовна, Козлов Даниил Александрович. - Заявка № 2023666511. Дата поступления 02 августа 2023 г. Дата государственной регистрации в Реестре программ для ЭВМ 18.08.2023 г.