TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
ООП – Автономные интеллектуальные системы
Отделение школы (НОЦ) – Отделение информационных технологий

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРАНТА

Тема работы
Обучение с подкреплением в виртуальных средах

УДК 004.853

Обучающийся

Группа	ФИО	Подпись	Дата
8ПМ2Л	Залогин Никита Евгеньевич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор (ОИТ ИШИТР)	Спицын В.Г.	д.т.н.		

Консультант

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ст. преподаватель (ОИТ ИШИТР)	Григорьев Д.М.			

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (БШ НСП)	Аникина Е.А.	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (ОКД ИШНКБ)	Перминов В.А.	д. ф.-м. н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП, должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (ОИТ ИШИТР)	Погребной А.В.	к.т.н.		

Томск – 2024 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП

по направлению 09.04.04 Программная инженерия

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК-2	Способен управлять проектом на всех этапах его жизненного цикла
УК-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК-4	Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия
УК-5	Способен анализировать и учитывать разнообразие культур в процессе меж- культурного взаимодействия
УК-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК-1	Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;
ОПК-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач;
ОПК-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;
ОПК-4	Способен применять на практике новые научные принципы и методы исследований;
ОПК-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем;
ОПК-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности;
ОПК-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях;
ОПК-8	Способен осуществлять эффективное управление разработкой программных средств и проектов.
Профессиональные компетенции	
ПК-9	Способен выбирать технологии и средства разработки программного обеспечения, включая системы управления исходным кодом
ПК-10	Способен исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта
ПК-11	Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач
ПК-12	Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Код компетенции	Наименование компетенции
ПК-13	Способен руководить проектами по созданию, поддержке и использованию системы искусственного интеллекта на основе нейросетевых моделей и методов
ПК-14	Способен выбирать, разрабатывать и проводить экспериментальную проверку работоспособности программных компонентов систем искусственного интеллекта по обеспечению требуемых критериев эффективности и качества функционирования
ПКО-1	Способен понимать фундаментальные принципы работы современных систем искусственного интеллекта, разрабатывать правила и стандарты взаимодействия человека и искусственного интеллекта и использовать их в социальной и профессиональной деятельности
ПКО-2	Способен разрабатывать алгоритмы и программные средства для решения за- дач в области создания и применения искусственного интеллекта
ПКО-3	Способен осуществлять эффективное управление проектами по разработке и внедрению систем искусственного интеллекта
ПКО-4	Способен применять методы системного анализа и программное обеспечение для системного моделирования с целью решения задач в сфере исследовательской деятельности

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
ООП/ОПОП – Автономные интеллектуальные системы
Отделение школы (НОЦ) – Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Погребной А.В.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ2Л	Залогин Никита Евгеньевич

Тема работы:

Обучение с подкреплением в виртуальных средах	
Утверждена приказом директора (дата, номер)	№ 39-15/с от 08.02.2024

Срок сдачи обучающимся выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Использование алгоритмов обучения с подкреплением в виртуальных средах.
Перечень разделов пояснительной записки подлежащих исследованию, проектированию и разработке *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Аналитический обзор 2. Алгоритмы обучения с подкреплением 3. Выбор платформы симуляции 4. Проведение экспериментов 5. Финансовый менеджмент. 6. Социальная ответственность. 7. Раздел на английском языке.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Демонстрации виртуальной среды 2. Топологии нейросетей 3. Графики обучения агентов 4. Диаграмма Ганта.

Консультанты по разделам выпускной квалификационной работы *(с указанием разделов)*	
Раздел	**Консультант**
Основная часть	Старший преподаватель ОИТ ИШИТР Григорьев Д.С.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент БШ НСП к.э.н. Аникина Е.А.
Социальная ответственность	Доцент ОКД ИШНКБ д. ф.-м. н. Перминов В.А.
Раздел на английском языке	Старший преподаватель ОИЯ ШОН Асадуллина Л.И.
Названия разделов, которые должны быть написаны на иностранном языке:	
Chapter 1. Analytical review	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель / консультант (при наличии):

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Профессор (ОИТ ИШИТР)	Спицын В.Г.	д.т.н.		
Ст. преподаватель (ОИТ ИШИТР)	Григорьев Д.С.			

Задание принял к исполнению обучающийся:

Группа	**ФИО**	**Подпись**	**Дата**
8ПМ2Л	Залогин Никита Евгеньевич		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
ООП/ОПОП – Автономные интеллектуальные системы
Уровень образования – Магистратура
Отделение школы (НОЦ) – Отделение информационных технологий
Период выполнения – Весенний семестр 2023/2024 учебного года

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ2Л	Залогин Никита Евгеньевич

Тема работы:

Обучение с подкреплением в виртуальных средах

Срок сдачи обучающимся выполненной работы:	

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
	Основная часть ВКР	60
	Раздел «Социальная ответственность»	20
	Раздел «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»	20

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор (ОИТ ИШИТР)	Спицын В.Г.	д.т.н.		

Консультант

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ст. преподаватель (ОИТ ИШИТР)	Григорьев Д.С.			

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (ОИТ ИШИТР)	Погребной А.В.	к.т.н.		

Обучающийся

Группа	ФИО	Подпись	Дата
8ПМ2Л	Залогин Никита Евгеньевич		

РЕФЕРАТ

Выпускная квалификационная работа содержит 110 страниц, 20 рисунков, 39 таблиц, 59 источников, 1 приложение.

Ключевые слова: обучение с подкреплением, сверточная нейронная сеть, робот манипулятор, PyBullet, DQN, PPO,

Объект исследования: Алгоритмы обучения с подкреплением.

Предмет исследования: Обучение агента в виртуальной среде с манипулятором.

Цель работы: Разработка метода на основе обучения с подкреплением для обучения агента управлению манипулятором.

В процессе исследования был проведен обзор и анализ существующих решений, реализованы несколько алгоритмов и произведено улучшение их реализаций, обучены агенты для взаимодействия с выбранной виртуальной средой.

В результате исследования были реализованы алгоритмы обучения с подкреплением для обучения агента манипулированию объектами в виртуальной среде KukaDiverseObjectEnv, а также проведены улучшения решений.

Степень внедрения: Выпущено 3 статьи по данной теме.

Область применения: Робототехника.

Экономическая эффективность/значимость работы: данное исследование позволяет автоматизировать управление роботом манипулятором в различных сложных средах на основе симуляции окружения и применения методов обучения с подкреплением.

В будущем планируется апробация других алгоритмов обучения с подкреплением, улучшение уже примененных решений, разработка авторской системы симуляции среды.

ОГЛАВЛЕНИЕ

ОСНОВНЫЕ СОКРАЩЕНИЯ И ОБОЗНАЧЕНИЯ

RL – Reinforcement Learning; Обучение с подкреплением.

DQN – Deep Q-Networks.

PPO – Proximal Policy Optimization.

TRPO – Region Policy Optimization.

LSTM – Long short-term memory.

ALE – The Arcade Learning Environment; Среда обучения аркадным играм.

NLP – Natural language processing; Обработка естественного языка.

xLSTM – Extended Long Short-Term Memory.

TD3 – Twin Delayed Deep Deterministic Policy Gradient.

SAC – Soft Actor Critic.

HER – Hindsight Experience Replay.

IPG – Interpolated Policy Gradients.

SRL – State Representation Learning.

IL – Imitation learning.

PCCL – Precision-Based Continuous Curriculum Learning.

SWOT – метод стратегического планирования, заключающийся в выявлении факторов внутренней и внешней среды организации и разделении их на четыре категории: Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы).

ВВЕДЕНИЕ

В последние годы робототехника приобретает все более значимое место в различных сферах, таких как производство, логистика и медицина. Роботы-манипуляторы благодаря своей высокой точности и повторяемости становятся незаменимыми инструментами для выполнения сложных промышленных задач. Одним из основных вызовов в робототехнике является обучение роботов новым навыкам. Традиционные методы программирования роботов требуют значительных временных и трудовых затрат, а также недостаточно эффективны для задач, требующих адаптивности к изменяющимся условиям [1, 2]. Для решения этой проблемы разработаны алгоритмы обучения с подкреплением (Reinforcement Learning, RL), которые представляют собой перспективный подход для создания агентов, способных самостоятельно обучаться через метод проб и ошибок и принимать решения в сложных и динамических средах, как виртуальных, так и реальных.

Виртуальные среды, такие как PyBullet [3] и Gazebo [4], предоставляют безопасные и удобные платформы для обучения роботов. Они позволяют быстро и эффективно создавать различные сценарии, тестировать алгоритмы RL и оценивать их эффективность.

Алгоритмы обучения с подкреплением представляют собой активно развивающуюся область исследований в сфере машинного обучения. Их применение и дальнейшее развитие могут привести к созданию более интеллектуальных и адаптивных систем, способных решать сложные задачи и обеспечивать оптимальное поведение в разнообразных условиях. В последние годы исследовательские группы, достигли значительных успехов в области RL для роботов, что подтверждают работы, такие как TossingBot [5] и Ravens - Transporter Networks [6].

Актуальность данной работы обусловлена стремительным развитием технологий обучения с подкреплением и их применением в виртуальных средах, что открывает новые возможности для автоматизации и оптимизации

сложных задач, таких как управление роботами-манипуляторами.

Для достижения целей данной работы и проверки гипотез использовались различные методы исследования. Анализ литературы, где был проведен обзор научных источников по темам обучения с подкреплением (RL), алгоритмов DQN и PPO, а также их применению в виртуальных средах. Изучались механизмы и архитектуры нейронных сетей, включая адаптацию механизмов внимания для улучшения алгоритма PPO. Программирование охватывало разработку моделей RL. Эксперименты и синтез были использованы для подбора гиперпараметров и итерационного улучшение алгоритма PPO. Сравнительный анализ проводился для оценки производительности исходных и модифицированных версий алгоритмов по ключевым метрикам, а также сравнения с аналогичными решениями.

Цель данной работы заключается в изучении алгоритмов и моделей машинного обучения с подкреплением, применяемых для создания агентов в виртуальных средах, а также в рассмотрении, использовании и улучшении примеров их реализации.

Для достижения поставленной цели были определены и решены следующие задачи:

1. Исследовать различные алгоритмы RL, применяемые для создания агентов в виртуальных или реальных средах;
2. Составить обзор актуальных статей и исследований, посвященных популярным алгоритмам RL и их модификациям, а также их применению в средах с манипуляторами. В рамках этого этапа будет выполнен анализ преимуществ и недостатков различных подходов, их эффективности и применимости в реальных и виртуальных средах. Такой обзор позволит выявить тенденции и перспективные направления в области RL для робототехники;
3. Провести сравнительный анализ популярных платформ для создания виртуальных сред, таких как PyBullet, Gazebo, MuJoCo и CoppeliaSim. Анализ будет включать оценку функциональности,

удобства использования, возможностей для интеграции с алгоритмами RL, производительности и поддержки сообщества. На основе этого анализа будет выбрана наиболее подходящая платформа для проведения дальнейших экспериментов;

4. Проведение экспериментов с выбранными алгоритмами RL в среде с манипулятором. Будет выполнено сравнение различных решений с точки зрения их эффективности, стабильности и способности к обучению;
5. Проведение сравнительного анализа обученных в ходе работы агентов между собой и с несколькими аналогичными решениями других авторов.

1 АНАЛИТИЧЕСКИЙ ОБЗОР

1.1 Обучение с подкреплением

Обучение с подкреплением (Reinforcement learning, RL) – это область машинного обучения, в которой агенты автономно учатся, взаимодействуя с окружающей средой. RL основан на принципе проб и ошибок, где агент получает вознаграждение за желаемое поведение и штрафы за нежелательное.

К основным преимуществам обучения с подкреплением относится:

- автономное обучение: агенты могут обучаться без явного программирования;
- универсальность: RL применим к широкому спектру задач, включая управление роботами, игры и финансы;
- эффективность: RL способен находить оптимальные или близкие к оптимальным стратегии для сложных задач.

К недостаткам же можно отнести:

- сложность: RL может требовать значительных вычислительных ресурсов для обучения;
- проблема переобучения (overfitting): RL-агенты могут слишком сильно подстраиваться под обучающие данные и плохо справляться с новыми условиями.
- проблема локальных оптимумов: RL-агенты могут застревать в локальных оптимумах и не достигать глобального оптимума.

Одним из популярных алгоритмов RL является Deep Q-Networks (DQN) (Volodymyr Mnih, и другие, 2013) [7], представленный в статье «Human-level control through deep reinforcement learning». Традиционно методы обучения с подкреплением испытывали трудности с многомерными сенсорными входными данными, подобными тем, которые встречаются в реальных условиях. В статье представлена сеть DQN –подход, который использует глубокие нейронные сети для прямого обучения эффективным политикам

управления на основе многомерных сенсорных входных данных.

В последнее время исследователей привлекает алгоритм Proximal Policy Optimization (PPO) (Schulman и другие, 2017) [8] предложенный в качестве усовершенствования алгоритма Trust Region Policy Optimization (TRPO) (Schulman и другие, 2017) [9]. PPO использует более простую в вычислительном плане целевую функцию, которая не напрямую оценивает, насколько хороша политика агента для достижения максимальной награды. Вместо этого, он использует функцию, которая приближает это значение. Это делает процесс обучения более стабильным и эффективным. Несмотря на это в статье показано, что PPO обладает большей эффективностью использования образцов, чем TRPO, во многих задачах управления. PPO также демонстрирует хорошую эмпирическую эффективность в среде обучения аркадным играм (ALE), которая включает игры Atari.

Отдельными разработчиками был реализован алгоритм PPO [10] с описанием 37 ключевых аспектов реализации алгоритма для различных окружений и задач

- 13 основных деталей реализации;
- 9 деталей для игр Atari;
- 9 деталей для робототехнических задач с непрерывным пространством действий;
- 5 деталей для моделей Long short-term memory (LSTM) (Sepp Hochreiter и другие, 1997) [11];
- 1 деталь для пространств многодискретных (MultiDiscrete) действий.

Авторы статьи [12] провели исследование 20 различных аспектов реализации на производительность алгоритмов градиентного обучения политики (deep policy gradients) на примере двух популярных методов: PPO, TRPO. В свою очередь в статье [13] исследование эффективности затронуло уже 68 различных аспектов реализации алгоритмов обучения с актором-критиком (A2C) на основе политик (on-policy) в задачах непрерывного

управления. Авторы обучили более 250000 агентов в пяти различных средах непрерывного управления с возрастающей сложностью и провели масштабное эмпирическое исследование, чтобы выявить, какие детали реализации наиболее влияют на производительность обученного агента.

В целом авторы статей [10, 12, 13] сделали следующие выводы:

- архитектура нейронной сети: выбор архитектуры и функции активации играет критическую роль в производительности;
- "Code-level optimizations": неочевидные детали реализации могут фундаментально влиять на производительность алгоритмов, меняя их поведение непредсказуемым образом;
- гиперпараметры: размер пакета, скорость обучения, коэффициент энтропии и дисконтирования и остальные являются крайне важными факторами как для производительности, так и для результативности;
- доверительная область: оптимизации часто определяют природу доверительной области, используемой алгоритмами;
- превосходство PPO над TRPO: большая часть преимуществ PPO над TRPO (и даже стохастическим градиентным спуском) обусловлена "code-level optimizations";
- модульное проектирование: крайне рекомендуется разрабатывать алгоритмы RL с модульной структурой для точной оценки влияния каждой детали реализации;
- тщательная оценка: рекомендуется тщательно анализировать методы RL, выходя за рамки простых сравнений производительности;
- среда: результаты обучения с одними и теми же гиперпараметрами могут сильно отличаться в зависимости от типа задачи.

Последним большим прорывом в области применения нейронных сетей для задач обучения с подкреплением стали модели трансформеры и их механизм внимания (Ashish Vaswani и другие, 2017) [14]. Данная работа

кардинально изменила подходы к обработке последовательностей данных и позволила достичь выдающихся результатов в задачах обработки естественного языка (NLP) и компьютерного зрения.

Дальнейшее свое развитие трансформеры получили в работе Decision Transformer: Reinforcement Learning via Sequence Modeling (Lili Chen и другие, 2021) [15] где исследователи предложили новую архитектуру Decision Transformer для задач RL. Decision Transformer переосмысливает задачу RL как задачу последовательного моделирования, предсказывая будущие действия агента на основе его текущего состояния и истории прошлых действий и вознаграждений. Этот подход позволяет использовать мощные методы из NLP, такие как механизмы внимания и самовнимания, для обучения эффективных политик RL.

Одним из конкурентов трансформеров является Long short-term memory (LSTM) (Sepp Hochreiter и другие, 1997) [11]. Несмотря на то, что LSTM значительно уступает трансформерам в эффективности и возможности параллелизации обучения из-за своей рекуррентной природы, они хорошо зарекомендовали себя в ряде задач RL. Например, модели AlphaStar для игры StarCraft II [16] и OpenAI Five для Dota 2 [17] использовали LSTM и показали выдающиеся результаты.

Также, представленная в 2024 году модель Extended Long Short-Term Memory (xLSTM) (Maximilian Beck и другие, 2024) [18] в которой исследователи добились значительных улучшений по сравнению с оригинальной моделью LSTM. В задачах NLP новая модель показала результаты, превосходящие или сопоставимые с распространенными аналогами, включая GPT-3 [19]. Предполагается, что xLSTM может стать достойным конкурентом трансформерам и другим моделям в различных областях, включая RL [20].

1.2 Обучение с подкреплением в средах с манипуляторами

Исследования в области RL агентов в виртуальных средах с манипуляторами получают широкое распространение. Это обусловлено тем, что традиционные методы программирования роботов требуют значительных временных и трудовых затрат, и, кроме этого, далеко не всегда подходят для задач, где требуется адаптивность к меняющимся условиям.

Так, например авторы статьей [21] проводят исследование различных алгоритмов RL на задачах достижения объекта манипулятором, где алгоритмы Twin Delayed Deep Deterministic Policy Gradient (TD3), Soft Actor Critic (SAC) и TRPO оказываются лидерами, а использование стратегии Hindsight Experience Replay (HER) для обогащения внутреннего сигнала вознаграждения в off-policy алгоритмах (SAC и TD3) слегка снижает стабильность обучения (повторяемость и эффективность выборки), но позволяет агенту обучаться эффективной политике действий в средах с инициализацией цели в случайных координатах, так же авторы столкнулись с тем, что перенос агента из симулятора в реальность оказался сложен из-за разницы в динамике (проблема sim-to-real). Реальный робот подвержен шуму и нестационарности, но частичное обучение на реальном конкретном роботе позволит политике адаптироваться к реальному миру.

Авторы [22] провели аналогичное исследование алгоритмов в двух других симуляционных средах робототехники с управлением на основе зрения, KukaDiverseObjectEnv и RaceCarZedGymEnv, где в качестве наблюдений используются RGB-изображения и имеют непрерывное пространство действий. В данных задачах наибольшую эффективность показал алгоритм Interpolated Policy Gradients (IPG), а использование стратегии HER, как и у авторов предыдущей статьи позволило значительно улучшить результат.

Автор [23] исследовал различные подходы для обучения агентов поднимать объекты, нажимать кнопки и управлять многопалой рукой. В

результате исследования было выведено, что оригинальный PPO эффективно решает базовые задачи захвата. Комбинированный подход (State Representation Learning (SRL) + RL) позволяет решать более сложные задачи нажатия кнопок, извлекая полезные знания из предварительного обучения состояний. Обучение с имитацией (Imitation learning IL) эффективно для обучения робота сложным многопалым манипуляциям путем подражания действиям человека.

Авторы [24] предлагают метод непрерывного обучения по учебному плану (Curriculum) на основе точности Precision-Based Continuous Curriculum Learning (PCCL) для ускорения обучения с подкреплением (RL) в многостепенных задачах достижения цели. В результате было выяснено, что PCCL улучшает эффективность обучения и производительность алгоритмов RL, особенно в сценариях с редким и двоичным вознаграждением

Авторы статьи [25] сосредоточились на вопросе безопасности обучения агента. Из-за трудностей в моделировании определенных форм поведения (например, взаимодействия с человеком, реальных сценариев дорожного движения и т.д.) обучение агента часто переносится в реальный мир, где вопросы безопасности имеют гораздо большое значение, поскольку, во время исследования среды, агент может совершить опасное для окружающих действие. Для решения данной проблемы авторы использовали алгоритм PPO с ограничениями Лагранжа. В результате алгоритм демонстрировал более длительное обучение, но в конечном итоге достигал той же эффективности, что и обычный PPO, при этом гарантируя большую безопасность своих действий для окружающих.

Авторы [5] разработали систему TossingBot на основе рекуррентных нейронных сетей для обучения робота-манипулятора хватать и бросать произвольные объекты за пределами зоны досягаемости робота. При этом агент одновременно обучается стратегиям хватания и броска. Ключевым элементом является использование «Остаточной физики» - гибридного контроллера, который применяет машинное обучение для предсказания

остаточных погрешностей поверх параметров управления, рассчитанных с помощью физической модели. Такой подход позволяет фокусировать обучение на тех аспектах динамики, которые сложно моделировать аналитически.

В более поздней работе, авторы [6] разработали подход к управлению роботами-манипуляторами на основе визуальной информации, позволяющий решать задачи:

- строительства пирамиды из блоков;
- сборки наборов с новыми, невиданными ранее объектами;
- перемещения куч мелких предметов с использованием замкнутой обратной связи.

В работе используется несколько моделей нейросетей для сегментации и распознавания объектов, но основной является модель Transporter Network на основе внимания, которая генерирует траектории движения робота и предсказывает смещение объекта в пространстве при определенных движениях робота. Данная модель обучается на наборе данных изображений объектов и их соответствующих смещений.

В целом авторы данной статьи предполагают, что такой подход позволит решать более сложные задачи, такие как управления роботами в режиме реального времени с высокой частотой или использование инструментов.

2 АЛГОРИТМЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

2.1 Deep Q-Network (DQN)

Deep Q-Network (DQN) [7] – это off-policy [26] алгоритм RL. Традиционные методы Q-обучения в условиях больших пространств состояний и действий сталкиваются с проблемой невозможности хранения Q-значений для каждого варианта из-за большого объема информации. Для решения данной проблемы DQN использует нейронную сеть для приближения функции Q, которая оценивает ожидаемый суммарное дисконтированное вознаграждение агента для каждой пары состояние-действие в среде, что позволяет ей работать в сложных ситуациях [27]. Схематичное представление нейронной сети DQN представлено на рисунке 2.1 [7].

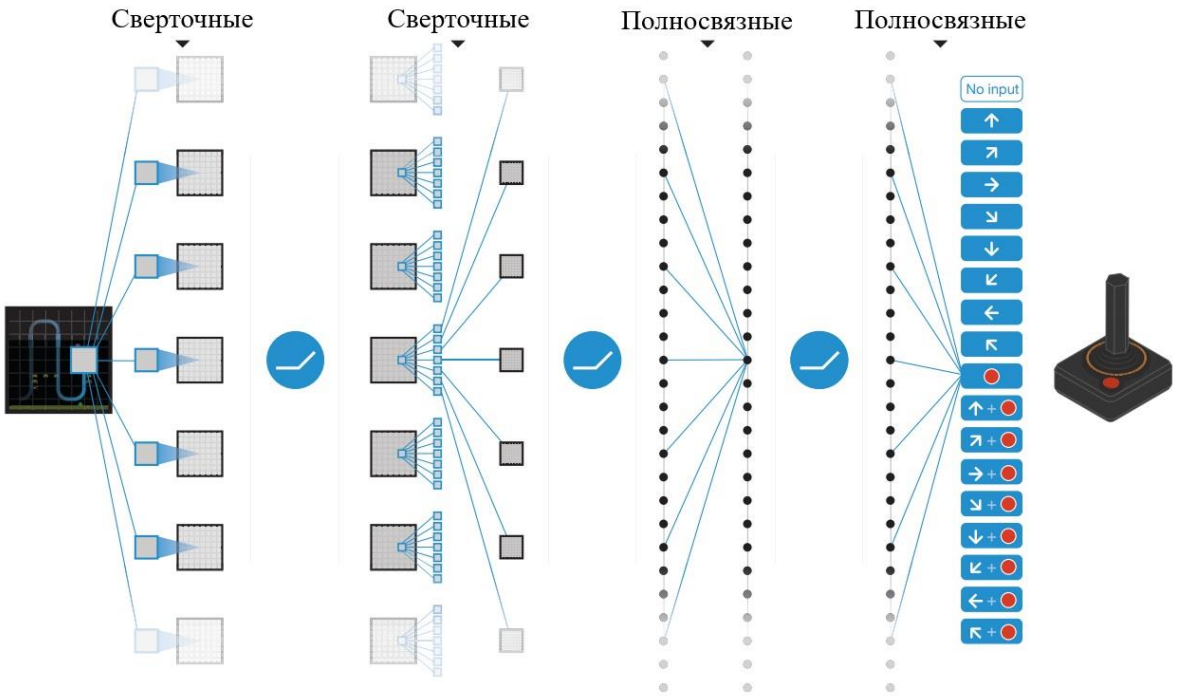


Рисунок 2.1 – Схематичное представление нейронной сети DQN

Главной целью DQN является поиск политики, направленной на максимизацию дисконтированного совокупного вознаграждения R_{t0} вычисляемого по формуле 2.1:

$$R_{t_0} = \sum_{t=t_0}^{?} \gamma^{t-t_0} r_0, \quad (2.1)$$

где γ – дисконтирующий коэффициент для будущих вознаграждений в пределах от 0 до 1;

r – награда исходящая от среды.

Основная идея Q-learning заключается в поиске функции, которая покажет какой была бы награда, если бы мы предприняли какое-либо действие в конкретном состоянии среды. В таком случае можно было бы легко разработать политику, которая максимизировала бы выгоду агента. Пример такой функции приведен в формуле 2.2

$$\pi^*(s) = \mathrm{a}rgmax_a Q^*(s,a), \quad (2.2)$$

где π – политика;

s – состояние среды;

a – действие в среде.

Однако алгоритм не содержит информации о среде в которой находится, поэтому у него нет доступа к значению Q функции. А так как нейронные сети являются универсальными аппроксиматорами функций, в алгоритме DQN было предложено создать такую сеть и обучить ее для предсказания значения Q функции.

Для тренировки нейросети DQN будет использовано два механизма для значительного улучшения и стабилизации процесса обучения:

1. Воспроизведение опыта (Experience Replay). Это позволит сохранить последовательности переходов et = (S_t,A_t,R_t,S_{t+1}) (где e – эпизод обучения, S – множество состояний среды, A – множество действий, R – множество наград), наблюдаемых агентом в буфер воспроизведения Dt = {e1,…,et}. Этот буфер содержит кортежи опыта из множества эпизодов. Во время обновлений Q-learning образцы случайным образом выбираются из буфера воспроизведения, что позволяет использовать один и тот же образец многократно. Опытное воспроизведение улучшает эффективность использования данных, устраняет корреляции в

последовательностях наблюдений и сглаживает изменения в распределении данных. Это значительно повышает стабильность и эффективность процесса обучения, так как агент обучается на более разнообразном и несвязном наборе данных.

2. Периодическое обновление целевой сети (Periodically Updated Target): Q-функция оптимизируется по целевым значениям, обновляемым с определенным периодом. Q-сеть клонируется и «замораживается» как целевая сеть каждые C шагов (C – гиперпараметр). Эта модификация делает процесс обучения более стабильным, так как позволяет избежать краткосрочных колебаний в значениях Q-функции.

Для используемого правила обновления политики будет использоваться тот факт, что каждая Q функция для некоторой политики подчиняется уравнению Беллмана приведенному в формуле 2.3.

$$Q^{\pi}(s,a) = r + \gamma Q^{\pi}(s', \pi(s')), \tag{2.3}$$

где π – политика;

s – состояние среды;

a – действие в среде.

Ошибка временной разницы вычисляется по формуле 2.4.

$$\delta = Q^{*}(s,a) - (r + \gamma max_{a} Q(s',a)), \tag{2.4}$$

Чтобы свести к минимуму данную ошибку, будет использоваться коэффициент потерь по Хуберу. Он действует как среднеквадратичная ошибка, когда ошибка мала, и как средняя абсолютная ошибка, когда ошибка велика. Такое свойство делает сеть более устойчивой к выбросам, когда оценки Q зашумлены. Функция потерь приведена в формулах 2.5 - 2.6

$$\zeta = \frac{1}{|B|} \sum_{(s,a,s',r)} \zeta\,(\delta), \tag{2.5}$$

$$\zeta(\delta) = \begin{cases} \frac{1}{2}\delta^2, & for\ |\delta| \le 1 \\ |\delta| - \frac{1}{2}, & otherwise \end{cases}, \tag{2.6}$$

где B – пакет траекторий.

Кратко алгоритм DQN представлен в виде псевдокода в [27].

К недостаткам алгоритма DQN можно отнести:

1. Смертельная триада (Deadly Triad): Основная проблема в DQN связана с сочетанием трёх ключевых компонентов: off-policy обучения, нелинейного приближения функций и бутстрепинга (bootstrapping). При совокупности всех трех элементов в одном алгоритме обучения с подкреплением, процесс обучения может становиться нестабильным и трудно сходящимся. Для решения данной проблемы были использованы механизмы воспроизведения опыта и периодического обновления целевой сети.
2. Требования к памяти и вычислительным ресурсам: Поддержание буфера воспроизведения и многократные обновления целевых сетей требуют значительных ресурсов памяти и вычислительной мощности. Это может стать узким местом при применении DQN в масштабных или ресурсоёмких задачах.
3. Проблемы с редкими и высокими вознаграждениями: DQN может сталкиваться с трудностями в задачах, где вознаграждения редки и высоки. Алгоритм может застревать в субоптимальных стратегиях, не обнаруживая глобально оптимальных действий. Это связано с тем, что агент недостаточно часто получает значительные вознаграждения, чтобы корректно обновить свою стратегию.
4. Чувствительность к гиперпараметрам: Эффективность DQN сильно зависит от выбора гиперпараметров, таких как размер буфера воспроизведения, коэффициенты обучения и частота обновления целевой сети. Неправильная настройка этих параметров может привести к нестабильности и ухудшению производительности.

2.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) [8] – это on-policy [26] алгоритм RL использующий архитектуру актор-критик (actor-critic) [28], который используется для обучения агентов действовать в сложных средах. PPO является более стабильной и эффективной версией алгоритма TRPO (Trust Region Policy Optimization) [29, 30]. PPO применяет две ключевые идеи для обеспечения стабильности и эффективности обучения.

Первая идея – это обновление политики на основе отношения вероятностей между старой и новой политиками, для ограничения её изменение. Отношение политик (r(θ)) вычисляется по формуле 2.7 [13]. Это означает, что обновление происходит малыми шагами, избегая значительных изменений, которые могут привести к нестабильности обучения. PPO вводит штраф за изменение политики, используя функцию клипирования, которая ограничивает величину обновления. Этот механизм контролирует отклонение новой политики от старой, обеспечивая более плавное и безопасное обучение.

Вторая идея – это использование значений преимуществ (advantages) для выравнивания обновлений политики, в противовес абсолютным вероятностям. Преимущество представляет собой разницу между фактическим вознаграждением и ожидаемым вознаграждением, что позволяет более точно оценивать эффективность действий агента. Это способствует более сбалансированному и стабильному обучению, так как обновления основываются на относительной выгоде действий, а не на их абсолютных значениях.

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, \tag{2.7}$$

где π – политика;

θ - параметры политики;

$\pi(a|s)$ определяет вероятность выбора действия a в состоянии s.

Без ограничения на расстояния между новой и старой политиками,

максимизация целевой функции привела бы к нестабильности с чрезвычайно большими обновлениями параметров и большими коэффициентами политики. PPO накладывает ограничение, заставляя оставаться в пределах небольшого интервала, используя функцию клипирования. Вычисляется целевая функция (J^{CLIP} (θ)) с этими изменениями по формуле 2.8 [13].

$$J^{CLIP}(\theta) = \mathrm{E}[min(r(\theta)\hat{A}_{\theta_{old}}(s,a), clip(r(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_{\theta_{old}}(s,a)], \quad (2.8)$$

где A(s,a) – функция преимуществ;

ε – гиперпараметр.

В дополнение к ограниченному вознаграждению, целевая функция дополняется членом ошибки при оценке значения и энтропией, чтобы стимулировать достаточное исследование среды. В итоге целевая функция (J^{CLIP} (θ)) будет вычисляться по формуле 2.9 [13].

$$J^{CLIP}(\theta) = \mathrm{E}[J^{CLIP}(\theta) - c_1(V_0(s) - V_{target})^2 + c_2 H(s, \pi_\theta(.))], \quad (2.9)$$

где c1 и c2 – константные гиперпараметры;

V(s) - ожидаемый возврат состояния s.

Вместо обновления политики на основе всех данных сразу, PPO использует мини-пакеты данных (mini-batch) и SGD (Stochastic Gradient Descent), что позволяет более эффективно использовать вычислительные ресурсы и улучшать обобщающую способность модели.

Кратко алгоритм PPO представлен в виде псевдокода в [30].

К недостаткам алгоритма PPO можно отнести:

1. На пространствах непрерывных действий стандартный PPO становится нестабильным, когда вознаграждения исчезают за пределами ограниченной области. В таких ситуациях алгоритм может испытывать трудности с обучением, так как отсутствие вознаграждений вне определённых границ приводит к недостатку сигналов для корректировки действий агента. Это ограничивает способность агента к исследованию новых стратегий и может приводить к преждевременной остановке обучения.
2. На пространствах дискретных действий с редкими высокими

вознаграждениями стандартный PPO часто застревает на субоптимальных действиях (локальных максимумах) и не находит глобально оптимальные действия. Алгоритм недостаточно часто сталкивается с вознаграждениями, чтобы скорректировать свою стратегию должным образом.

3. Политика чувствительна к начальной инициализации, когда вблизи начальной точки имеются локально оптимальные действия. Алгоритм может быстро стабилизироваться на этих действиях, не исследуя более выгодные стратегии. Это может привести к тому, что агент будет застревать в субоптимальных стратегиях, которые не являются наилучшими в долгосрочной перспективе.
4. Для достижения хорошей производительности алгоритм PPO требует тщательной настройки множества гиперпараметров, таких как коэффициент клипирования, размер батча и коэффициенты обучения. Неправильная настройка этих параметров может привести к нестабильности обучения или недостаточной эффективности.
5. Высокие вычислительные затраты. PPO использует mini-batch, SGD и другие методы, требующие значительных вычислительных ресурсов. Это может сделать обучение медленным и дорогостоящим, особенно для сложных задач или задач с высокими размерностями действий.
6. Точность функции стоимости играет критическую роль в производительности PPO. Ошибки в оценке стоимости могут привести к неправильным обновлениям политики, что отрицательно сказывается на процессе обучения.

3 ВЫБОР ПЛАТФОРМЫ СИМУЛЯЦИИ

3.1 PyBullet

PyBullet [3] – это библиотека на языке Python с открытым исходным кодом для физического моделирования и симуляции роботов, основанная на движке Bullet Physics. PyBullet обладает следующими ключевыми особенностями:

- высокая скорость: обеспечивает быструю и эффективную симуляцию, что позволяет проводить эксперименты в реальном времени.
- функциональность: содержит широкий набор функций, включая:
  - физическое моделирование роботов и объектов;
  - визуализация симуляции;
  - управление роботами;
  - имитация различных датчиков и сенсоров;
  - экспорт данных симуляции.
- открытый исходный код: распространяется под лицензией zlib, что позволяет свободно использовать, модифицировать и распространять его

3.2 Gazebo

Gazebo [4] – это платформа симуляции роботов с открытым исходным кодом, разработанный Open Robotics. Он позволяет создавать реалистичные симуляции роботов в разнообразных средах. Gazebo обладает следующими ключевыми особенностями:

- реалистичность: использует передовые методы физического моделирования, что обеспечивает реалистичное поведение роботов в симуляции;

– функциональность: содержит обширный набор функций, включая:
  - поддержку различных роботов, датчиков и сенсоров;
  - моделирование различных сред, таких как городские пейзажи, поля и леса;
  - планирование траекторий и управление роботами;
  - поддержку многопользовательского режима;
  - визуализацию симуляции;
  - экспорт данных симуляции;
– открытый исходный код: Gazebo распространяется под лицензией Apache 2.0, что позволяет свободно использовать, модифицировать и распространять его.

3.3 MuJoCo

MuJoCo (Multi-Joint Dynamics with Contact) [31] – это высокоскоростной физический движок и симулятор роботов, разработанный DeepMind. Несмотря на его платность, MuJoCo предоставляет высокую точность и скорость, что делает его идеальным инструментом для исследований в области робототехники и машинного обучения. MuJoCo обладает следующими ключевыми особенностями:

– высокая скорость: обеспечивает симуляции в реальном времени, что важно для обучения роботов и разработки алгоритмов управления;
– точность: использование передовых методов физического моделирования гарантирует реалистичное поведение роботов в симуляции.
– функциональность: содержит широкий спектр функций, включая:
  - поддержку различных роботов, датчиков и сенсоров;
  - моделирование различных сред, таких как городские пейзажи,

поля и леса;

- планирование траекторий и управление роботами;
- поддержку многопользовательского режима;
- визуализацию симуляции;
- экспорт данных симуляции.

– открытый исходный код: MuJoCo распространяется под лицензией Apache 2.0, что позволяет свободно использовать, модифицировать и распространять его.

3.4 CoppeliaSim

CoppeliaSim (ранее известный как V-REP) [32]– это платная платформа для симуляции роботов с открытым исходным кодом, разработанная компанией Coppelia Robotics. Она позволяет создавать реалистичные симуляции роботов и других механических систем в различных средах. CoppeliaSim обладает следующими преимуществами:

– реалистичность: использует передовые методы физического моделирования, что обеспечивает реалистичное поведение роботов и других объектов в симуляции;

– функциональность: содержит широкий набор функций, включая:

- поддержку различных роботов, датчиков, сенсоров и механизмов;
- моделирование различных сред, таких как городские пейзажи, поля, леса и производственные цеха;
- планирование траекторий и управление роботами;
- поддержку многопользовательского режима.
- визуализацию симуляции в 3D.
- экспорт данных симуляции.

– простота использования: обладает интуитивно понятным

интерфейсом, что делает его доступным даже для начинающих пользователей;

- актуальность: является активно развивающимся проектом, в который регулярно добавляются новые функции.
- открытый исходный код: CoppeliaSim распространяется под лицензией GNU GPL, что позволяет свободно использовать, модифицировать и распространять его.

3.5 Сравнительный анализ

Для проведения сравнительного анализа и выбора лучше из описанных выше платформ выведем важные метрики и их ценность от 0 до 1. После чего оценим каждую платформу по описанным метрикам от 0 до 10 и вычислим итоговую оценку. В таблицах 3.1 и 3.2 приведены метрики и их ценность, а также оценки каждой платформы по каждой метрике.

Таблица 3.1 – Метрики для оценки платформ и их ценность

Метрика	Ценность
Цена	1
Визуализация	0,6
Открытый исходный код	0,8
Экспорт данных симуляции	0,9
Поддержка различных датчиков и сенсоров	0,8
Актуальные обновления	0,8
Простота использования	0,8
Сообщество	0,8
Производительность	1
Дополнительные возможности	0,5

Таблица 3.2 – Оценка платформ по метрикам

Метрика	PyBullet	Gazebo	MuJoCo	CoppeliaSim
Цена	10	10	0	2
Визуализация	8	8	9	8
Открытый исходный код	10	10	10	10
Экспорт данных симуляции	8	8	9	8
Поддержка различных датчиков и сенсоров	8	10	9	9
Актуальные обновления	9	9	8	9
Простота использования	9	6	6	6
Сообщество	10	10	7	9
Производительность	9	8	10	7
Дополнительные возможности	8	8	7	9
Итоговая оценка	71.8	70	59	59,9

Как видно из таблицы 3.2 PyBullet и Gazebo являются лидерами по совокупности баллов. Обе платформы – это отличный выбор как для исследователей, так и для разработчиков робототехники. Но PyBullet отличается легкостью в освоении, производительностью и популярностью среди научного сообщества. В связи с этим выбор был сделан в пользу PyBullet.

4 ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТОВ

4.1 Среда обучения агента

KukaDiverseObjectEnv [33] – виртуальная среда симуляции с открытым исходным кодом на платформе PyBullet, разработанная для обучения агентов манипулированию объектами. В данной среде используется робот-манипулятор с семью степенями свободы. На рисунке 4.1 представлена среда KukaDiverseObjectEnv.

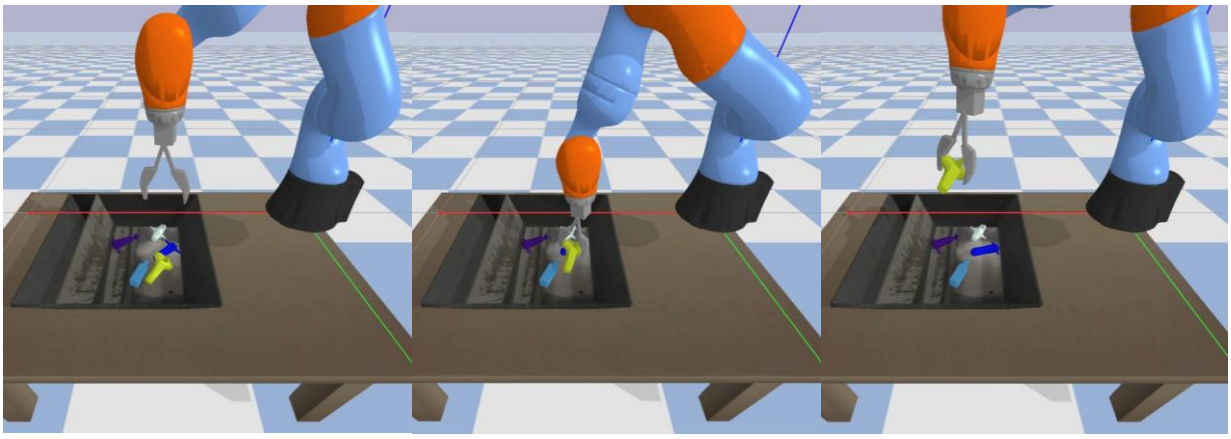


Рисунок 4.1 – Среда KukaDiverseObjectEnv

Среда KukaDiverseObjectEnv содержит следующие объекты:

1. Стол с контейнером;
2. Объекты различной формы, с которыми робот будет взаимодействовать;
3. Робот Kuka iiwa, с помощью которого нейросетевой агент будет поднимать объекты.

Главная цель – научить робота корректно захватывать и поднимать объекты. Агенту необходимо принимать решение между 3 действиями – по одному действию для перемещения по каждой из осей x и y, а также одно действие для угла поворота захвата. При каждом шаге в среде манипулятор автоматически опускается. Таким образом агенту необходимо захватить и

поднять любой объект из корзины. В данной задаче награда является бинарной, и выдается, только если один из объектов находится выше заданной высоты к концу эпизода. Входными данными, предоставляемыми агенту, являются трехканальные изображения состояния среды размером (48, 48, 3), пример такого изображения приведен на рисунке 4.2. Благодаря редкой двоичной награде, графическим входным данным и сложности задачи в целом, среда становится достаточно интересной для исследования.

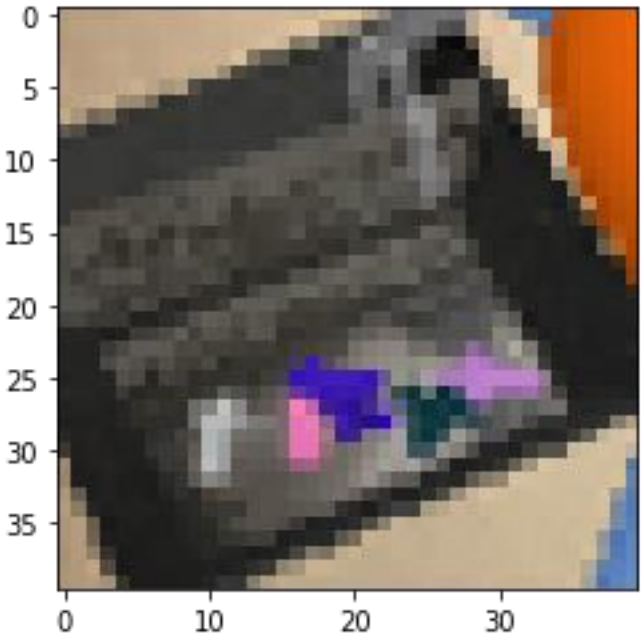


Рисунок 4.2 – Пример изображения с виртуальной камеры робота

Среда PyBullet KukaDiverseObjectEnv обладает следующими преимуществами:

- реалистичность: Среда использует реалистичную модель робота Kuka iiwa и физического движка PyBullet. А входными данными служит изображение с виртуальной камеры.
- разнообразие: Среда содержит различные объекты, что позволяет роботу учиться манипулировать объектами разных форм, размеров и весов.
- сложность: Среда достаточно сложна, чтобы представлять интерес

для исследователей и разработчиков робототехники.

Поскольку высокая точность в подобных средах достигается за крайне продолжительное время, для сокращения времени обучения точкой ранней остановки для всех экспериментов будет являться достижение точности равной 50% за последние 100 эпизодов.

4.2 Реализация

4.2.1 Deep Q-Network (DQN)

Среда KukaDiverseObjectEnv симулирует задачи, в которых робот KUKA должен взаимодействовать с различными объектами. Она предоставляет агенту доступ к визуальной информации с виртуальной камеры робота. В данном алгоритме входное изображение будет преобразовано в grayscale.

Основной цикл обучения реализован в соответствии [27]. Для реализации метода DQN была построена UML-диаграмма классов, изображенная на рисунке 4.3.

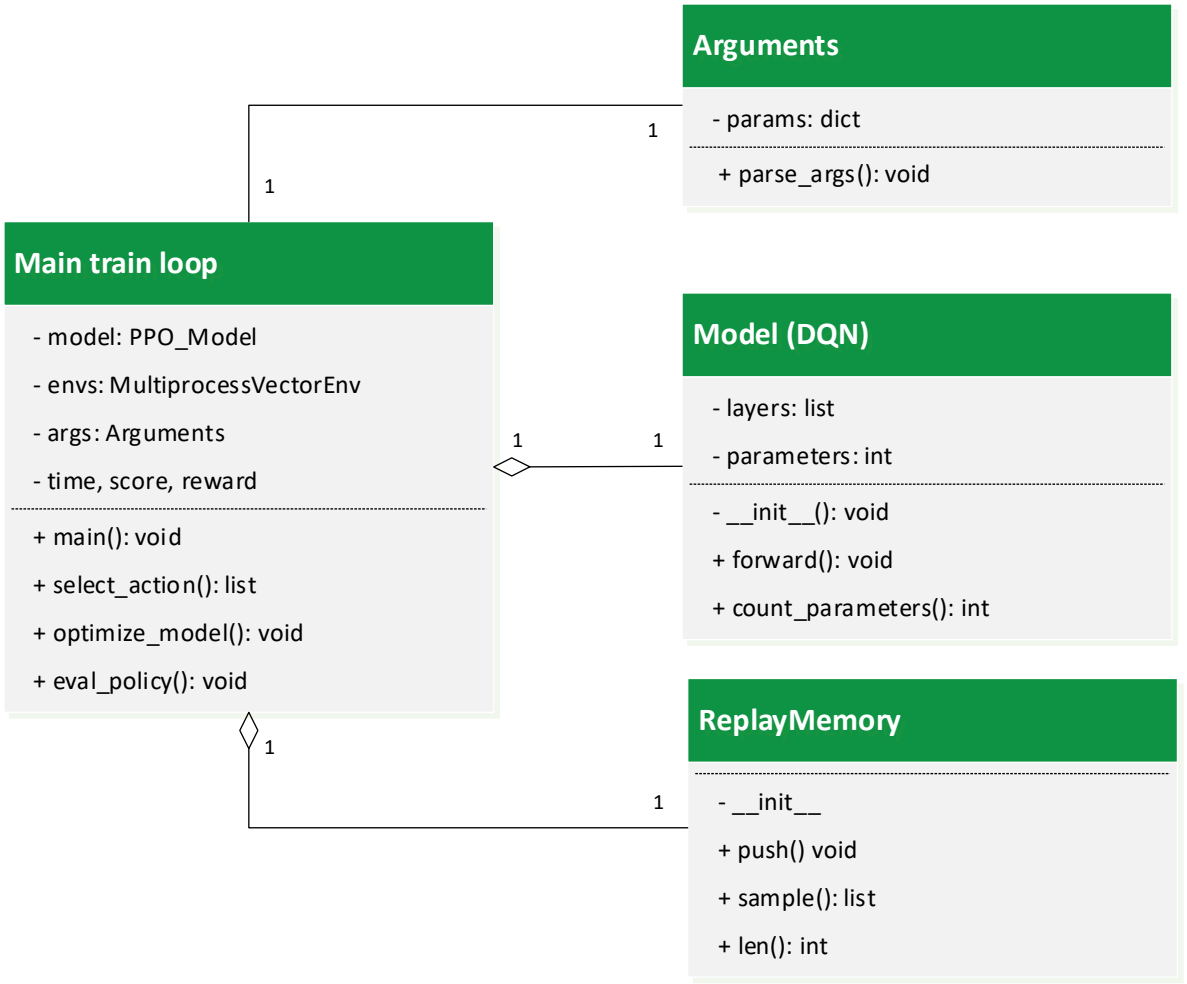


Рисунок 4.3 – UML-диаграмма классов DQN

Нейронная сеть DQN использует сверточные слои для обработки визуальных входных данных и полносвязные слои. Архитектура сети включает:

1. Входной слой, принимающий состояние среды.
2. Несколько сверточных слоев для извлечения признаков.
3. Несколько полносвязных слоев для обработки состояния и вычисления Q-значений для каждого действия.

Топология нейронной сети агента DQN представлена на рисунке 4.4

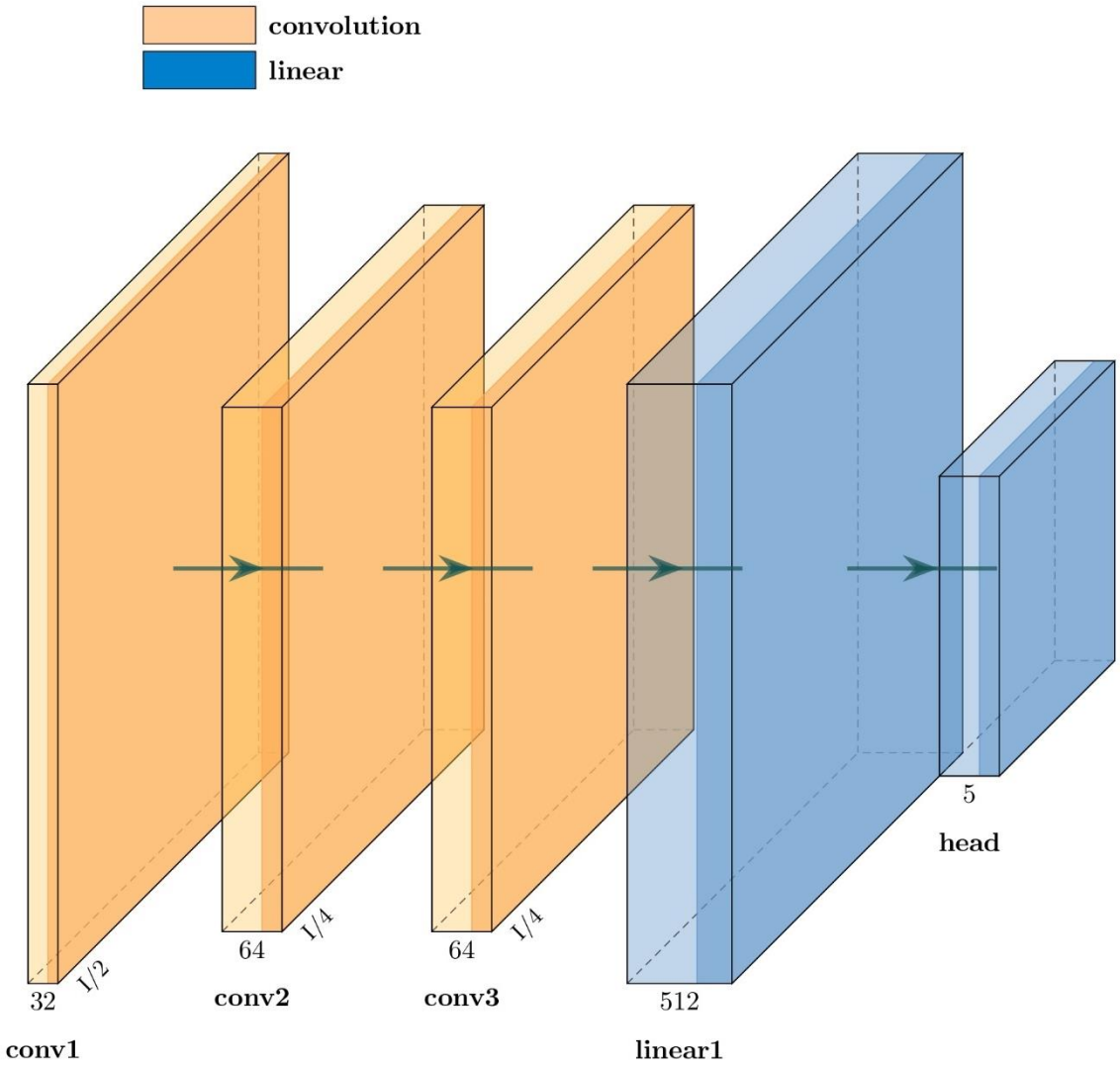


Рисунок 4.4 – Топология нейронной сети агента DQN

Гиперпараметры агента DQN представлены в таблице 4.2.

Таблица 4.1 – Гиперпараметры агента DQN

Параметр	Значение	Описание
Количество агентов	1	Количество агентов, обучаемых одновременно
Шаги за эпизод	20	Максимальное количество шагов за эпизод
Эпохи	10	Количество эпох обучения на одну выборку
Размер пакета	32	Размер пакета, выбираемый из сохраненных траекторий
Коэффициент дисконтирования (gamma)	0.99	Коэффициент дисконтирования
Скорость обучения	1e-4	Скорость обучения модели
Память воспроизведения	10000	Размер памяти для воспроизведения повторов
Оптимизатор	Adam	Метод оптимизации

В результате обучения агента DQN был получен график средней награды по циклам обучения агента изображенный на рисунке 4.5.

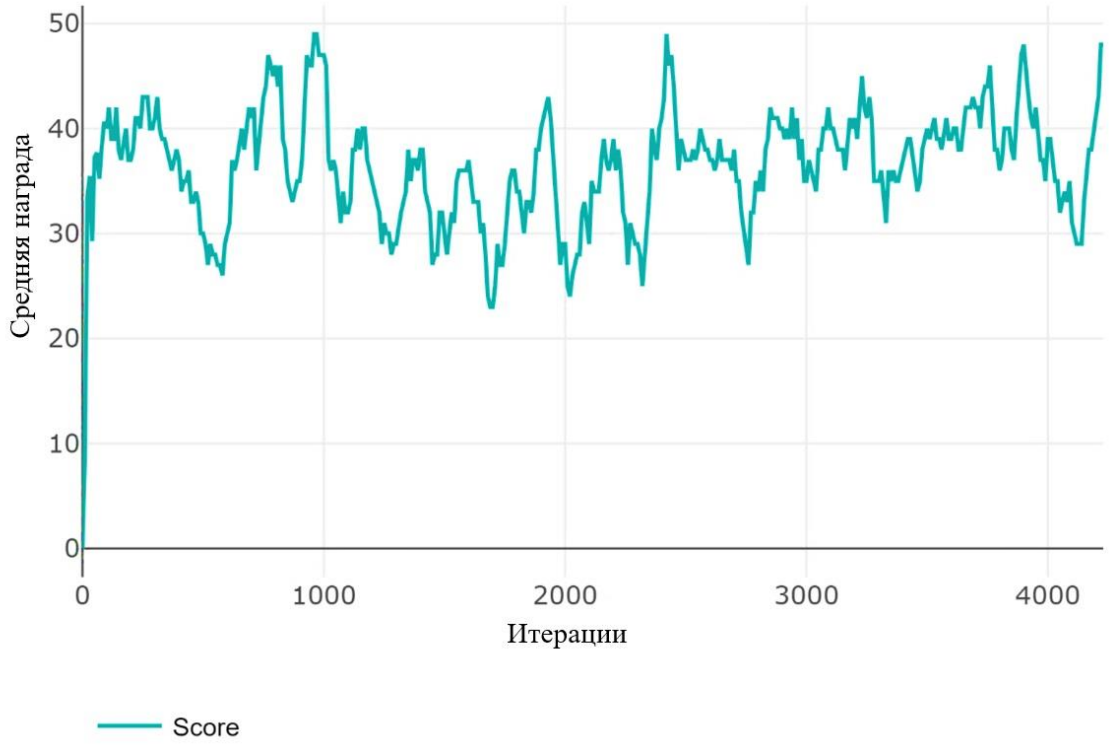


Рисунок 4.5 – График средней награды агента DQN

Таким образом был реализован метод DQN, который в последствии будет тестироваться и исследоваться.

4.2.2 Proximal Policy Optimization (PPO)

В алгоритме PPO используется входное изображение в формате RGB. Основной цикл обучения реализован в соответствии с [30]. Для реализации метода PPO была построена UML-диаграмма классов, изображенная на рисунке 4.6.

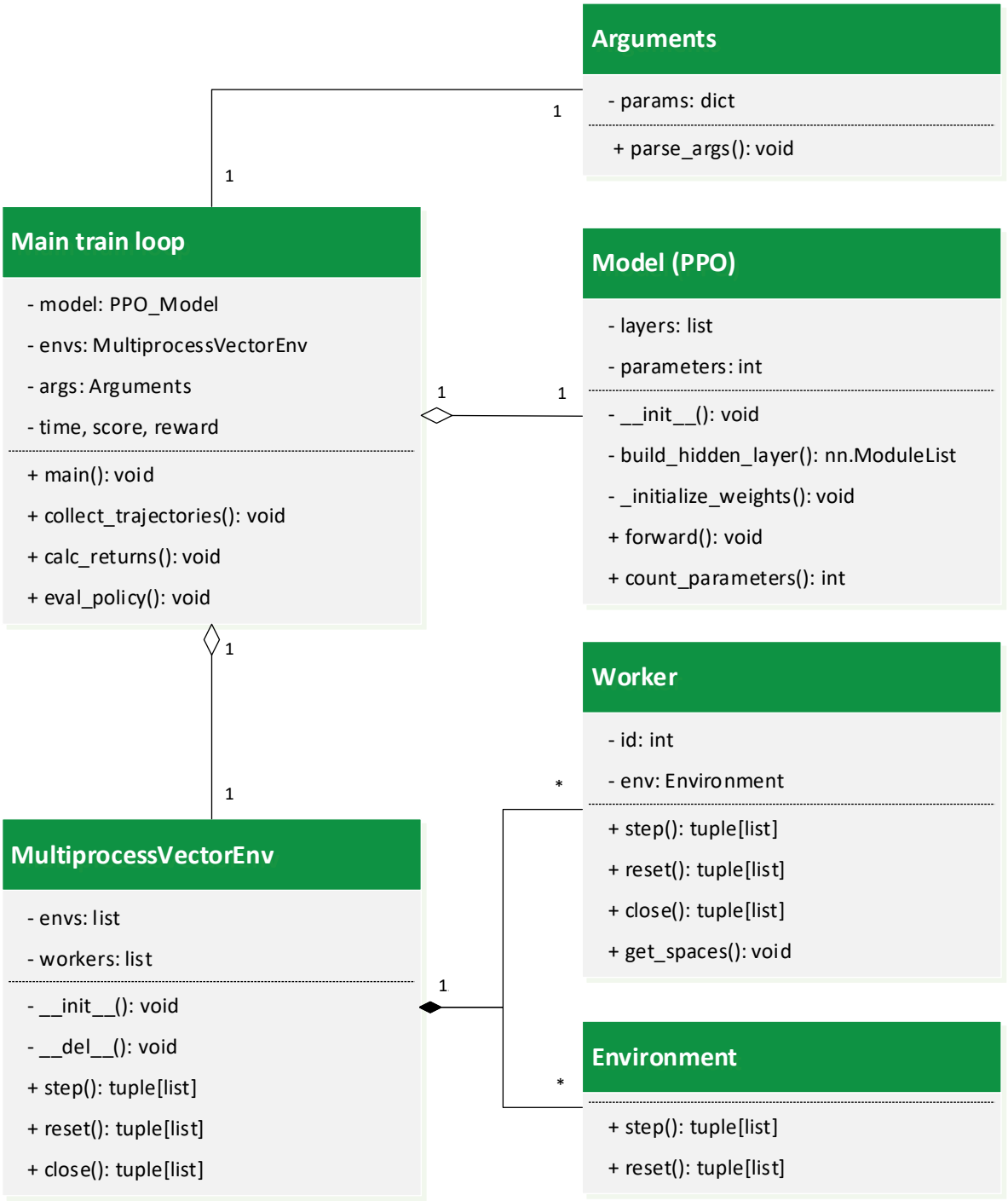


Рисунок 4.6 – UML-диаграмма классов PPO

Нейронная сеть DQN использует сверточные слои для обработки визуальных входных данных и полносвязные слои. Архитектура сети включает:

1. Входной слой, принимающий состояние среды.
2. Несколько сверточных слоев для извлечения признаков.
3. Несколько полносвязных слоев для обработки состояния.

4. Слои actor-critic. Это включает две нейронные сети:
   – Actor (политика): генерирует вероятности выбора действий, основываясь на текущем состоянии.
   – Critic (ценностная функция): оценивает ценность состояния, что используется для вычисления значений преимуществ (advantages).

Топология нейронной сети агента PPO представлена на рисунке 4.7.

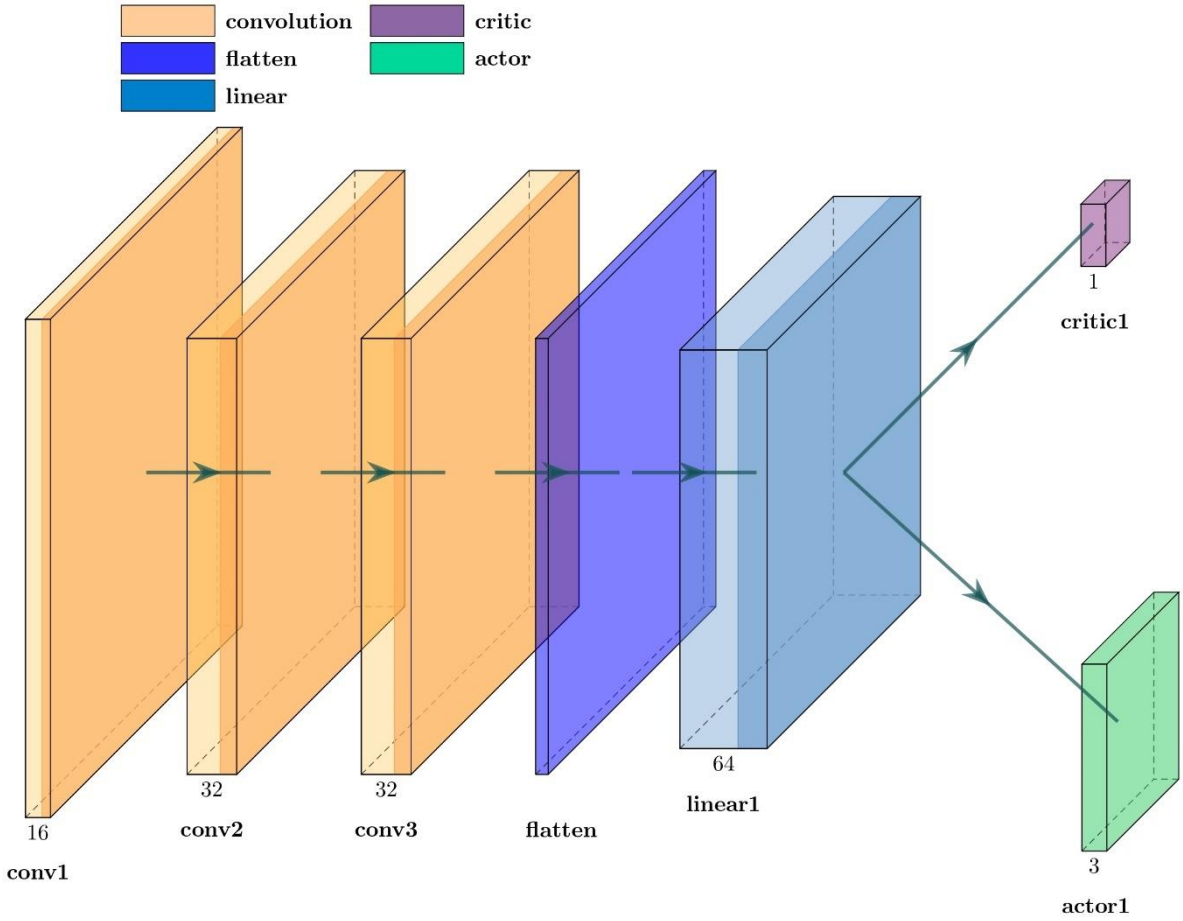


Рисунок 4.7 – Топология нейронной сети агента PPO

Для оптимизации и обновления сети на основе функции потерь используется метод SGD или если быть точнее адаптивный его вариант Adam. Обновления производятся на мини-пакетах данных, выбранных из общего пакета данных.

Гиперпараметры агента PPO представлены в таблице 4.3.

Таблица 4.2 – Гиперпараметры агента PPO

Параметр	Значение	Описание
Количество агентов	1	Количество агентов, обучаемых одновременно
Шаги за эпизод	1024	Максимальное количество шагов за сезон
Эпохи	10	Количество эпох обучения на одну выборку
Режим выбора пакета	Shuffle	Режим выборки за сезон
Размер пакета	128	Размер выборки, взятой из накопленных траекторий
Коэффициент дисконтирования (gamma)	0.993	Коэффициент дисконтирования
Ограничение градиента	10.0	Максимальная норма градиента
Уменьшение	False	Уменьшение значений Epsilon, Beta и скорости обучения
Epsilon	0.07	Коэффициент, используемый для клипирования r = new_probs/old_probs во время обучения
Beta	0.01	Коэффициент энтропии
Метод оценивания преимуществ	GAE	Тип оценивания преимущества
Tau	0.95	Коэффициент Tau в GAE
Нормализация	False	Нормализация преимуществ
Скорость обучения	2e-4	Скорость обучения
Оптимизатор	Adam	Метод оптимизации
Активатор	Tahn	Метод активации

В результате обучения агента PPO был получен график средней награды по циклам обучения изображенный на рисунке 4.8.

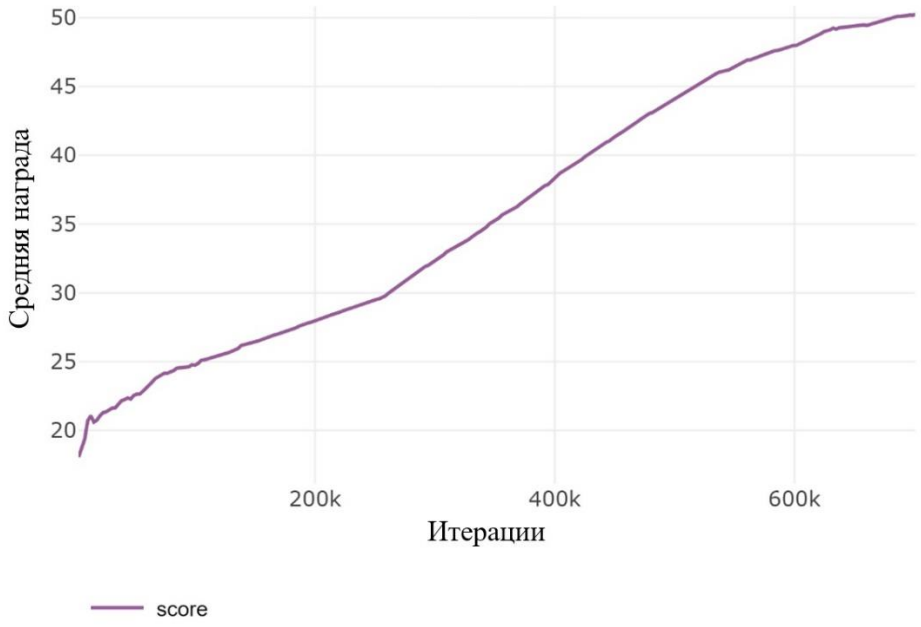


Рисунок 4.8 – График средней награды агента PPO

Таким образом был реализован метод PPO, который в последствии будет тестироваться и исследоваться.

4.2.3 Улучшения агента PPO

4.2.3.1 PPOv1

Алгоритм PPOv1 представляет собой первое улучшение ранее описанного PPO. В данном варианте модель специально адаптирована для взаимодействия со средой KukaDiverseObjectEnv.

Помимо структурных изменений в обучающем цикле, произведены изменения в пространствах наблюдений и действий. Входные данные представлены изображениями размером (84, 84, 3), а количество доступных действий увеличено до пяти: по два действия для перемещения по каждой из осей x и y, а также одно действие для угла поворота захвата.

Изменения также затронули архитектуру нейронной сети.

– был добавлен дополнительный второй линейный слой перед слоями актора-критика;

– в трех сверточных слоях изменены размеры ядра и шаг. Изначальные размеры ядра составляли 8, 4 и 3, в то время как новые значения ядра были изменены на 5 для всех слоев. Аналогично, шаг был изменен с первоначальных значений 4, 2 и 1 на новые значения 2 для всех слоев.
– применена пакетная нормализация (BatchNorm2d) после каждого сверточного слоя.
– в качестве метода инициализации весов нейронной сети использован метод равномерного распределения Ксавье (xavier-uniform).

Данные изменения должны стабилизировать и ускорить обучение агента.

Топология нейронной сети агента PPOv1 представлена на рисунке 4.9

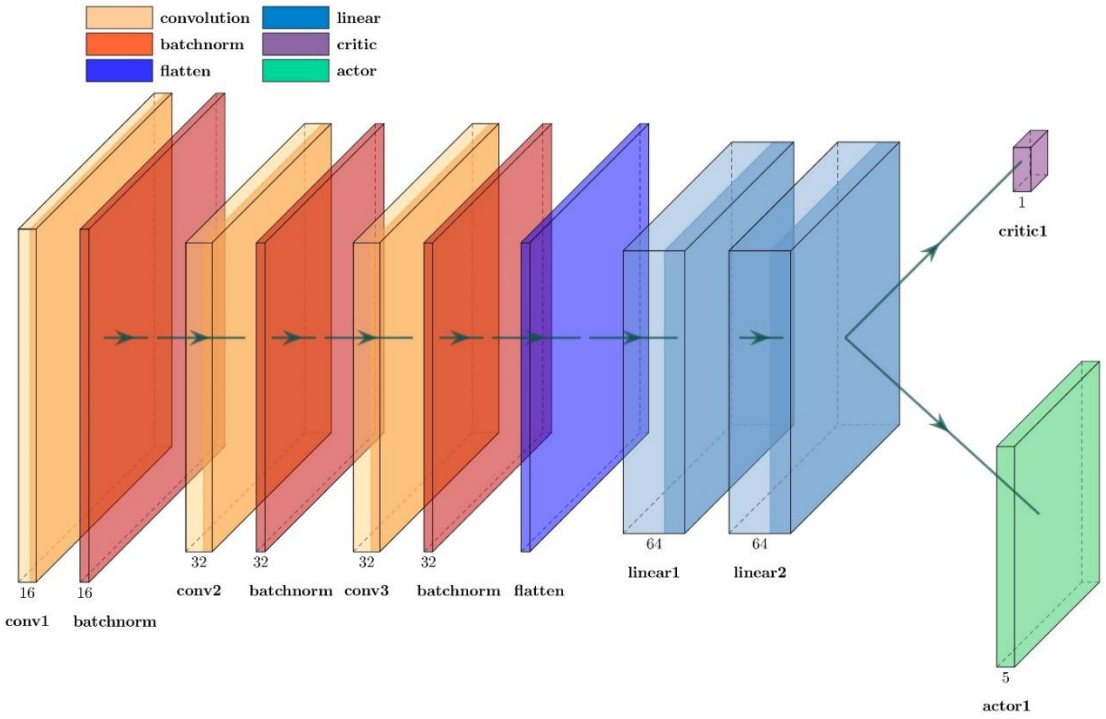


Рисунок 4.9 – Топология нейронной сети агента PPOv1

Таким образом был реализован метод PPOv1, который в последствии будет тестироваться и исследоваться.

4.2.3.2 PPOv2

Алгоритм PPOv2, являющийся второй модификацией PPO, вносит дополнительные изменения к PPOv1. В данном варианте используется синхронное параллельное выполнение действий в среде. Этот подход существенно ускоряет процесс сбора траекторий действий в среде и, следовательно, улучшает производительность обучения агента в целом. Реализация параллельных сред выполнена с использованием библиотеки multiprocessing [34], при этом число параллельных сред составляет 20, что соответствует количеству потоков процессора на рабочей станции.

4.2.3.3 PPOv3

Алгоритм PPOv3 представляет собой третью модификацию PPO, продолжающую развитие PPOv2. В этой версии алгоритма были оптимизированы некоторые гиперпараметры благодаря использованию модульной системы обучающего цикла. В частности, главными изменениями стали уменьшение размера пакета данных и увеличение скорости обучения (Learning rate), что позволило сократить длительной эпизодов обучения и ускорить процесс обновления весов модели, что также ускорило обучение. Важным изменением является увеличение глубины слоев с большим количеством нейронов для сверточные сети, включая общие слои (shared layers) и слои actor-critic.

Подобранные гиперпараметры агента PPOv3 приведены в таблице 4.4

Таблица 4.3 – Гиперпараметры агента PPOv3

Параметр	Значение	Описание
Количество агентов	20	Количество агентов, обучаемых одновременно
Шаги за эпизод	1024	Максимальное количество шагов за сезон
Эпохи	10	Количество эпох обучения на одну выборку
Режим выбора пакета	Shuffle	Режим выборки за сезон
Размер пакета	32	Размер выборки, взятой из накопленных траекторий

Параметр	Значение	Описание
Коэффициент дисконтирования (gamma)	0.99	Коэффициент дисконтирования
Ограничение градиента	10.0	Максимальная норма градиента
Уменьшение	True	Уменьшение значений Epsilon, Beta и скорости обучения
Epsilon	0.07	Коэффициент, используемый для клипирования r = new_probs/old_probs во время обучения
Beta	0.01	Коэффициент энтропии
Метод оценивания преимуществ	GAE	Тип оценивания преимущества
Tau	0.95	Коэффициент Tau в GAE
Нормализация	True	Нормализация преимуществ
Скорость обучения	3e-4	Скорость обучения
Оптимизатор	Adam	Метод оптимизации
Активатор	Tahn	Метод активации

Топология нейронной сети агента PPOv3 представлена на рисунке 4.10

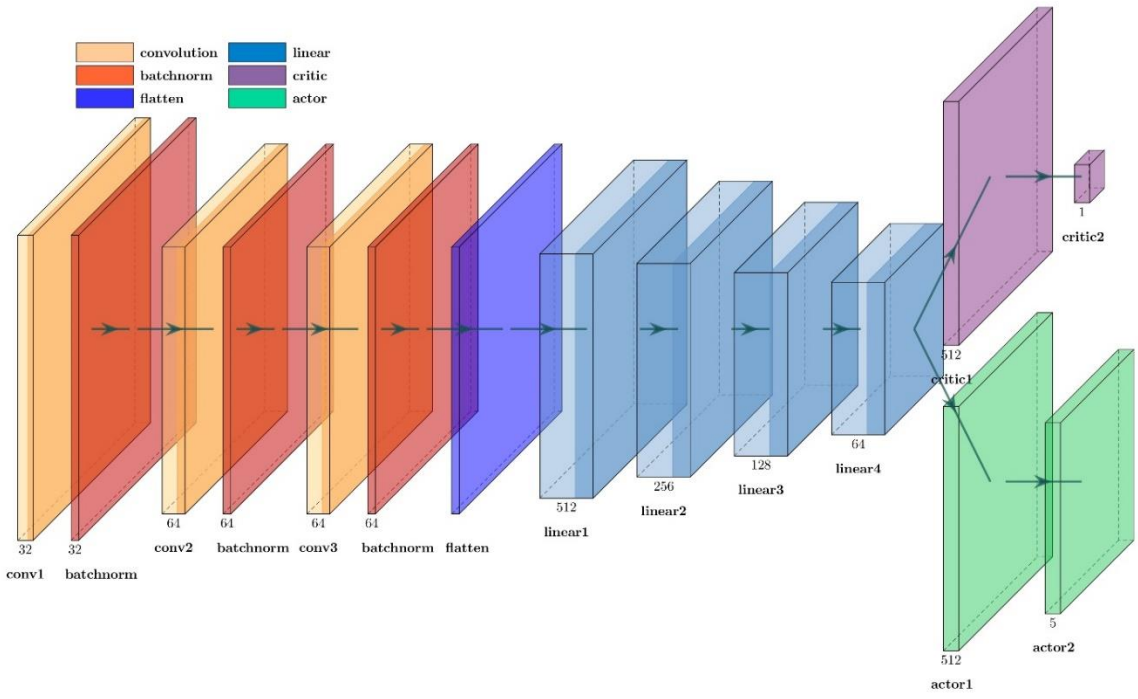


Рисунок 4.10 – Топология нейронной сети агента PPOv3

Таким образом был реализован метод PPOv2, который в последствии будет тестироваться и исследоваться.

4.2.3.4 PPOv4

Алгоритм PPOv4 представляет собой четвертую модификацию PPO, продолжающую развитие PPOv3. Нововведением данной версии стало добавление механизма внимания, который был представлен в работе [14]. Механизм внимания позволяет модели фокусироваться на различных частях входных данных, взвешивая их значимость для текущей задачи. В контексте робота-манипулятора это означает, что модель может выделять важные признаки объектов и их окружения, что критически важно для точного выполнения манипуляций. Топология нейронной сети агента PPOv4 представлена на рисунке 4.11

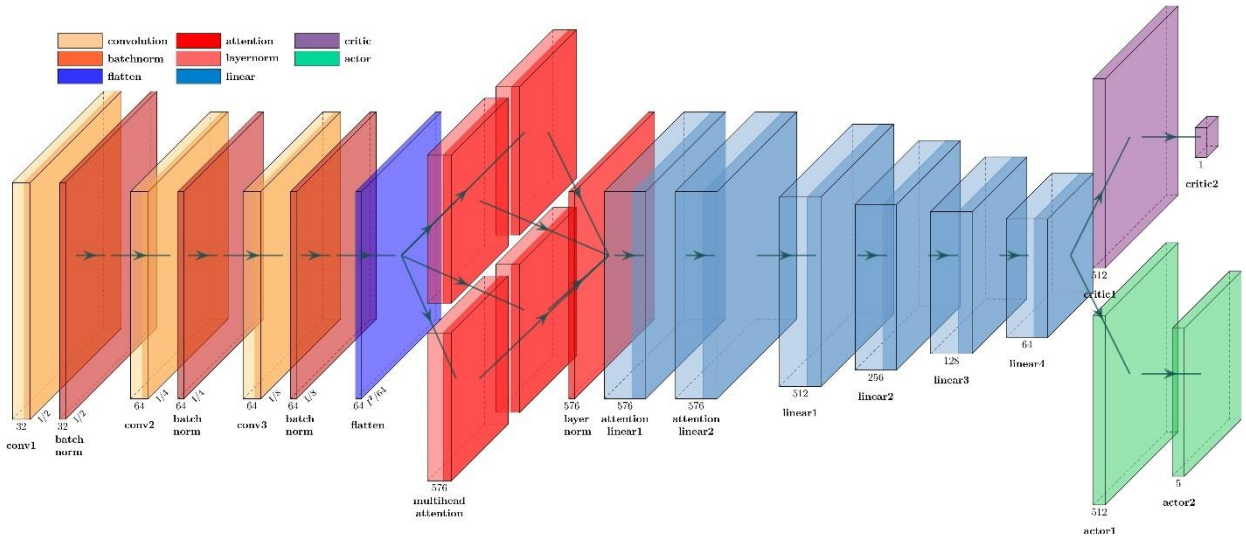


Рисунок 4.11 – Топология нейронной сети агента PPOv4

Сначала сверточные слои обрабатывают изображения с камеры робота, извлекая пространственные признаки объектов и их относительного расположения. Результат обработки представляет собой многомерные тензоры, содержащие информацию о высокоуровневых признаках. Затем эти тензоры подаются в слои внимания. Механизм внимания вычисляет весовые коэффициенты для каждого признака, определяя их относительную важность. В результате, внимание распределяется между различными частями

изображения, например, на манипуляторе робота и объектах, которые нужно захватить.

Наглядная демонстрация графиков обучения агентов с использованием механизма внимания и без представлена на рисунке 4.12.

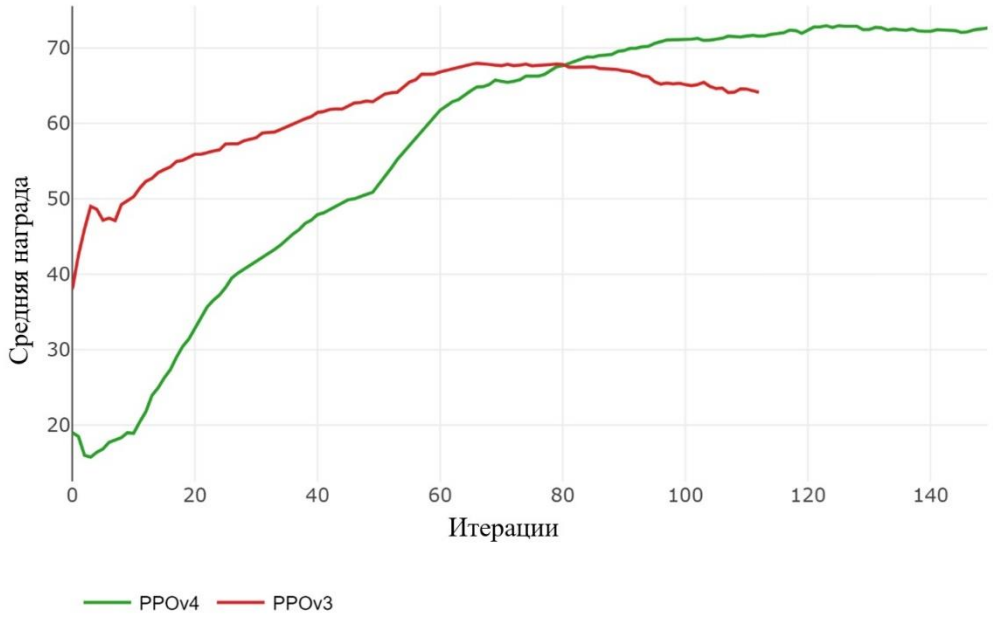


Рисунок 4.12 – Графики средней награды агентов PPO с использование механизма внимания и без

В ходе тестирования было установлено, что добавление механизма внимания после сверточных слоев способствует стабилизации процесса обучения. Это обусловлено снижением переобучения, поскольку модель меньше полагается на случайные шумы и незначительные детали входных данных. Вместо этого внимание сосредотачивается на ключевых аспектах задачи. Несмотря на улучшение стабильности и предсказуемости обучения, механизм внимания не оказывает значимого влияния на конечную точность модели. Более того, на начальных этапах обучение модели начинается с существенно более низких значений награды, однако скорость обучения увеличивается и приводит к тому, что со временем их производительность выравнивается.

5 РЕЗУЛЬТАТЫ РАБОТЫ

5.1 Тестирование агентов

Для тестирования всех полученных агентов использовалось по 1000 эпизодов среды в тестовом режиме. Результаты тестирования приведены в таблице 5.1.

Таблица 5.1 – Результаты работы агентов со средой в тестовом режиме

Агент	DQN	PPO	PPOv1	PPOv2	PPOv3	PPOv4
True episode	47,8%	49,7%	50,2%	49,1%	52,4%	48,9%
False episode	52,2%	50,3%	49,8%	50,9%	47,6%	51,1%

Как видно из таблицы 5.1 все агенты достигают поставленной цели и имеют удовлетворительную точность.

5.2 Сравнительный анализ обученных агентов

Полученные результаты обучения агентов DQN, PPO и его модификаций представлены на рисунках 4.4, 4.6, 4.10 и 5.1 и в таблице 5.2

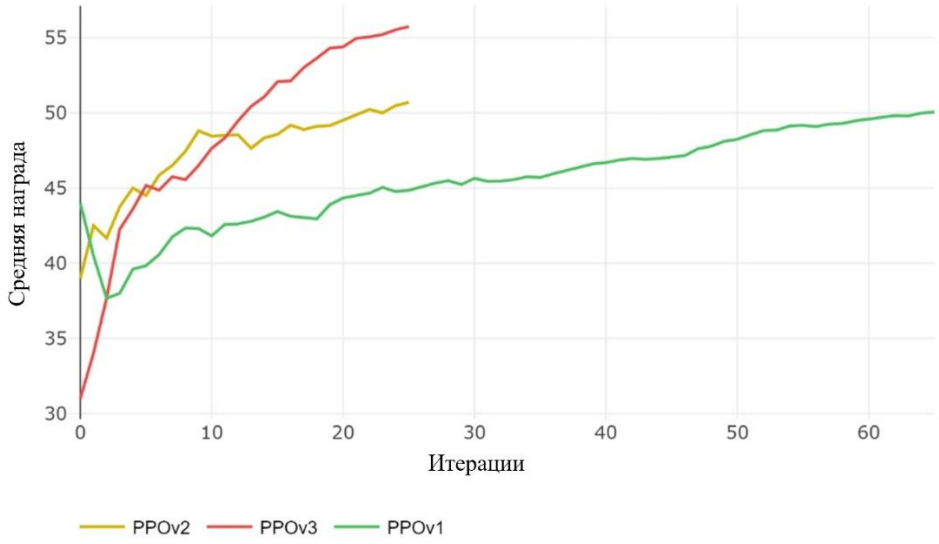


Рисунок 5.1 – Графики средней награды агентов PPOv1, PPOv2, PPOv3

Таблица 5.2 – Сравнение метрик обучения всех полученных агентов

Агент	Средняя награда	Шаги обучения	Время обучения
DQN	51,000	4239	2:06:54,878
PPO	50,202	696320	5:12:04,500
PPOv1	50,690	65 (66560)	1:52:07,842
PPOv2	50,061	26 (26624)	0:35:12,836
PPOv3	50,067	13 (13312)	0:09:34,629
PPOv4	50,229	47 (48128)	0:32:01,375

Исходя из рисунков 4.4, 4.6, 4.10, 5.1 и таблиц 5.1 и 5.2 можно сказать, что:

- Все агенты справились с поставленной задачей и достигают точности в 50 процентов как при обучении, так и при тестировании;
- Агент DQN затрачивает примерно в 2.5 раза меньше времени на обучение по сравнению с агентом PPO. Вероятно, это вызвано большой вычислительной сложностью алгоритма PPO;
- Благодаря улучшениям в PPOv1 удалось сравнять время обучения с алгоритмом DQN;
- Параллелизация сред значительно ускорила обучения нейросети и сократила время обучения практически в 4 раза;
- Оптимизация гиперпараметров является критичной для алгоритма PPO. Их подбор в PPOv3 стал значительным вкладом в эффективность и скорость обучения модели. Глубина нейросети также сыграла большую роль;
- Добавление механизма внимания в данном контексте значительно замедляет процесс обучения. Но если не ограничиваться наградой в 50 процентов модель в целом сравнивается по метрике награды с описанными конкурентами.

Лучшем по эффективности и скорости обучения можно считать алгоритм PPO, а конкретно его версию PPOv3. Но стоит отметить, хотя PPOv3

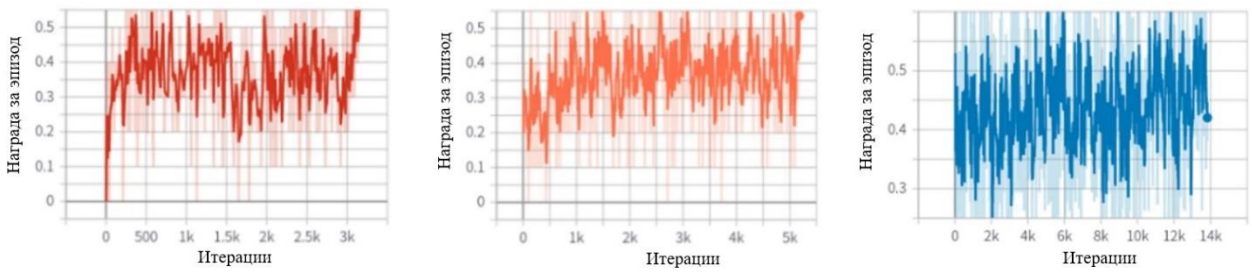
обучается быстрее, PPOv4 обладает большей стабильностью и потенциалом к дальнейшему обучению и улучшению.

5.3 Сравнение с аналогичными решениями

В сравнении с аналогичными решениями будут участвовать агент DQN и лучшие агенты PPOv3 и PPOv4. Сравнение будет проводится с результатами, полученными в трех работах:

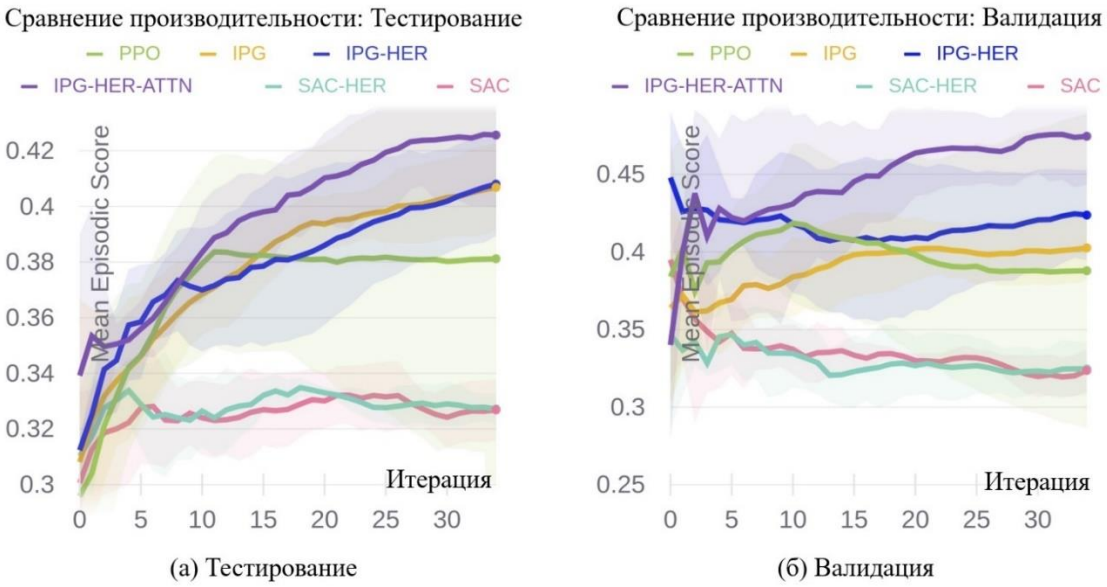
- В работе [23] исследовалась точность алгоритмов DQN, PPO и параллельной его реализации «PPO parallel». Обучение агентов проводилось в среде KukaDiverseObjectEnv до достижения 50% точности;
- В работе [35] исследовалась точность алгоритмов PPO, IPG, IPG-HER, IPG-HER-ATTN, SAC, SAC-HER. В данном сравнении алгоритмы SAC, SAC-HER не использовались поскольку не показали удовлетворительных результатов. Обучение агентов проводилось в среде KukaDiverseObjectEnv до достижения 30 эпизодов обучения, что соответствует 30720 шагам в среде;
- В работе [5] исследовалась точность алгоритмов Regression, Regression-PoP, Physics-only, Residual-physics для задач захвата и броска объекта как в виртуальной среде, так и в реальном мире. В данном сравнении используются полученные метрики точности только для захвата объекта в виртуальной среде, поскольку другие задачи и среды не соответствуют тематике текущей работы. Обучение агентов проводилось в специализированной среде для задачи TossingBot до достижения максимально высокой точности.

На рисунке 5.2-5.4 и таблицах 5.3-5.6 представлены результаты обучения агентов, приведенные в выбранных для сравнения работах.



(а) DQN: 51.00 (б) PPO: 50.08 (в) PPO parallel: 50.13

Рисунок 5.2 – Графики средней награды агентов в работе [23]



(а) Тестирование (б) Валидация

Рисунок 5.3 – Графики средней награды агентов в работе [35]

Method	Balls	Cubes	Rods	Hammers	Seen	Unseen
Regression	99.4	99.2	89.0	87.8	95.6	69.4
Regression-PoP	99.2	98.0	89.8	87.0	96.4	70.6
Physics-only	99.4	99.2	87.6	85.2	96.6	64.0
Residual-physics	98.8	99.2	89.2	84.8	96.0	74.6

Рисунок 5.4 – Точность захвата различных объектов манипулятором в работе [5]

Таблица 5.3 – Метрики, полученные в текущей работе

Агент	Средняя награда (%)	Максимальная награда (%)	Шаги обучения
DQN	51,000	51,000	4239
PPOv3	50,067	67,960	13 (13312)
PPOv4	50,229	73,420	47 (48128)

Таблица 5.4 – Метрики, полученные в работе [23]

Агент	Средняя награда (%)	Шаги обучения
DQN	51,000	~3300
PPO	50,080	~5200
PPO parallel	50,130	~13900

Таблица 5.5 – Метрики, полученные в работе [35]

Агент	Средняя награда (%)	Шаги обучения
IPG-HER-ATTN	~48,500	30 (30720)
IPG-HER	~42,800	30 (30720)
IPG	~40,100	30 (30720)
PPO	~42,600	10 (10240)

Таблица 5.6 – Метрики для задачи захвата, полученные в работе [5]

Агент	Средняя награда (%)			Шаги обучения
	Знакомый объект	Не знакомый объект	Среднее значение	
Regression	95,6	69,4	82,5	–
Regression-PoP	96,4	70,6	83,5	–
Physics-only	96,6	64,0	80,3	–
Residual-physics	96,0	74,6	85,3	–

Исходя из результатов приведенных на рисунках 5.2-5.4 и таблицах 5.3-5.6 можно сказать:

- Агент DQN из текущей работы затратил большее количество шагов в среде для своего обучения нежели аналогичный агент DQN из работы [23];
- Агент PPOv3 из текущей работы в целом схож с агентом PPO parallel из работы [23] и вероятно немного превосходит PPO из работы [35];
- Агент PPOv4 из текущей работы затрачивает большее количество шагов в среде, но вероятно более стабилен в обучении и не склонен к переобучению по сравнению со своими аналогами из работы [23, 35] поскольку в них не используется механизм внимания;

– Агент PPOv4 показал максимальную точность примерно на 10% хуже, чем алгоритмы, предложенные в работе [5]
– Алгоритм IPG из работы [35] был немного улучшен с помощью стратегии HER. В то же время механизм внимания использованный вместе с IPG-HER значительно ускорил рост точности модели, что не коррелирует с результатами полученными в данном исследование. Вероятно, это связано из-за особенностей алгоритмов PPO и IPG, а также других аспектов реализации;

В целом, при сравнении результатов, полученных в аналогичных решениях, можно сказать, что обученные в данной работе агенты справляются с поставленной задачей. Кроме этого, они имеют ряд преимуществ над некоторыми сравниваемыми решениями, такие как точность, стабильность.

ЗАДАНИЕ К РАЗДЕЛУ
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Обучающемуся:

Группа	ФИО
8ПМ2Л	Залогин Никита Евгеньевич

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	магистратура	Направление/ООП/ОПОП	09.04.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Энергетические ресурсы: электрическая энергия (4.78р/1кВт); информационные ресурсы: учебники по теме исследований; человеческие ресурсы: системный администратор (1 человек)*
2. *Нормы и нормативы расходования ресурсов*	*30% премии; 20% надбавки; 16% накладные расходы; 30% районный коэффициент*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Коэффициент отчислений во внебюджетные фонды – 30%*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Описание продукта. Анализ потенциальных сфер применения результатов НИР. Проведение SWOT-анализа. Оценка конкурентоспособности технических решений*
2. *Разработка устава научно-технического проекта*	*Инициация проекта: определение заинтересованных сторон проекта, целей, ожиданий, требований и результатов проекта*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*Описание этапов работ с учетом ограничений и ресурсной обеспеченности процессов. Анализ ограничений и рисков. Определение бюджета НТИ*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Проведение оценки экономической эффективности, ресурсоэффективности и сравнительной эффективности различных вариантов исполнения*
Перечень графического материала:	
1. *Матрица SWOT* 2. *Цели и результаты проекта* 3. *Перечень этапов работы и распределение исполнителей* 4. *Иерархическая структура работ* 5. *График проведения НТИ* 6. *Расчетная стоимость исследований* 7. *Оценка ресурсной, финансовой и экономической эффективности НТИ*	

Дата выдачи задания к разделу в соответствии с календарным учебным графиком	

Задание выдал консультант по разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (БШ НСП)	Аникина Екатерина Алексеевна	к.э.н., доцент		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ2Л	Залогин Никита Евгеньевич		

6 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

6.1 Предпроектный анализ

6.1.1 Потенциальные потребители результатов исследования

Целью данной работы является разработка метода на основе обучения с подкреплением (Reinforcement Learning, RL) для обучения интеллектуального нейросетевого агента управлению манипулятором.

Разработка заключается в реализации и улучшении алгоритмов обучения с подкреплением. Для обучения агентов используется специализированная виртуальная среда с роботом-манипулятором. Результатом работы агента является поднятый с помощью манипулятора объект в виртуальной среде.

Целевым рынком для данной разработки являются промышленные компании и научно-исследовательские лаборатории и центры, а также образовательные организации.

Сегментация рынка услуг проводится по степени развития и отрасли применения. Результат сегментации представлен в таблице 6.1.

Таблица 6.1 – Карта сегментации степени развития и отрасли применения

		Отрасль применения разработки		
		Научная	Промышленная	Образовательная
Размер компании	Большой			
	Средний			
	Малый			

	Фирма А		Фирма Б		Фирма В

Таким образом рассмотрена возможная сегментация рынка по использованию машинного обучения в задачах RL с роботом-манипулятором.

6.1.2 Анализ конкурентных технических решений с позиции ресурсоэффективности и ресурсосбережения

В наше время значительное внимание уделяется решению задач управления роботами-манипуляторами в виртуальных средах с использованием алгоритмов обучения с подкреплением. Современные работы сосредоточены на достижении высокой эффективности и адаптивности алгоритмов обучения с подкреплением, таких как Deep Q-Network (DQN) и Proximal Policy Optimization (PPO). Однако не всегда внимания уделяется оптимизации вычислительной эффективности этих алгоритмов. Стремление к повышению адаптивности и точности управления приводит к созданию алгоритмов с высокой вычислительной сложностью, что снижает их алгоритмическую эффективность. Примерами работ, направленных на повышение эффективности алгоритмов обучения с подкреплением в задачах управления роботами, являются исследования [5, 21, 22, 23, 24, 35]. Согласно исследованиям, наиболее используемыми и актуальными архитектурами являются:

1. DQN
2. PPO;
3. IPG;
4. Residual-physics.

В качестве набора технических критериев рассматриваются следующие требования:

1. Технические критерии оценки ресурсоэффективности:
    1.1. Точность: точность выполнения манипуляции;
    1.2. Вычислительная сложность: насколько требователен алгоритм к вычислительным ресурсам рабочей системы;

1.3. Универсальность: насколько универсален обученный агент;

1.4. Востребованность: насколько востребован обученный агент;

2. Экономические критерии оценки эффективности:

2.1. Стоимость разработки: насколько дорогой является разработка.

Экспертная оценка основных технических характеристик решений представлена в таблице 6.2.

Таблица 6.2 – Оценочная карта для сравнения конкурентных технических решений

Критерии оценки	Вес критерия, B_i	Баллы				Конкуренто-способность			
		$Б_1$	$Б_2$	$Б_3$	$Б_4$	$К_1$	$К_2$	$К_3$	$К_4$
Технические критерии оценки ресурсоэффективности									
1. Точность	0,35	3	4	3	5	1,05	1,4	1,05	1,75
2. Вычислительная сложность	0,25	5	4	5	2	1,25	1	1,25	0,5
3. Универсальность	0,1	2	4	2	4	0,2	0,4	0,2	0,4
4. Востребованность	0,1	4	5	4	5	0,4	0,5	0,4	0,5
Экономические критерии оценки эффективности									
1. Стоимость разработки	0,2	4	3	4	2	0,8	0,6	0,8	0,4
Итого	1	18	20	18	18	3,7	3,9	3,7	3,55

На основе проведенного анализа можно сделать вывод, что конкурентные решения для обучения агентов имеют различные преимущества. Некоторые имеют более высокую точность, но при этом и высокую вычислительную сложность, некоторые же наоборот более низкую точность и низкую вычислительную сложность. Кроме того, алгоритмы PPO обладают схожей универсальностью и востребованностью с более сложным в разработке Residual-physics. На основе этого можно сделать вывод, что алгоритм PPO является наиболее выгодным вариантом. Это подтверждает решение использовать PPO качестве основы для исследования.

6.1.3 SWOT – анализ

SWOT – Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) – представляет собой комплексный анализ научно-исследовательского проекта. SWOT–анализ применяют для исследования внешней и внутренней среды проекта.

SWOT – анализ состоит из трех этапов. В первом этапе мы анализируем сильные и слабые стороны исследовательской работы (внутренняя среда), а также возможности и угрозы (внешняя среда). Описание выполняется с помощью факторов, не имеющих количественной оценки. Начальная матрица SWOT-анализа приведена в таблице 6.3.

Таблица 6.3 – SWOT-анализ

	Strengths: S1. Соотношение точности и сложности обучения модели; S2. Низкие требования к вычислительным ресурсам; S3. Высокая скорость обучения модели; S4. Возможность дообучения модели.	Weaknesses: W1. Ограниченность виртуальной среды; W2. Вероятность ошибочного выполнения действия; W3. Сложность разработки; W4. Отсутствие финансирования.
Opportunities: O1. Совершенствование технологий и метода обучения модели; O2. Возможность точной настройки гиперпараметров; O3. Возможность перехода из виртуальной среды в реальную.		
Threats: T1. Отсутствие интереса к проекту; T2. Отсутствие рыночного спроса данный вид автоматизации; T3. Появление лучших аналогов.		

Второй этап состоит в выявлении соответствия сильных и слабых сторон научно-исследовательского проекта внешним условиям окружающей среды. Это соответствие или несоответствие должны помочь выявить степень необходимости проведения стратегических изменений.

В рамках данного этапа была построена интерактивную матрицу проекта приведенная в таблице 6.4. Ее использование помогает разобраться с различными комбинациями взаимосвязей областей матрицы SWOT. Каждый фактор помечен либо знаком «+» (означает сильное соответствие сильных сторон возможностям), либо знаком «-» (что означает слабое соответствие); «0» – если есть сомнения в том, что поставить «+» или «-».

Таблица 6.4 – Интерактивная матрица проекта

		Сильные стороны				Слабые стороны			
		S1	S2	S3	S4	W1	W2	W3	W4
Возможности	O1	+	+	+	+	-	-	0	-
	O2	+	+	+	+	-	-	0	-
	O3	+	+	+	+	-	-	0	-
Угрозы	T1	-	-	-	0	0	0	-	-
	T2	-	-	-	-	0	0	-	-
	T3	0	0	0	0	0	0	0	0

В третьем этапе составляется итоговая матрица SWOT – анализа исходя из полученной информации в таблице 6.4. Итоговая матрица SWOT приведена в таблице 6.5

Таблица 6.5 – SWOT-анализ

	Strengths: S1. Соотношение точности и сложности обучения модели; S2. Низкие требования к вычислительным ресурсам; S3. Высокая скорость обучения модели; S4. Возможность дообучения модели.	Weaknesses: W1. Ограниченность виртуальной среды; W2. Вероятность ошибочного выполнения действия; W3. Сложность разработки; W4. Отсутствие финансирования.
Opportunities: O1. Совершенствование технологий и метода обучения модели; O2. Возможность точной настройки гиперпараметров; O3. Возможность перехода из виртуальной среды в реальную.	O1O2O3S1S2S3S4. Все сильные стороны крайне благотворно влияют на возможности дальнейшего развития проекта.	O1O2O3W1W2W4. Несмотря на возможности, ограниченность виртуальной среды, вероятность ошибок и отсутствие финансирования остаются проблемами, которые требуют решения для полного использования возможностей.
Threats: T1. Отсутствие интереса к проекту; T2. Отсутствие рыночного спроса данный вид автоматизации; T3. Появление лучших аналогов.	T1T2S1S2S3. Сильные стороны проекта, такие как точность, низкие требования к ресурсам и высокая скорость обучения, могут не преодолеть отсутствие интереса и рыночного спроса; T2S4. Возможность дообучения модели может помочь преодолеть отсутствие рыночного спроса, но это требует стратегического продвижения и адаптации проекта.	T1T2W3W4. Сложность разработки и отсутствие финансирования, вместе с отсутствием интереса и рыночного спроса, представляют серьезные угрозы для проекта, требующие активного управления и поиска ресурсов.

Основываясь на результатах SWOT-анализа, можно заключить, что проект демонстрирует значительный потенциал благодаря своим преимуществам, которые хорошо сочетаются с возможностями. Однако он сталкивается с многими ограничениями и угрозами. Для успешной реализации необходимо сосредоточиться на использовании сильных сторон и возможностей, привлекая дополнительные ресурсы и стратегически продвигая проект для преодоления слабых сторон и внешних угроз.

6.1.4 Оценка готовности разработки к коммерциализации

Для исследовательского проекта важно оценить его готовность к коммерциализации, а также уровень знаний, необходимых для его реализации, что поможет сделать вывод о готовности исследовательского проекта к коммерциализации, а также о необходимости каких-либо усовершенствований. Перечень вопросов и оценка показателей степени проработки проекта с точки зрения как коммерциализации, так и научной составляющей, и компетенций разработчика представлены в таблице 6.6.

Таблица 6.6 – Оценка степени готовности научного проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1	Определен имеющийся научно-технический задел	4	4
2	Определены перспективные направления коммерциализации научно-технического задела	3	4
3	Определены отрасли и технологии (товары, услуги) для предложения на рынке	3	3
4	Определена товарная форма научно- технического задела для представления на рынок	2	2
5	Определены авторы и осуществлена охрана их прав	3	4
6	Проведена оценка стоимости интеллектуальной собственности	4	4
7	Проведены маркетинговые исследования рынков сбыта	2	3
8	Разработан бизнес-план коммерциализации научной разработки	3	3
9	Определены пути продвижения научной разработки на рынок	3	3
10	Разработана стратегия (форма) реализации научной разработки	4	4
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	2	2
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	2	2
13	Проработаны вопросы финансирования коммерциализации научной разработки	2	3
14	Имеется команда для коммерциализации научной разработки	1	2
15	Проработан механизм реализации научного проекта	4	4
	Итого	42	47

Значения оценки степени проработанности и знаний говорит о том, что перспективы развития и знания разработчика немного выше среднего. Для увеличения показателей необходимо проконсультироваться с экспертами в сферах, затрагиваемых разработкой, сформировать команду и найти финансирование.

6.1.5 Методы коммерциализации результатов научно-технического исследования.

При коммерциализации научно-технических разработок (НТ разработок) продавец (владелец интеллектуальной собственности) преследует цель, зависящую от дальнейшего использования полученного коммерческого эффекта. Целями могут быть получение средств для НИОКР (финансирование, оборудование, материалы, другие НТ разработки), одноразовая выгода (финансирование или накопление) или постоянный доход (поток финансовых средств).

Выбор метода коммерциализации напрямую влияет на срок вывода товара на рынок. Целью данного раздела является выбор метода коммерциализации объекта исследования и обоснование его целесообразности.

В данной работе возможны следующие методы коммерциализации научных разработок:

1. Торговля патентными лицензиями, т.е. передача третьим лицам права использования объектов интеллектуальной собственности на лицензионной основе. При этом в патентном законодательстве выделяющие виды лицензий: исключительные (простые), исключительные, полные лицензии, сублицензии, опционы.
2. Инжиниринг как самостоятельный вид коммерческих операций предполагает предоставление на основе договора инжиниринга одной стороной, именуемой консультантом, другой стороне,

именуемой заказчиком, комплекса или отельных видов инженерно-технических услуг, связанных с проектированием, строительством и вводом объекта в эксплуатацию, с разработкой новых технологических процессов на предприятии заказчика, усовершенствованием имеющихся производственных процессов вплоть до внедрения изделия в производство и даже сбыта продукции.

3. Передача интеллектуальной собственности в уставной капитал предприятия.

В рамках данной работы наиболее целесообразным методом коммерциализации результатов разработки является инжиниринг. Применение ресурсов и инфраструктуры заказчика в процессе разработки технологического решения позволяет оптимизировать его временные и финансовые затраты. Вовлеченность заказчика с самого начала минимизирует риски, связанные с несоответствием решения его потребностям и ожиданиям. Совместная работа над проектом способствует более быстрому внедрению разработанного решения в производственный процесс. Глубокое понимание заказчиком специфики задачи и особенностей рынка, в котором он оперирует, увеличивает шансы на успешную коммерциализацию разработки.

Инжиниринг может стать отправной точкой для долгосрочного и взаимовыгодного сотрудничества между разработчиком и заказчиком, что открывает возможности для реализации новых совместных проектов.

6.2 Инициация проекта

6.2.1 Цели и результаты проекта

В этом разделе представлена информация о заинтересованных сторонах проекта, иерархии целей проекта и критериях достижения целей.

Заинтересованные стороны проекта – это лица или организации,

которые активно вовлечены в проект или чьи интересы могут быть затронуты положительно или отрицательно в результате выполнения или завершения проекта. Информация о заинтересованных сторонах проекта представлена в таблице 6.7.

Таблица 6.7 – Заинтересованные стороны проекта

Заинтересованные стороны проекта	Ожидания заинтересованных сторон
Разработчик	Получение опыта, получение заработной платы.
Научный руководитель	Завершение выпускной квалификационной работы.
Компания-пользователь, научно-исследовательская лаборатория	Получение результатов обучения виртуальных агентов, разработка новых моделей RL, разработка систем под управлением интеллектуальных агентов.
Промышленное предприятие	Получение интеллектуальных агентов способных точно выполнять различные манипуляции для автоматизации производственных процессов
Образовательная организация	Получение материалов для формирования учебных процессов

Цель и результаты проекта представлены в таблице 6.8.

Таблица 6.8 – Цели и результат проекта

Цели проекта:	1. Исследовать различные алгоритмы RL; 2. Составить обзор актуальных статей и исследований; 3. Выбрать платформу для создания виртуальных сред; 4. Реализация, изучение и улучшение выбранных алгоритмов RL; 5. Проведение сравнительного анализа и представление результатов исследования, выводов и предложений.
Ожидаемые результаты проекта	1. Были реализованы и обучены модели RL для управления роботом-манипулятором; 2. Выпускная квалификационная работа завершена.
Критерии приёмки результата проекта:	1. Успешное тестирование агента в среде в тестовом режиме; 2. Удовлетворительные значения метрик точности манипулирования.
Требования к результату проекта:	1. Точность выполнения задачи среды должна быть приближена к 50% в среде в тестовом режиме; 2. Скорость обучения модели должна быть быстрее чем у стандартных вариантов реализации выбранных алгоритмов.

6.2.2 Организационная структура проекта

Состав рабочей группы данного проекта, роль каждого участника в данном проекте, а также функции, выполняемые каждым из участников и их трудозатраты в проекте представлены в таблице 6.9.

Таблица 6.9 – Рабочая группа проекта

№ п/п	ФИО, основное место работы, должность	Роль в проекте	Функции	Трудо-затраты, час.
1	Спицын Владимир Григорьевич, д.т.н., Профессор, ОИТ, ИШИТР.	Руководитель проекта	Постановка глобальных целей и задач; Проведение консультаций по целям и задачам выпускной квалификационной работы; Оценка эффективности полученных результатов	72
2	Григорьев Дмитрий Сергеевич, ст. преподаватель, ОИТ, ИШИТР.	Консультант проекта	Постановка локальных целей и задач; Разработка и мониторинг календарного плана; Оценка результатов технологической практики; Консультации по необходимым изменениям в выпускной квалификационной работе.	116
3	Аникина Екатерина Алексеевна, к.э.н., Доцент, БШ, НСП.	Эксперт проекта	Консультирование по разделу выпускной квалификационной работы «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение».	48
4	Перминов Валерий Афанасьевич, д. ф.-м. н., Доцент, ОКД.	Эксперт проекта	Консультирование по разделу выпускной квалификационной работы «Социальная безопасность».	48
5	Асадуллина Лилия Ильгизовна, ст. преподаватель, ОИЯ, ШОН.	Эксперт проекта	Консультации по разделу на английском языке выпускной квалификационной работы.	24
6	Залогин Никита Евгеньевич, Магистрант, Инженер.	Исполнитель по проекту	Исследование алгоритмов RL; Составление обзора актуальных статей и исследований; Выбор платформы для создания виртуальных сред; Реализация, изучение и улучшение выбранных алгоритмов RL; Проведение сравнительного анализа и представление результатов исследования, выводов и предложений. Составление пояснительной записки.	880
Итого				1188

6.2.3 Ограничения и допущения проекта

Ограничения проекта – это все факторы, которые могут ограничивать степень свободы членов команды проекта. Эти факторы перечислены в таблице 6.10

Таблица 6.10 – Ограничения проекта

Фактор	Ограничения / Допущения
Источник финансирования	Средства ТПУ
Сроки реализации проекта	30.01.2024 – 14.06.2024
Дата утверждения плана проект	03.03.2024
Дата завершения	14.06.2024
Продолжительность работы над проектом в день	8 часов

6.3 Планирование управления научно-техническим проектом

6.3.1 Иерархическая структура работ проекта

Иерархическая структура работ (ИСР) – детализация укрупнённой структуры работ. В процессе создания ИСР структурируется и определяется содержание всего проекта. Диаграмма ИСР показана на рисунке 6.1.

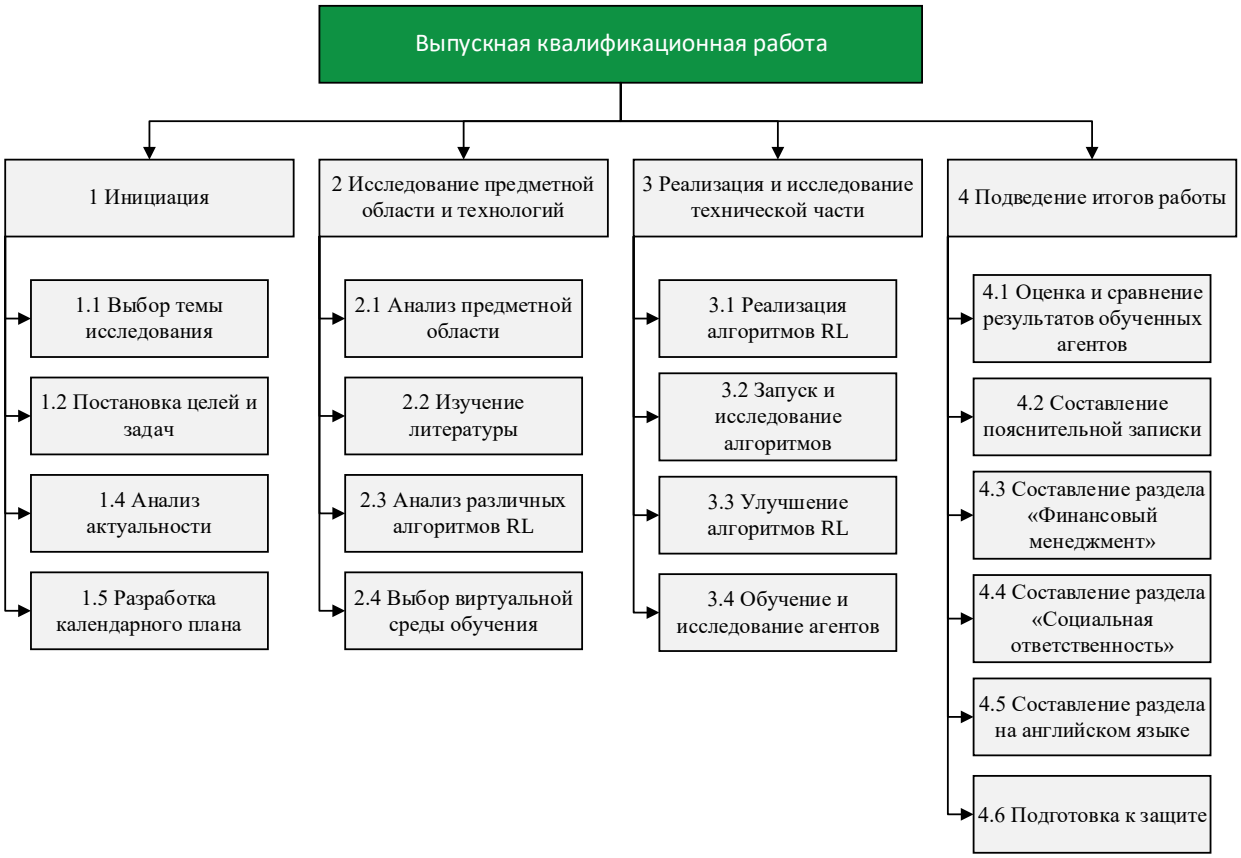


Рисунок 6.1 – Иерархическая структура работ (ИСР)

Таким образом ИСР позволила эффективно и понятно структурировать данную работу.

6.3.2 План проекта

В процессе организации работ по внедрению необходимо обоснованно планировать разделение труда и рабочего времени каждого участника процесса. Рабочие дни были рассчитаны исходя из шестидневной рабочей недели, с учётом праздничных дней Российской Федерации. Список участников представлен ниже:

– НР – Научный руководитель (Руководитель проекта), Спицын Владимир Григорьевич;
– НК – Научный консультант проекта, Григорьев Дмитрий Сергеевич;

– ФМ – Эксперт по разделу Финансового Менеджмента, Аникина Екатерина Алексеевна;
– СО – Эксперт по разделу Социальной Ответственности, Перминов Валерий Афанасьевич;
– АЧ – Эксперт по разделу на английском языке, Асадуллина Лилия Ильгизовна;
– ИС – Исполнитель по проекту (Магистрант), Залогин Никита Евгеньевич.

Для планирования разделения труда был составлен полный перечень работ, представленный в таблице 6.11.

Таблица 6.11 – Календарный план проекта

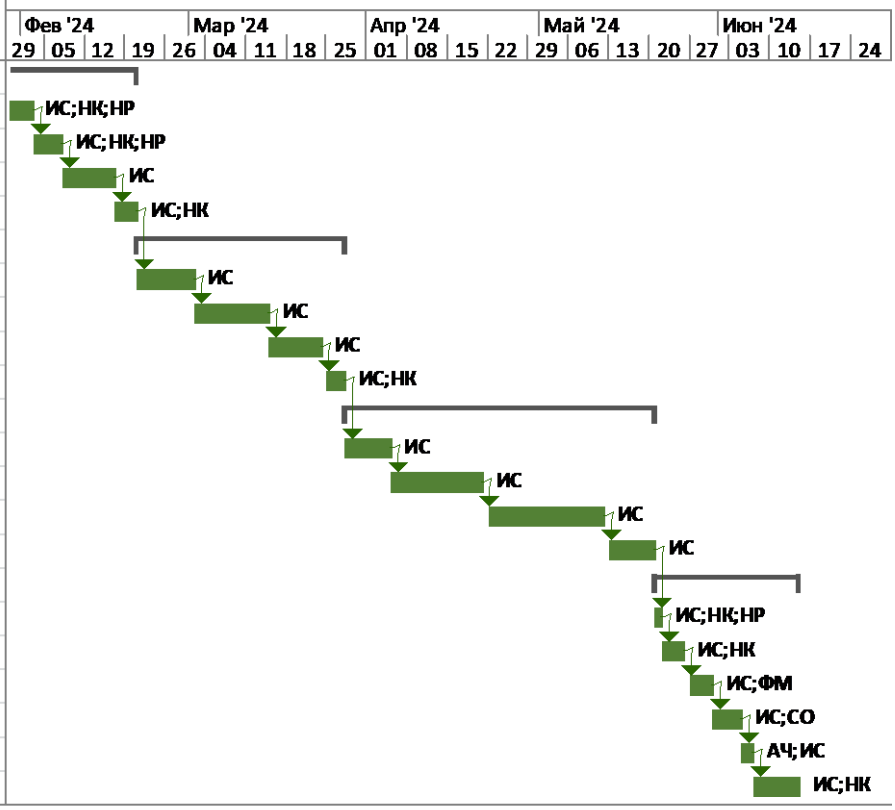
№	Название	Длитель-ность, раб. дни	Дата начала работ	Дата окончания работ	Состав участников
1	Инициация ВКР	19 дней	Вт 30.01.24	Вт 20.02.24	
1.1	Выбор темы исследования	4 дней	Вт 30.01.24	Пт 02.02.24	ИС;НК;НР
1.2	Постановка целей и задач	4 дней	Сб 03.02.24	Ср 07.02.24	ИС;НК;НР
1.3	Анализ актуальности	8 дней	Чт 08.02.24	Пт 16.02.24	ИС
1.4	Разработка календарного плана	3 дней	Сб 17.02.24	Вт 20.02.24	ИС;НК
2	Исследование предметной области и технологий	29 дней	Ср 21.02.24	Ср 27.03.24	
2.1	Анализ предметной области	8 дней	Ср 21.02.24	Пт 01.03.24	ИС
2.2	Изучение литературы	10 дней	Сб 02.03.24	Чт 14.03.24	ИС
2.3	Анализ различных алгоритмов RL	8 дней	Пт 15.03.24	Сб 23.03.24	ИС
2.4	Выбор виртуальной среды обучения	3 дней	Пн 25.03.24	Ср 27.03.24	ИС;НК
3	Реализация и исследование технической части	42 дней	Чт 28.03.24	Пн 20.05.24	
3.1	Реализация алгоритмов RL	7 дней	Чт 28.03.24	Чт 04.04.24	ИС
3.2	Запуск и исследование алгоритмов	14 дней	Пт 05.04.24	Сб 20.04.24	ИС
3.3	Улучшение алгоритмов RL	14 дней	Пн 22.04.24	Сб 11.05.24	ИС
3.4	Обучение и исследование агентов	7 дней	Пн 13.05.24	Пн 20.05.24	ИС
4	Подведение итогов работы	22 дней	Вт 21.05.24	Пт 14.06.24	
4.1	Оценка и сравнение результатов агентов	1 день	Вт 21.05.24	Вт 21.05.24	ИС;НК;НР
4.2	Составление пояснительной записки	4 дней	Ср 22.05.24	Сб 25.05.24	ИС;НК
4.3	Составление раздела «Финансовый менеджмент»	4 дней	Пн 27.05.24	Чт 30.05.24	ИС;ФМ

№	Название	Длитель-ность, раб. дни	Дата начала работ	Дата окончания работ	Состав участников
4.4	Составление раздела «Социальная ответственность»	4 дней	Пт 31.05.24	Вт 04.06.24	ИС;СО
4.5	Составление раздела на английском языке	2 дней	Ср 05.06.24	Чт 06.06.24	АЧ;ИС
4.6	Подготовка к защите	7 дней	Пт 07.06.24	Пт 14.06.24	ИС;НК

Диаграмма Ганта — это тип столбчатой диаграммы, иллюстрирующей график проекта. На этой диаграмме изображаются задачи и их временные интервалы. Диаграмма Ганта показана в таблице 6.12.

Таблица 6.12 – Диаграмма Ганта

Ид.	№	Исполнители	Длительность
2	**1**		**19 дней**
3	1.1	ИС;НК;НР	4 дней
4	1.2	ИС;НК;НР	4 дней
5	1.3	ИС	8 дней
6	1.4	ИС;НК	3 дней
7	**2**		**29 дней**
8	2.1	ИС	8 дней
9	2.2	ИС	10 дней
10	2.3	ИС	8 дней
11	2.4	ИС;НК	3 дней
12	**3**		**42 дней**
13	3.1	ИС	7 дней
14	3.2	ИС	14 дней
15	3.3	ИС	14 дней
16	3.4	ИС	7 дней
17	**4**		**22 дней**
18	4.1	ИС;НК;НР	1 день
19	4.2	ИС;НК	4 дней
20	4.3	ИС;ФМ	4 дней
21	4.4	ИС;СО	4 дней
22	4.5	АЧ;ИС	2 дней
23	4.6	ИС;НК	7 дней



6.4 Бюджет научного исследования

6.4.1 Специальное оборудование для научных (экспериментальных) работ

В эту статью включены все затраты, связанные с приобретением специального оборудования, необходимого для выполнения работ. Стоимость специального оборудования определяется в соответствии с действующими прейскурантами, затраты по которым приведены в таблице 6.14.

Таблица 6.14 – Расчёт стоимости специального оборудования

№ п/п	Наименование оборудования	Кол-во единиц оборудования	Цена единицы оборудования, руб.	Общая стоимость оборудования, руб.
1	Персональный компьютер	1	95 488	95 488
			Всего за материалы	95 488

Таким образом, итоговые затраты на специальное оборудование для выполнения научных работ составили 95 488 руб.

6.4.2 Основная заработная плата

В данном разделе рассматриваются оклад, стимулирующие и надбавки. Расчёт основан на сложности каждого этапа, а также на месячной зарплате исполнителя. Расчёт основной заработной платы представлен в таблице 6.15.

Таблица 6.15 – Расчёт основной заработной платы

№ п/п	Наименование этапов	Исполнители по категориям	Трудоёмкость, чел.-дн.	Заработная плата, приходящаяся на один чел.-дн., руб.	Всего заработная плата по тарифу, руб.
1	Инициация ВКР	НР	2	6 469,08	12938,16
2	Инициация ВКР	НК	3	3 780,80	11342,4
3	Инициация ВКР	ИС	14	595,17	20622,56
4	Исследование предметной области и технологий	НК	1	3 780,80	3780,8
5	Исследование предметной области и технологий	ИС	28	595,17	41245,12
6	Реализация и исследование технической части	ИС	42	595,17	61867,68
7	Подведение итогов работы	НР	1	6 469,08	6469,08
8	Подведение итогов работы	НК	3	3 780,80	11342,4
9	Подведение итогов работы	ФМ	1	4 627,79	4627,79
10	Подведение итогов работы	СО	1	4 627,79	4627,79
11	Подведение итогов работы	АЧ	1	4 627,79	4627,79
12	Подведение итогов работы	ИС	17	1 473,04	25041,68
				Итого	208533,25

Статья включает основную заработную плату работников, непосредственно занятых выполнением проекта, (включая премии, доплаты) и дополнительную заработную плату. Заработная плата рассчитывается по формуле 6.2:

$$C_{\text{зп}} = З_{\text{осн}} + З_{\text{доп}} \quad (6.2)$$

где $З_{\text{осн}}$ – основная заработная плата, руб.;

$З_{\text{доп}}$ – дополнительная заработная плата, руб.;

Основная заработная плата рассчитывается по формуле 6.3:

$$З_{\text{осн}} = З_{\text{дн}} + T_{\text{раб}} \quad (6.3)$$

где $З_{\text{дн}}$ – среднедневная заработная плата работника, руб.;

$Т_{\text{раб}}$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.

Среднедневная заработная плата рассчитывается по формуле 6.4:

$$З_{\text{дн}} = \frac{З_{\text{м}} * M}{F_{\text{д}}} \quad (6.4)$$

где $З_{\text{м}}$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года (при отпуске в 48 раб. дней М = 10,4 месяца, 6-дневная неделя);

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн. (252 дня).

В таблице 6.16 показано количество календарных дней, нерабочих и праздничных дней, дней в связи с потерей рабочего времени и фактический годовой фонд рабочего времени.

Таблица 6.16 – Баланс рабочего времени

Показатели рабочего времени	Работник
Календарное число дней	365
Количество нерабочих дней: – выходные дни; – праздничные дни.	 54 11
Потери рабочего времени: – отпуск; – невыходы по болезни.	 48 0
Действительный годовой фонд рабочего времени	252

Месячный должностной оклад работника рассчитывается по формуле 6.5:

$$З_м = З_б * (k_{пр} + k_д) * k_р \quad (6.5)$$

где $З_б$ — базовый оклад, руб.;

$k_{пр}$ — премиальный коэффициент;

$k_д$ — коэффициент доплат и надбавок;

$k_р$ — районный коэффициент (для Томска равен 1,3).

Таким образом, расчёты месячной заработной платы представлены в таблице 6.17.

Таблица 6.17 – Расчёт основной заработной платы

Исполнители	$З_б$, руб.	$k_{пр}$	$k_д$	$k_р$	$З_м$, руб.	$З_{дн}$, руб.	$Т_{раб}$, дн.	$З_{осн}$, руб.
Научный руководитель	52 700	1,1	1,188	1,3	156 750,88	6 469,08	3	19 407,25
Научный консультант	30 800				91 611,52	3 780,79	7	26 465,55
Эксперт по разделу Финансового Менеджмента	37 700				112 134,88	4 627,79	1	4 627,79
Эксперт по разделу Социальной Ответственности	37 700				112 134,88	4 627,79	1	4 627,79
Эксперт по английской части	37 700				112 134,88	4 627,79	1	4 627,79
Исполнитель по проекту (Магистрант)	12 000				35 692,80	1 473,04	101	148 776,66
							Итого	208 532,82

Таким образом, общая сумма основной заработной платы равна 208 532,82 руб.

6.4.3 Дополнительная заработная плата

В данную статью включается сумма выплат, предусмотренных законодательством о труде, например, оплата очередных и дополнительных отпусков; оплата времени, связанного с выполнением государственных и общественных обязанностей; выплата вознаграждения за выслугу лет и т.п. Дополнительная заработная плата рассчитывается на основе 10-15% от основного оклада работников по формуле 6.6:

$$З_{доп} = З_{осн} * k_{доп} \quad (6.6)$$

где $k_{доп}$ – коэффициент дополнительной зарплаты (10%).

В таблице 6.18 представлен расчёт дополнительной заработной платы.

Таблица 6.18 – Расчёт дополнительной заработной платы

Исполнители	Основная зарплата, руб	Дополнительная зарплата, руб.	Зар. плата исполнителя, руб..
Научный руководитель	19 407,25	1 940,73	21 347,98
Научный консультант	26 465,55	2 646,56	29 112,11
Эксперт по разделу Финансового Менеджмента	4 627,79	462,78	5 090,57
Эксперт по разделу Социальной Ответственности	4 627,79	462,78	5 090,57
Эксперт по английской части	4 627,79	462,78	5 090,57
Исполнитель по проекту (Магистрант)	148 776,66	14 877,67	163 654,32
Итого	208 532,82	20 853,28	229 386,11

Таким образом, общая сумма дополнительной заработной платы равна 20 853,28 руб.

6.4.4 Отчисления на социальные нужды

Отчисления на социальное страхование (так называемый трудовой налог) во внебюджетные фонды являются обязательными по нормам, установленным законодательством Российской Федерации на государственное социальное страхование, пенсионный фонд и медицинское страхование от расходов работников. Платёж во внебюджетные фонды определяется по формуле 6.7:

$$C_{\text{внеб}} = k_{\text{внеб}} * \left(З_{\text{осн}} + З_{\text{доп}}\right) \quad (6.7)$$

где $k_{\text{внеб}}$ – коэффициент отчислений на уплату во внебюджетные фонды (составляет 30%).

В таблице 6.19 приведён расчёт отчислений на социальные нужды с коэффициентом отчислений, равным 30%.

Таблица 6.19 – Расчёт отчислений на социальные нужды

Исполнители	Основная зарплата, руб	Дополнительная зарплата, руб.	Отчисления, руб.
Научный руководитель	19 407,25	1 940,73	6 404,39
Научный консультант	26 465,55	2 646,56	8 733,63
Эксперт по разделу Финансового Менеджмента	4 627,79	462,78	1 527,17
Эксперт по разделу Социальной Ответственности	4 627,79	462,78	1 527,17
Эксперт по английской части	4 627,79	462,78	1 527,17
Исполнитель по проекту (Магистрант)	148 776,66	14 877,67	49 096,30
Итого	208 532,83	20 853,30	68 815,84

Таким образом, общая сумма отчислений на социальные нужды равна 68 815,84 руб.

6.4.5 Накладные расходы

В эту статью включаются затраты на управление и хозяйственное обслуживание, которые могут быть отнесены непосредственно на конкретную тему. Кроме того, сюда относятся расходы по содержанию, эксплуатации и ремонту оборудования, производственного инструмента и инвентаря, зданий, сооружений и др. В расчётах эти расходы принимаются в размере 70-90% от суммы основной заработной платы научно-производственного персонала данной научно-технической организации.

Накладные расходы составляют 16% от суммы основной и дополнительной заработной платы, работников, непосредственно участвующих в выполнение темы. Расчёт накладных расходов ведётся по формуле 6.8:

$$C_{\text{накл}} = k_{\text{накл}} * \left(З_{\text{осн}} + З_{\text{доп}}\right) \quad (6.8)$$

где $k_{\text{внеб}}$ – коэффициент накладных расходов (составляет 16%).

В таблице 6.20 представлен расчёт накладных расходов.

Таблица 6.20 – Расчёт накладных расходов

Исполнители	Основная зарплата, руб	Дополнительная зарплата, руб.	Накладные расходы, руб.
Научный руководитель	19 407,25	1 940,73	3415,6768
Научный консультант	26 465,55	2 646,56	4657,9376
Эксперт по разделу Финансового Менеджмента	4 627,79	462,78	814,4912
Эксперт по разделу Социальной Ответственности	4 627,79	462,78	814,4912
Эксперт по английской части	4 627,79	462,78	814,4912
Исполнитель по проекту (Магистрант)	148 776,66	14 877,67	26184,6928
Итого	208 532,83	20 853,30	36 701,78

Таким образом, общая сумма накладных расходов равна 36 701,78 руб.

6.4.6 Формирование бюджетных расходов

Сформированные бюджетные расходы представлены в таблице 6.21.

Таблица 6.21 – Бюджет затрат на проект

Статьи	Стоимость, руб.
Расходы на специальное оборудование для научных экспериментов	95 488,00
Основная заработная плата	208 533,25
Дополнительная зарплата	20 853,28
Социальные отчисления	68 815,84
Накладные расходы	36 701,78
Итого запланированные расходы	430 392,15

Таким образом, общая сумма запланированные расходов равна 430 392,15 руб

6.5 Оценка сравнительной эффективности исследования

Определение эффективности происходит на основе расчёта интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин:

финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат трёх (или более) вариантов исполнения исследования. Для этого наибольший интегральный показатель принимается за базу расчёта, с которым соотносятся финансовые значения по всем вариантам исполнения. Интегральный финансовый показатель разработки определяется по формуле 6.9:

$$I_{\text{ф}}^{p} = \frac{\Phi_{pi}}{\Phi_{max}} \tag{6.9}$$

где Φ_{pi} – стоимость i-го варианта исполнения;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить формулами 6.10 и 6.11:

$$I_{m}^{p} = \sum_{i=1}^{n} a_i * b_i^{p}, \tag{6.10}$$

$$I_{m}^{a} = \sum_{i=1}^{n} a_i * b_i^{a} \tag{6.11}$$

где a_i – весовой коэффициент i-го параметра;

b_i^{p} b_i^{a} – бальная оценка i-го параметра для аналога и разработки, устанавливается экспертным путём по выбранной шкале оценивания;

n – число параметров сравнения.

Результаты определения интегрального показателя ресурсоэффективности приведены в таблице 6.22.

Таблица 6.22 – Определение интегрального показателя ресурсоэффективности

Критерии	Весовой коэффициент параметра	Текущий проект.	Аналог 1	Аналог 2	Аналог 3
Точность	0,35	4	3	3	5
Вычислительная сложность	0,25	4	5	5	2
Универсальность	0,1	4	2	2	4
Востребованность	0,1	5	4	4	5
Стоимость разработки	0,2	3	4	4	2
Итого	1	3,9	3,7	3,7	3,55

Интегральный показатель эффективности разработки и аналога определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формулам 6.12 и 6.13:

$$I^{p}_{\text{финр}} = \frac{I^{p}_{m}}{I^{p}_{\text{ф}}}, \tag{6.12}$$

$$I^{a}_{\text{финр}} = \frac{I^{a}_{m}}{I^{a}_{\text{ф}}} \tag{6.13}$$

где $I^{p}_{\text{финр}}$ $I^{a}_{\text{финр}}$ – интегральный показатель эффективности вариантов;

Сравнение интегрального показателя эффективности текущего проекта и аналогов позволит определить сравнительную эффективность проекта. Сравнительная эффективность проекта определяется по формуле 6.14:

$$\text{Э}_{\text{ср}} = \frac{I^{a}_{\text{финр}}}{I^{p}_{\text{финр}}}, \tag{6.14}$$

Результаты расчётов представлены в таблице 6.23

Таблица 6.23 – Сравнительная оценка характеристик вариантов исполнения проекта

Показатели	Текущий проект	Аналог 1	Аналог 2	Аналог 3
Интегральный финансовый показатель разработки	0,38	0,372	0,42	1
Интегральный показатель ресурсоэффективности разработки	3,9	3,7	3,7	3,55
Интегральный показатель эффективности разработки	10,264	9,94	8,809	3,55
Сравнительная эффективность вариантов исполнения	1	0,968	0,858	0,346

Сравнение значений интегральных показателей эффективности показывает, что текущий проект и два аналога, в целом, схожи по эффективности, но аналоги являются менее точными и менее быстрыми.

6.6 Выводы по разделу

В данном разделе были рассмотрены этапы проектирования и создания конкурентоспособных разработок, соответствующих требованиям ресурсоэффективности и ресурсосбережения.

Целью данной работы является разработка метода на основе обучения с подкреплением (Reinforcement Learning, RL) для обучения агента управлению манипулятором.

Была составлена карта потенциальных потребителей, показавшая возможные компании, использующие машинное обучение. Проведён SWOT-анализ, описывающий сильные и слабые стороны проекта, а также возможности и угрозы. Оценка готовности проекта к коммерциализации показала перспективы выше среднего. Для повышения готовности проекта к коммерциализации необходимо провести консультации с экспертами, сформировать команду и найти финансирование.

На этапе инициации проекта определены внутренние и внешние заинтересованные стороны, а также цели и ожидаемые результаты. График проекта представлен на диаграмме Ганта, показывающей распределение задач между исполнителями (руководитель, заказчик, эксперты по разделам ВКР и исполнитель) и временные рамки выполнения.

Стоимость материалов включена в бюджет проекта, рассчитана основная и дополнительная заработная плата исполнителей. Бюджет проекта составил 430 392,15 руб. после вычетов на социальные нужды и накладные расходы. Определена сравнительная эффективность проекта, которая, согласно расчётам, не уступает аналогам.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Обучающемуся:

Группа	ФИО
8ПМ2Л	Залогин Никита Евгеньевич

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	магистратура	Направление/специальность	09.04.04 Программная инженерия

Тема ВКР:

Обучение с подкреплением в виртуальных средах	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования: Алгоритмы обучения с подкреплением* *Область применения: Роботы-манипуляторы* *Рабочая зона: офис* *Размеры помещения: 10*8*3,5м.* *Количество и наименование оборудования рабочей зоны: десять персональных компьютеров, десять устройств ввода и вывода информации, двадцать ЖК мониторов.* *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне: присутствует окно, через которое может производится вентиляция помещения, принудительная вентиляция отсутствует; в зимнее время помещение отапливается; в помещении используется комбинированное освещение.*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018); ГОСТ 12.2.032-78 Рабочее место при выполнении работ сидя; ГОСТ 21889-76 Система "Человек-машина". Кресло человека-оператора; ГОСТ 22269-76 Рабочее место оператора. Взаимное расположение элементов рабочего места; ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения; СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.
2. Производственная безопасность при разработке проектного решения: – Анализ выявленных вредных и опасных производственных факторов – Расчет уровня опасного или вредного производственного фактора	**Вредные факторы:** Недостаточная освещенность рабочей зоны, нарушения микроклимата, повышенный уровень шума и электромагнитного излучения, вибрации, эмоциональные перегрузки, умственное перенапряжение. **Опасные факторы:** Опасность поражения электрическим током, короткое замыкание, статическое электричество. **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** беруши, наушники. **Расчет:** расчет системы искусственного освещения.
3. Экологическая безопасность при разработке проектного решения.	**Воздействие на селитебную зону:** потребление электроэнергии, шум. **Воздействие на литосферу:** образования отходов. **Воздействие на гидросферу:** потребление воды и электроэнергии, образование отходов, включая сточные

	воды и электронные отходы. **Воздействие на атмосферу:** потребление электроэнергии, тепловыделение.
4. Безопасность в чрезвычайных ситуациях <u>при разработке проектного решения.</u>	**Возможные ЧС:** морозы, несанкционированное проникновение посторонних на рабочее место, пожар. **Наиболее типичная ЧС:** Пожар.
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент (ОКД ИШНКБ)	Перминов Валерий Афанасьевич	д. ф.-м. н.		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ2Л	Залогин Никита Евгеньевич		

7 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

Введение

Целью данной работы является разработка метода на основе обучения с подкреплением для обучения агента управлению манипулятором.

В данной работе реализованы, исследованы и улучшены алгоритмы обучения с подкреплением, такие как Deep Q-Network (DQN) и Proximal Policy Optimization (PPO). Полученные решения позволяют решать задачи управления роботами-манипуляторами в виртуальных средах, что может найти применение в различных областях. Потенциальные пользователи полученных результатов могут быть другие исследователи в данной области и производственные предприятия. Результаты могут применяться в других виртуальных средах как цельное решение или часть большей системы. География использования не ограничена.

В данном разделе рассматриваются опасные и вредные факторы, их воздействие на работника и окружающую среду, а также юридические и организационные аспекты, включая меры безопасности в чрезвычайных ситуациях.

Работа проводилась в помещение размером 6*10*3.5м с использованием персональной электронной вычислительной машины (ПЭВМ) и сети интернет.

7.1 Правовые и организационные вопросы обеспечения безопасности

7.1.1 Правовые нормы трудового законодательства

Основные правовые нормы, обеспечивающие безопасность производства, описаны в Трудовом кодексе Российской Федерации (ТК РФ) [36]. Эти нормы определяют права и обязанности работников и работодателей,

а также регулируют вопросы трудоустройства, охраны труда, оплаты труда и другие аспекты трудовой деятельности.

В ходе научно-исследовательского проекта воздействие вредных и опасных производственных факторов на работника не превышает установленных норм. Изменения функционального состояния организма восстанавливаются во время предусмотренного отдыха. Условия труда, связанные с разработкой программного обеспечения (ПО), соответствуют установленным нормам в Трудовом кодексе РФ [36].

Продолжительность рабочего времени не должна превышать 40 часов в неделю. Для определенных категорий работников установлены сокращенные нормы:

- в возрасте до 16 лет − не более 24 часов в неделю;
- в возрасте от 16 до 18 лет − не более 35 часов в неделю;
- для инвалидов I или II группы − не более 35 часов в неделю;
- при вредных условиях труда 3 или 4 степени или опасных условиях труда – не более 36 часов в неделю.

Продолжительность рабочего времени работника определяется трудовым договором и устанавливается трудовым договором в различных форматах рабочей недели.

Рабочий режим включает в себя график работы, продолжительность рабочей недели, работу в гибкий график, время начала и окончания рабочего дня, перерывы, число смен, а также чередование рабочих и нерабочих дней.

Время отдыха разделяется на перерывы в течение рабочего дня, ежедневный отдых, выходные дни, нерабочие праздничные дни и отпуска. В рабочее время предусмотрен перерыв для отдыха и питания продолжительностью от 30 минут до двух часов не входящий в рабочее время.

Работникам предоставляется основной оплачиваемый отпуск продолжительностью 28 календарных дней с сохранением места работы и среднего заработка. Дополнительный отпуск предусмотрен для работников, занятых на опасных или вредных работах.

Оплата труда работника должна быть описана в трудовом договоре. К разработчику ПО применяется повременная система оплаты, с учетом отработанного времени. Нормирование труда используется для определения справедливой оплаты и оценки времени на каждый этап проекта.

7.1.2 Эргономические требования к правильному расположению и компоновке рабочей зоны

Выполнение научно-исследовательского проекта связано с работой на компьютере и требует соблюдения правовых норм к ПЭВМ и норм организации рабочего места, описанных в ГОСТ 12.2.032-78 [37], ГОСТ 21889-76 [38], ГОСТ 22269-76 [39], ГОСТ Р 50923-96 [40] и СанПиН 1.2.3685-21 [41].

Главными компонентами рабочего места оператора ПЭВМ являются рабочий стол, кресло, дисплей и клавиатура. Требования к рабочему месту описываются в ГОСТ Р 50923-96 [40].

Конструкция рабочего стола должна обеспечивать подходящее размещение оборудования и документов. Этот стол может быть с регулируемой или нерегулируемой высотой. Форму рабочей поверхности, в соответствии с ГОСТ 12.2.032-78, следует выбирать, учитывая характер работы.

Конструкция рабочего места должна учитывать антропометрические, физиологические и психологические особенности, а также характер выполняемой работы. Регулировка пространства для ног или использование подставки для ног может обеспечить оптимальное положение работающего при нерегулируемой высоте рабочей поверхности, а также при регулировке высоты стола и кресла. Стол разработчика соответствует требованиям, которые представлены в таблице 7.1.

Таблица 7.1 – Требования к рабочему месту

Параметр	Требования	Стол оператора
Высота рабочей поверхности	от 680 до 800 мм.	740 мм.
Глубина рабочей поверхности	не менее 600 мм.	700 мм.
Ширина рабочей поверхности	не менее 1200 мм.	2000 мм.
Высота пространства для ног	не менее 600 мм.	710 мм.
Ширина пространства для ног	не менее 500 мм.	1900 мм.
Глубина на уровне колен	не менее 450 мм.	700 мм.
Глубина на уровне вытянутых ног	не менее 650 мм.	700 мм.

Кресло оператора, согласно ГОСТ 21889-76 [38], должно содержать сиденье, спинку и подлокотники, а также может иметь подголовник и подставку для ног. Оно обязано обеспечивать регулировку высоты сиденья и угла наклона спинки, а также может быть подвижным, позволяя вращение на 180-360 градусов вокруг вертикальной оси. Кресло также должно поддерживать физиологически оптимальную позу оператора, обеспечивать изменение позы для снижения напряжения мышц и предотвращения нарушений циркуляции крови в нижних конечностях. Требования к креслу оператора описаны в таблице 7.2.

Таблица 7.2 – Требования к креслу оператора

Параметр	Требования	Кресло оператора
Высота поверхности сиденья	не менее 400 мм.	от 420 до 520 мм.
Глубина поверхности сиденья	не менее 400 мм.	540 мм.
Ширина поверхности сиденья	от 400 до 550 мм.	440 мм.
Высота спинки кресла	не менее 380 мм.	800 мм.
Ширина спинки кресла	от 280 мм.	480 мм.
Длинна подлокотников	не менее 250 мм.	280 мм.
Ширина подлокотников	от 50 до 70 мм.	70 мм.

Согласно ГОСТ 22269-76 [39], взаимное расположение элементов рабочего места должно обеспечивать возможность свободного выполнения всех необходимых движений и перемещений для работы с оборудованием. Оно должно способствовать оптимальному режиму работы и отдыха, уменьшению утомления и предотвращению ошибок.

При работе за ПЭВМ органами управления являются клавиатура и компьютерная мышь. Их следует размещать в ближней зоне, чтобы исключить перекрещивание рук. Редко используемые элементы, такие как принтер, звуковые колонки и системный блок, должны быть расположены в дальней зоне. Клавиатура должна быть размещена на расстоянии от 100 до 300 мм от переднего края стола или на регулируемой по высоте поверхности для обеспечения оптимальной видимости экрана.

Дисплей компьютера должен соответствовать требованиям ГОСТ Р 50923-96 [40] и быть установлен ниже уровня глаз оператора так, чтобы изображение было различимо без изменения положения головы. Угол наблюдения экрана не должен превышать 60 градусов относительно горизонтальной линии зрения. Освещенность рабочего места должна составлять от 300 до 500 лк, обеспечиваемая общим искусственным освещением. Для освещения зоны документов допускается использование местного освещения. Рабочий стол следует располагать так, чтобы окно находилось сбоку.

7.2 Производственная безопасность

7.1.1 Вредные факторы

7.1.1.1 Недостаточная освещенность рабочей зоны

Плохое естественное и искусственное освещение рабочего места оказывает влияние на физическое и психологическое состояние пользователя, что неблагоприятно сказывается на его работе. Не надлежащее качество освещения может привести к ухудшению зрения.

Согласно СП 52.13330.2016 [42] при работах III зрительного разряда и подразряда г (работы высокой точности) освещенность при системе общего освещения должна быть не ниже E = 200 Лк.

Расчет общего равномерного искусственного освещения горизонтальной рабочей поверхности выполняется методом коэффициента использования светового потока, учитывающим световой поток, отраженный от потолка и стен. Длина помещения A = 10м., ширина В = 8м., высота H = 3.5м.

Площадь помещения вычисляется по формуле 7.1:

$$S = A * B = 10 * 8 = 80\text{м}^2 \quad (7.1)$$

Коэффициент отражения стен, оклеенных светлыми обоями с окнами, без штор p_c = 30%, потолка светлой поверхности $p_{\text{п}}$ = 50%. Коэффициент запаса, учитывающий загрязнение светильника, для помещений с малым выделением пыли равен $K_{\text{з}}$ = 1,5. Коэффициент неравномерности для люминесцентных ламп Z = 1,1.

Для освещения рабочего помещения применяется три светильниками типа ОДОР-2-40 с двумя лампами ЛД-40 со световым потоком $\Phi_{\text{п}}$ = 2300 Лм. По паспорту длина светильника $A_{\text{св}}$ = 1227 мм., ширина $B_{\text{св}}$ = 265 мм. Мощность лампы $P_{\text{л}}$ 40 Вт.

Расстояние светильников от перекрытия (свес): h_c = 0,5 м. Высота рабочей поверхности над полом hp = 0,8 м.

Высота светильника над полом определяется по формуле 7.2:

$$h_{\text{п}} = H - h_c = 3{,}5 - 0{,}5 = 3\text{м} \quad (7.2)$$

Высота светильника над рабочей поверхностью определяется по формуле 7.3:

$$h = h_{\text{п}} - h_{\text{р}} = 3 - 0{,}8 = 2{,}2\text{м} \quad (7.3)$$

Интегральным критерием оптимальности расположения светильников является величина λ, которая для люминесцентных светильников с защитной решеткой лежит в диапазоне 1,1-1,3. Примем λ = 1,2.

Расстояние между светильниками определяется по формуле 7.4:

$$L = \lambda - h = 1{,}2 * 2{,}2 = 2{,}64\text{м} \quad (7.4)$$

Оптимальное расстояние от крайнего ряда светильников до стены рекомендуется принимать равным L/3, в данном случае это 0,88 м.

Индекс помещения определяется по формуле 7.5:

$$i = \frac{A*B}{h*(A+B)} = \frac{10*8}{2{,}2*(10+8)} = 2{,}02 \qquad (7.5)$$

Из методической таблицы коэффициент использования светового потока, показывающий какая часть светового потока ламп попадает на рабочую поверхность, для светильников типа ОДОР с люминесцентными лампами при p_c = 30%, $p_{\text{п}}$ = 50% и индексе помещения i = 1,11 равен η = 47%.

План помещения и размещения светильников с люминесцентными лампами представлен на рисунке 7.1.

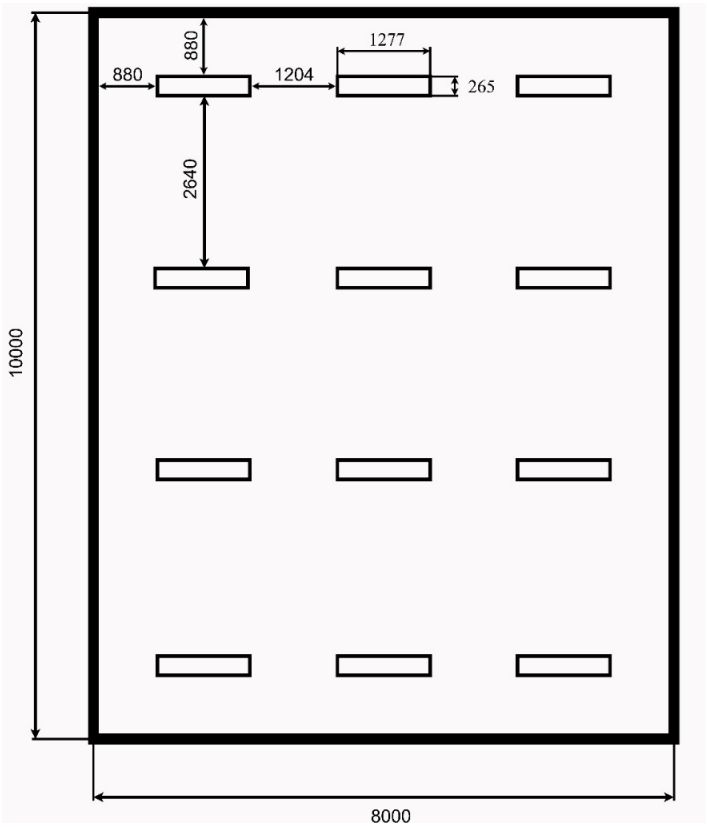


Рисунок 7.1 – с люминесцентными лампами в производственном помещении

Общее число светильников: $N_{\text{св}}$ = 12. Соответственно количество люминесцентных ламп: $N_{\text{лл}}$ = 24. Расчет светового потока группы люминесцентных ламп светильника определяется по формуле 7.6:

$$\Phi_{\text{рас}} = \frac{E*A*B*K_{\text{з}}*Z}{N_{\text{лл}}*\eta} = \frac{200*10*8*1.5*1.1}{24*0{,}47} = 2340 \qquad (7.6)$$

Проверку выполнения условия по формуле 7.7:

$$-10\% \leq \frac{\Phi_{\text{п}}-\Phi_{\text{рас}}}{\Phi_{\text{п}}} * 100\% = \frac{2300-2340}{2300} * 100\% = -1{,}73\% \leq 20\% \qquad (7.7)$$

Таким образом, полученный световой поток светильника не выходит за пределы требуемого диапазона.

Мощность осветительной установки рассчитывается по формуле 7.8:

$$P_{\text{уст}} = N * P_{\text{л}} = 24 * 40 = 960\text{Вт}. \qquad (7.8)$$

Удельная мощность осветительной установки рассчитывается по формуле 7.9:

$$P_{\text{уд}} = \frac{N*P_{\text{л}}}{S} = \frac{24*40}{80} = 12\,\frac{\text{Вт}}{\text{м}^2} \qquad (7.9)$$

Общие требования и рекомендации к организации освещения на рабочем месте:

– Рабочие места следует располагать так, чтобы естественный свет падал в основном слева, а мониторы были ориентированы боковой стороной к окнам.
– система общего равномерного освещения должна регулировать искусственное освещение в помещениях для эксплуатации ПЭВМ.

Вышеперечисленные меры полностью соблюдаются, что позволяет сохранить зрение и избежать пагубного воздействия на глаза во время разработки и эксплуатации результатов выпускной квалификационной работы.

7.1.1.2 Отклонение показателей микроклимата в помещении

Комфортные условия для работы создаются оптимальным сочетанием температуры, относительной влажности и скорости движения воздуха. На рабочих местах, где использование ПЭВМ является основным (диспетчерские, операторские, расчетные, кабины и посты управления, залы вычислительной техники и др.) и связана с нервноэмоциональным напряжением должны обеспечиваться оптимальные параметры микроклимата для категории работ 1а в соответствии с СанПиН 1.2.3685-21 [41]. Согласно этому документу,

должны быть соблюдены требования, описанные в таблицах 7.3 и 7.4.

Таблица 7.3 – Оптимальные нормы микроклимата

Период года	Температура воздуха, С°	Температура поверхностей, С°	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	22-24	21-25	60-40	0,1
Теплый	23-25	23-25	60-40	0,1

Таблица 7.4 – Допустимые нормы микроклимата

Период года	Температура воздуха, С°		Температура поверхностей, С°	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	
	Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			Диапазон температур воздуха ниже оптимальных величин	Диапазон температур воздуха выше оптимальных величин
Холодный	20 - 21,9	24,1 - 25	19 - 26	15 - 75	0.1	0.1
Теплый	21 - 22,9	25,1 - 28	20 - 29	15 - 75	0.1	0.2

Общая площадь рабочего помещения составляет 80 м², объем – 280 м³. Согласно санитарным нормам СП 2.2.3670-20 [43], на одного человека должно приходиться не менее 4,5 м² площади и 15 м³ объема. В соответствии с этими нормативами, размер помещения позволяет разместить необходимое количество рабочих мест, соблюдая санитарные нормы.

В помещении осуществляется естественная вентиляция через открываемые оконные проемы (форточки) и дверные проемы, что обеспечивает общеобменную вентиляцию. Основной недостаток данного типа вентиляции заключается в том, что приточный воздух поступает в помещение без предварительной очистки и нагревания.

Параметры микроклимата поддерживаются в холодное время года с помощью систем водяного отопления, нагревающего воду до 100°С, а в теплое время года с помощью кондиционирования.

7.1.1.3 Превышенный уровня шума

Шум на производстве, вызванный различным оборудованием и устройствами, является серьезным вредным фактором, отрицательно влияющим на здоровье и производительность сотрудников. Повышенные уровни шума могут привести к необратимым изменениям в слухе и оказать негативное воздействие на нервную систему, что снижает концентрацию, память и реакцию работников, увеличивая риск ошибок в работе.

Требования к допустимому уровню шума определены в ГОСТ 12.1.003-2014 [44] и СанПиН 1.2.3685-21 [41]. В помещениях операторов ЭВМ (без дисплеев) допустимый уровень шума не должен превышать 65 дБА, а на рабочих местах с шумными агрегатами вычислительных машин - 75 дБА. В процессе выполнения работ в рабочем помещении основным источником шума являлась ПЭВМ с системой воздушного охлаждения уровень шума которой при простое составлял в среднем 27 дБА, в режиме высокой нагрузки - около 60 дБА, иногда достигая 70 дБА.

При превышении допустимых уровней шума необходимо применять как средства коллективной защиты (СКЗ), так и средства индивидуальной защиты (СИЗ).

Средства коллективной защиты:

1. Устранение причин шума или его значительное ослабление непосредственно в источнике образования;
2. Изоляция источников шума от окружающей среды с использованием глушителей, экранов и звукопоглощающих строительных материалов;
3. Применение средств, уменьшающих шум и вибрацию на пути их распространения.

Средства индивидуальной защиты:

1. Использование специальной одежды и средств защиты органов слуха, таких как наушники, беруши и антифоны.

Для обеспечения защиты от шума при выполнении данной работы использовались средства индивидуальной защиты, в частности охватывающие наушники закрытого типа с достаточным уровнем пассивного шумоподавления.

7.1.1.4 Повышенный уровень электромагнитного излучения

Источником электромагнитных излучений в нашем случае являются дисплеи ПЭВМ. Монитор включает в себя излучения рентгеновской, ультрафиолетовой и инфракрасной области, а также широкий диапазон электромагнитных волн других частот. Согласно СанПиН 1.2.3685-21 [41] напряженность электромагнитного поля по электрической составляющей на расстоянии 50 см. вокруг видеодисплейного терминала (ВДТ) не должна превышать 25 В/м. в диапазоне от 5 Гц. до 2 кГц., 2,5 В/м. в диапазоне от 2 до 400 кГц. Плотность магнитного потока не должна превышать в диапазоне от 5 Гц. до 2 кГц 250 нТл., и 25 нТл. в диапазоне от 2 до 400 кГц. Поверхностный электростатический потенциал не должен превышать 500 В.

В ходе работы использовался монитор ПЭВМ типа Acer VG270 со следующими характеристиками: напряженность электромагнитного поля 2,5 В/м.; поверхностный потенциал составляет 450 В. (основы противопожарной защиты предприятий ГОСТ 12.1.004-91 [45] и ГОСТ 12.1.010-76 [46]).

При длительном постоянном воздействии электромагнитного поля (ЭМП) радиочастотного диапазона при работе на ПЭВМ у человеческого организма возможны сердечно-сосудистые, респираторные и нервные расстройства, головные боли, усталость, ухудшение состояния здоровья, гипотония, изменения сердечной мышцы проводимости. Тепловой эффект ЭМП характеризуется увеличением температуры тела, локальным селективным нагревом тканей, органов, клеток за счет перехода ЭМП на теплую энергию.

Предельно допустимые уровни (ПДУ) облучения (по ОСТ 54 30013-83) [47]:

- до 10 мкВт/см^2., время работы (8 часов);
- от 10 до 100 мкВт/см^2., время работы не более 2 часов;
- от 100 до 1000 мкВт/см^2., время работы не более 20 мин. при условии использования защитных очков;
- для населения в целом плотность потока мощности (ППМ) не должна превышать 1 мкВт/см2.

Защита человека от опасного воздействия электромагнитного излучения как средствами коллективной защиты (СКЗ), так и средствами индивидуальной защиты (СИЗ).

Средства коллективной защиты:

1. Защиту временем - ограничение времени пребывания в зоне воздействия излучения;
2. Защиту расстоянием - максимальное удаление от источника излучения;
3. Снижение интенсивности излучения в самом источнике;
4. Экранирование источника излучения или рабочего места;
5. Заземление экрана вокруг источника;

Средства индивидуальной защиты:

1. Очки и специальная одежда, выполненная из металлизированной ткани (кольчуга) для краткосрочной защиты;
2. Вместо обычных стекол используют стекла, покрытые тонким слоем золота или диоксида олова (SnO2).

7.1.1.5 Нервно-эмоциональное напряжение

Работа с ПЭВМ сопряжена с воздействием вредных психофизиологических факторов, таких как нервно-психические перегрузки. Эти перегрузки представляют собой совокупность изменений в

психофизиологическом состоянии организма, возникающих после выполнения работы и приводящих к временному снижению эффективности труда. Состояние утомления характеризуется специфическими объективными показателями и субъективными ощущениями.

Нервно-психические перегрузки подразделяются на умственное перенапряжение, перенапряжение анализаторов, монотонность труда и эмоциональные перегрузки. При первых признаках психического перенапряжения рекомендуется установление периодов для расслабления и восстановления, рациональное чередование работы и отдыха, занятия спортом, соблюдение регулярного режима сна и обращение за медицинской помощью при серьезных симптомах.

В соответствии с МР 2.2.9.2311-07 [48], работа с ПЭВМ классифицируется в зависимости от вида и категории трудовой деятельности, где виды работы подразделяются на три группы: считывание информации, ввод данных и творческая работа. Категории тяжести и напряженности работы с ПЭВМ определяются по суммарному числу считываемых или вводимых знаков либо по времени непосредственной работы, не превышающему 6 часов за смену. Если в рабочей смене выполняются работы разных видов, основной работой с ПЭВМ считается та, которая занимает не менее 50% времени.

Суммарное время регламентированных перерывов при работе с ПЭВМ представлено в таблице 7.5.

Таблица 7.5 – Суммарное время перерывов в зависимости от категории работы и нагрузки

Категория работы с ПЭВМ	Уровень нагрузки за рабочую смену при видах работ с ПЭВМ			Суммарное время регламентированных перерывов при 8-часовой смене, мин
	Группа А, количество знаков	Группа Б, количество знаков	Группа В, часов	
I	До 20000	До 15000	До 2	50
II	До 40000	До 30000	До 4	70
III	До 60000	До 40000	До 6	90

Согласно МР 2.2.9.2311-07 [48], если работа требует непрерывного взаимодействия с ВДТ и исключает возможность переключения на другие виды деятельности, рекомендуется делать перерывы продолжительностью 10-15 минут каждые 45-60 минут работы.

При проведении исследований уровень нагрузки относился к группе В, категория работы III. Согласно таблице, суммарное время перерывов необходимо установить не менее 90 минут.

7.1.2 Опасные факторы

7.1.2.1 Электроопасность, класс электроопасности помещения, безопасные номиналы I, U, Rзаземления

Несоблюдение правил безопасности при работе с ПЭВМ, в соответствии с ГОСТ 12.1.038-82 [49] может иметь серьезные последствия, включая травмы или даже летальный исход, из-за разнообразного воздействия электрического тока на организм человека, такого как термическое, электролитическое, биологическое и механическое.

Допустимые значения напряжения и силы тока прикосновения варьируются в зависимости от частоты тока. В нашем рабочем помещении, хотя используется однофазный электрический ток напряжением 220 В и частотой 50 Гц, отсутствуют факторы повышенной опасности, такие как высокая влажность или токопроводящая пыль, что классифицирует его как помещение без повышенной опасности поражения электрическим током.

Безопасными номиналами являются: сила тока $I < 0{,}1$ А., напряжение $U < (2–36)$ В., сопротивление заземления $R_{зазем} < 4$ Ом.

Для обеспечения защиты от поражения электрическим током используются как средства коллективной защиты (СКЗ), так и средства индивидуальной защиты (СИЗ). Средства коллективной защиты включают защитное заземление и зануление, применение малого напряжения,

электрическое разделение сетей, защитное отключение, изоляцию токоведущих частей, оградительные устройства, а также использование щитов, барьеров, клеток, ширм, заземляющих и шунтирующих штанг, а также специальных знаков и плакатов. Средства индивидуальной защиты включают использование диэлектрических перчаток, изолирующих клещей и штанг, слесарных инструментов с изолированными рукоятками, применение указателей напряжения, ношение диэлектрической обуви и использование подставок и ковриков.

7.1.2.2 Пожароопасность, категория пожароопасности помещения, марки огнетушителей, их назначение и ограничение применения, схема эвакуации

Помещения по взрывопожарной и пожарной опасности классифицируются на категории А, Б, В1-В4, Г и Д. Рабочее помещение относится к категории В согласно НПБ 105-03 [50], включающей горючие и трудногорючие вещества. Помещение имеет 1-ю степень огнестойкости по СНиП 21-01-97* [51] (кирпич, группа В1).

Для предотвращения пожаров необходимо учитывать их электрические (короткое замыкание, перегрузка) и неэлектрические причины (неосторожное обращение с огнем). Для тушения пожаров используются водопенные (ОХВП-10), углекислотные (ОУ-2) и порошковые огнетушители (ОП-5), размещаемые на каждом этаже в количестве не менее двух штук и располагаемые на видных местах вблизи от выходов из помещений на высоте не более 1,35 м. Первичные средства пожаротушения должны быть размещены так, чтобы не мешать безопасной эвакуации людей. Кроме того, для предотвращения пожаров необходимо обеспечить специальные помещения для хранения легковоспламеняющихся жидкостей с вентиляцией в соответствии с ГОСТ 12.4.021-75 [52] и СП 60.13330.2020 [53], изолированные помещения для хранения пылеобразной канифоли, а также установить автоматические

сигнализаторы (типа СВК-3 М 1) и обеспечить первичные средства пожаротушения.

Рабочее помещение полностью соответствует требованиям пожарной безопасности, а именно, наличие охранно-пожарной сигнализации, плана эвакуации, изображенного на рисунке 7.2, порошковых огнетушителей с поверенным клеймом, табличек с указанием направления к запасному (эвакуационному) выходу

Рисунок 7.2 – План эвакуации

7.3 Экологическая безопасность

На рабочем месте выявлен источник загрязнения литосферы и атмосферы из-за неправильной утилизации отходов вычислительной и оргтехники. Нормативы экологической безопасности установлены ГОСТ Р 70280-2022 [54] и ГОСТ Р 53692-2023 [55].

Согласно ГОСТ Р 53692-2023 [55], такие отходы вычислительной и оргтехники относятся к IV классу опасности и требуют специальной утилизации, при которой более 90% перерабатывается, а менее 10%

отправляется. Утилизация проходит в два этапа:

1. Переработка или полная утилизация обезвреженных отходов,
2. Безопасное размещение на полигонах или уничтожение.

Литиевые батареи и люминесцентные лампы требуют особой утилизации из-за токсичных металлов, таких как кобальт, никель и марганец, которые могут загрязнять водные источники и экосистемы. Предельно допустимые концентрации вредных веществ определены СанПиН 1.2.3685-21 [41]. Такие отходы по регламенту упаковываются и после накопления объемом в 1 транспортную единицу сдаются на переработку на соответствующее предприятие.

7.4 Безопасность в чрезвычайных ситуациях

Природная чрезвычайная ситуация — обстановка на определенной территории, вызванная природными факторами, которая может привести к человеческим жертвам, ущербу здоровью, материальным потерям и нарушению условий жизнедеятельности.

Производство находится в городе Томске с континентально-циклоническим климатом, без землетрясений, наводнений, засух и ураганов. Возможные ЧС — сильные морозы и диверсии. Возможными ЧС на объекте, могут быть сильные морозы и диверсия.

Для Сибирского региона в зимнее время года характерны сильные морозы. Достижение критически низких температур приводит к авариям систем жизнеобеспечения. В этом случае при подготовке к зиме следует предусмотреть:

- газобаллонные калориферы (запасные обогреватели);
- дизель или бензогенераторы;
- запасы питьевой и технической воды (не менее 30 л на 1 человека);
- теплый транспорт для доставки работников на работу и с работы домой в случае отказа муниципального транспорта.

Для предупреждения вероятности осуществления диверсии предприятие необходимо оборудовать системой видеонаблюдения, круглосуточной охраной, пропускной системой, надежной системой связи, а также исключения распространения информации о системе охраны объекта, расположении помещений и оборудования в помещениях, системах охраны, сигнализаторах, их местах установки и количестве. Должностные лица раз в полгода проводят тренировки по отработке действий на случай экстренной эвакуации.

Для предотвращения пожара необходимо соблюдать технику пожарной безопасности согласно ГОСТ 12.1.004-91 [45], включая регулярную проверку исправности электрических приборов и проводов, соблюдение техники безопасности при работе с электроприборами, а также заземление частей электроприборов для снижения статического заряда.

При возникновении пожара необходимо немедленно сообщить в пожарную охрану, эвакуировать людей, отключить электроэнергию и приступить к тушению первичными средствами.

Для тушения следует использовать углекислотные огнетушители (ОУ-5) и пожарный кран внутреннего противопожарного водопровода, избегайте химических пенных огнетушителей. Категорически запрещается тушить возгорания в помещениях офиса при помощи химических пенных огнетушителей [45].

7.5 Выводы по разделу

Значение всех производственных факторов на изучаемом рабочем месте соответствует нормам.

Категория тяжести труда в лаборатории по СанПиН 1.2.3685-21 [41] относится к категории Iа (работы, производимые сидя и сопровождающиеся незначительным физическим напряжением).

Для минимизации влияния физиологического и психофизиологического

воздействия на организм человека, достаточно соблюдать меры, приведённые в МР 2.2.9.2311-07 [48].

Категория помещения по электробезопасности, согласно ПУЭ, соответствует первому классу — «помещения без повышенной опасности» [56]. Персонал должен обладать I группой допуска по электробезопасности. Присвоение I группы проводится сотрудником с III группой по электробезопасности или специалистом по охране труда с IV группой, назначенным распоряжением руководителя организации [57].

Помещение относится к категории В1-В4: горючие и трудногорючие жидкости, твёрдые горючие и трудногорючие вещества и материалы, способные при взаимодействии с водой, кислородом воздуха или друг с другом только гореть, при условии, что помещения не относятся к категории А или Б [58].

Согласно постановлению «Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий», объект оказывает незначительное негативное воздействие на окружающую среду и относится к объектам III категории [59].

ЗАКЛЮЧЕНИЕ

В ходе данной преддипломной работы была достигнута поставленная цель. А также были решены следующие задачи:

1. Изучены алгоритмы обучения с подкреплением DQN и PPO, применяемые для создания агентов в виртуальных средах.
2. Составлен аналитический обзор статей по популярным алгоритмам и их модификациям, а также их применению в виртуальных или реальных средах с манипуляторами;
3. Выбрана платформа PyBullet для симуляции виртуальной среды с помощью сравнительного анализа среди одних из популярных платформ.
4. Проведены эксперименты с решениями DQN, PPO и их улучшениями для обучения агента в средах с манипулятором.
5. Проведено сравнение полученных агентов между собой, где было выявлено, что для поставленной задачи лучшим является агент PPOv3, а агент PPOv4 наиболее перспективным для дальнейшего обучения и улучшения. Полученные агенты сравнивались с аналогичными решениями, где продемонстрировали ряд преимуществ над некоторыми конкурентами, такие как точность и стабильность.

В будущем предполагается дальнейшее улучшение алгоритма PPO и общего процесса обучения для достижения лучших результатов и решения проблемы переобучения. Апробация алгоритмов LSTM и xLSTM для данной задачи. Рассмотрение различных дополнительных стратегий на подобие HER или PCCL для off-policy алгоритмов использованные авторами статей упомянутых в пункте 2, может расширить возможности и повысить производительность агентов. Замена среды обучения на уже существующие варианты или разработка авторской системы симуляции среды может открыть новые направления исследований в области обучения с подкреплением.

СПИСОК ПУБЛИКАЦИЙ ОБУЧАЮЩЕГОСЯ

1. Залогин Н.Е. Обучение агентов в виртуальной среде KukaDiverseObjectEnv // Сборник трудов XXI Международной научно-практической конференции студентов и молодых ученых «Молодежь и современные информационные технологии». Томск, 15-18 апреля 2024. Томск: изд-во ТПУ. 2024;
2. Залогин Н.Е. Обучение агентов в виртуальной среде KukaDiverseObjectEnv // Сборник научных трудов XX Международной конференции студентов, аспирантов и молодых ученых «Перспективы развития фундаментальных наук». Томск, 23-26 апреля 2024. Томск: изд-во ТПУ. 2024 - Том. 7. IT-технологии и электроника;
3. Залогин Н.Е. Обучение агентов в виртуальной среде KukaDiverseObjectEnv // Сборник научных трудов XI Международная молодежная научная конференция «Математическое и программное обеспечение информационных, технических и экономических систем». Томск, 24 – 27 мая 2024 г. Томск: изд-во ТГУ. 2024.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Arulkumaran K. et al. Deep reinforcement learning: A brief survey / K. Arulkumaran et al. – Текст : электронный // IEEE Signal Processing Magazine. – 2017. – Т. 34. – №. 6. – С. 26-38. – URL: https://ieeexplore.ieee.org/abstract/document/8103164 (дата обращения: 19.05.2024). – Режим доступа: ieeexplore.ieee ieeexplore.ieee.org;

2. Wang H. et al. Deep reinforcement learning: a survey / H. Wang et al. – Текст : электронный // Frontiers of Information Technology & Electronic Engineering. – 2020. – Т. 21. – №. 12. – С. 1726-1744. URL: https://link.springer.com/article/10.1631/fitee.1900533 (дата обращения: 19.05.2024). – Режим доступа: link.springer link.springer.com;

3. Bullet Real-Time Physics Simulation : официальный сайт. – URL: https://pybullet.org/ (дата обращения: 19.05.2024). – Текст : электронный;

4. Gazebo : официальный сайт. – URL: https://gazebosim.org/ (дата обращения: 19.05.2024). – Текст : электронный;

5. Andy Zeng. et al. TossingBot: Learning to Throw Arbitrary Objects with Residual Physics / Zeng Andy. et al. – Текст : электронный //arXiv preprint arXiv:1903.11239. – 2020. URL: https://arxiv.org/abs/1903.11239 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org;

6. Andy Zeng. et al. Transporter Networks: Rearranging the Visual World for Robotic Manipulation / Zeng Andy. et al. – Текст : электронный //arXiv preprint arXiv:2010.14406. – 2022. URL: https://arxiv.org/abs/2010.14406 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org;

7. Mnih V. et al. Human-level control through deep reinforcement learning //nature. – 2015. – Т. 518. – №. 7540. – С. 529-533. URL: https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf (дата обращения: 19.05.2024). – Режим доступа: googleapis googleapis.com

8. John Schulman. et al. Proximal Policy Optimization Algorithms / Schulman John. et al. – Текст : электронный //arXiv preprint arXiv:1707.06347. –

2017. URL: https://arxiv.org/abs/1707.06347 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

9. John Schulman. et al. Trust Region Policy Optimization / Schulman John. et al. – Текст : электронный //arXiv preprint arXiv:1502.05477. – 2017. URL: https://arxiv.org/abs/1502.05477 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

10. Huang. et al., The 37 Implementation Details of Proximal Policy Optimization / Huang. et al. – Текст : электронный // ICLR Blog Track – 2022. URL: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/ (дата обращения: 19.05.2024). – Режим доступа: The ICLR Blog Track iclr-blog-track.github.io.

11. Sepp Hochreiter et al. Long Short-term Memory / Hochreiter Sepp. et al. – Текст : электронный //Neural computation. – 1997. – Т. 9. – №. 8. – С. 1735-1780. URL: https://www.researchgate.net/publication/13853244_Long_Short-term_Memory (дата обращения: 19.05.2024). – Режим доступа: google research research.google.

12. Logan Engstrom. et al. Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO / Engstrom Logan. et al. – Текст : электронный //arXiv preprint arXiv:2005.12729. – 2020. URL: https://arxiv.org/abs/2005.12729 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

13. Marcin Andrychowicz. et al. What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study / Andrychowicz Marcin. et al. – Текст : электронный //International conference on learning representations. – 2020. URL: https://research.google/pubs/what-matters-for-on-policy-deep-actor-critic-methods-a-large-scale-study/ (дата обращения: 19.05.2024). – Режим доступа: google research research.google.

14. Vaswani A. et al. Attention is all you need / Ashish Vaswani. – Текст : электронный //arXiv preprint arXiv:1706.03762. – 2017. URL: https://arxiv.org/abs/1706.03762 (дата обращения: 19.05.2024). – Режим

доступа: arxiv arxiv.org.

15. Lili Chen. et al. Decision Transformer: Reinforcement Learning via Sequence Modeling / Chen Lili. et al. – Текст : электронный //arXiv preprint arXiv:2106.01345 – 2021. URL: https://arxiv.org/abs/2106.01345 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

16. Oriol Vinyals. et al. StarCraft II: A New Challenge for Reinforcement Learning / Vinyals Oriol. et al. – Текст : электронный //arXiv preprint arXiv:1708.04782 – 2017. URL: https://arxiv.org/abs/1708.04782 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

17. Christopher Berner. et al. Dota 2 with Large Scale Deep Reinforcement Learning / Berner Christopher. et al. – Текст : электронный //arXiv preprint arXiv:1912.06680– 2019. URL: https://arxiv.org/abs/1912.06680 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

18. Maximilian Beck. et al. xLSTM: Extended Long Short-Term Memory / Beck Maximilian. et al. – Текст : электронный //arXiv preprint arXiv:2405.04517– 2024. URL: https://arxiv.org/abs/2405.04517 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

19. Tom B. Brown et al. Language Models are Few-Shot Learners / Brown Tom B. et al. – Текст : электронный //arXiv preprint arXiv:2005.14165– 2020. URL: https://arxiv.org/abs/2005.14165 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

20. Погружение в xLSTM – обновленную LSTM, которая может оказаться заменой трансформера // Data Secrets. – 2024. – URL: https://datasecrets.ru/articles/10 (дата обращения: 19.05.2024);

21. Pierre Aumjaud. et al. Reinforcement Learning Experiments and Benchmark for Solving Robotic Reaching Tasks / Aumjaud Pierre. et al. – Текст : электронный //arXiv preprint arXiv:2011.05782. – 2020. URL: https://arxiv.org/abs/2011.05782 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

22. Swagat Kumar. et al. Benchmarking Deep Reinforcement Learning

Algorithms for Vision-based Robotics / Kumar Swagat. et al. – Текст : электронный //arXiv preprint arXiv:2201.04224. – 2022. URL: https://arxiv.org/abs/2201.04224v1 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

23. Chen H. Robotic manipulation with reinforcement learning, state representation learning, and imitation learning (student abstract) / H. Chen. – Текст : электронный //Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Т. 35. – №. 18. – С. 15769-15770. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17881 (дата обращения: 19.05.2024). – Режим доступа: ojs.aaai ojs.aaai.org.

24. Sha Luo. et al. Accelerating Reinforcement Learning for Reaching using Continuous Curriculum Learning / Luo Sha. et al. – Текст : электронный //arXiv preprint arXiv:2002.02697. – 2020. URL: https://arxiv.org/abs/2002.02697 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

25. Luka K. et al. Safe Reinforcement Learning in a Simulated Robotic Arm / K. Luka. et al. – Текст : электронный //arXiv preprint arXiv:2312.09468. – 2024. URL: https://arxiv.org/abs/2312.09468 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

26. OpenAi Spinning Up Algorithms // OpenAI. – 2018 – 2024. – URL: https://spinningup.openai.com/en/latest/user/algorithms.html (дата обращения: 19.05.2024);

27. Lilian Weng A (Long) Peek into Reinforcement Learning // lil'log. – 2018 – 2024. URL: https://lilianweng.github.io/posts/2018-02-19-rl-overview/#deep-q-network (дата обращения: 19.05.2024);

28. Actor-Critic Algorithm in Reinforcement Learning // geeksforgeeks. – 2022 – 2024. URL: https://www.geeksforgeeks.org/actor-critic-algorithm-in-reinforcement-learning/ (дата обращения: 19.05.2024);

29. Lilian Weng Policy Gradient Algorithms // Lil'Log – 2018 – 2024. URL: https://lilianweng.github.io/posts/2018-04-08-policy-gradient/#ppo (дата обращения: 19.05.2024);

30. Josh Achiam Proximal Policy Optimization // OpenAI. – 2018 – 2024. – URL: https://spinningup.openai.com/en/latest/algorithms/ppo.html (дата обращения: 19.05.2024);

31. MuJoCo : официальный сайт. – URL: https://mujoco.org/ (дата обращения: 19.05.2024). – Текст : электронный;

32. CoppeliaSim : официальный сайт. – URL: https://www.coppeliarobotics.com/ (дата обращения: 19.05.2024). – Текст : электронный;

33. KukaDiverseObjectEnv : сайт. GitHub, Inc. – 2024. – URL: https://github.com/bulletphysics/bullet3/blob/master/examples/pybullet/gym/pybullet_envs/bullet/kuka_diverse_object_gym_env.py (дата обращения: 19.05.2024). – Текст: электронный;

34. multiprocessing – Process-based parallelism // Python Software Foundation. – 2001 – 2023. URL: https://docs.python.org/3.10/library/multiprocessing.html (дата обращения: 19.05.2024);

35. Swagat K. et al. Benchmarking Deep Reinforcement Learning Algorithms for Vision-based Robotics / K. Swagat et al. – Текст : электронный //arXiv preprint arXiv:2201.04224. – 2024. URL: https://arxiv.org/abs/2201.04224 (дата обращения: 19.05.2024). – Режим доступа: arxiv arxiv.org.

36. Кодекс 197-ФЗ Трудовой кодекс Российской Федерации: дата введения 30-12-2001 (ред. от 06.04.2024) – Текст : непосредственный;

37. ГОСТ 12.2.032-78 Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования: дата введения 26-04-1978 – Текст : непосредственный;

38. ГОСТ 21889-76 Система "человек-машина". Кресло человека-оператора. Общие эргономические требования: дата введения 25-05-1976 – Текст : непосредственный;

39. ГОСТ 22269-76 Система "человек-машина". Рабочее место оператора. Взаимное расположение элементов рабочего места. Общие

эргономические требования: дата введения 22-12-1976 – Текст : непосредственный;

40. ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения: дата введения 10-07-1996 – Текст : непосредственный;

41. СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания: дата введения 28-01-2021 – Текст : непосредственный;

42. СП 52.13330.2016. Естественное и искусственное освещение: дата введения 07-11-2016 – Текст : непосредственный: дата введения – Текст : непосредственный;

43. СП 2.2.3670-20 Санитарно-эпидемиологические требования к условиям труда: дата введения 02-12-2020 – Текст : непосредственный;

44. ГОСТ 12.1.003-2014 Система стандартов безопасности труда. Шум. Общие требования безопасности: дата введения 29-12-2014 – Текст : непосредственный;

45. ГОСТ 12.1.004-91 Система стандартов безопасности труда. Пожарная безопасность. Общие требования: дата введения 14-06-1991 – Текст : непосредственный;

46. ГОСТ 12.1.010-76 Система стандартов безопасности труда. Взрывобезопасность. Общие требования: дата введения 28-06-1976 – Текст : непосредственный;

47. ОСТ54-30013-83 Система стандартов безопасности труда. Электромагнитные излучения СВЧ. Предельно допустимые уровни облучения. Требования безопасности: дата введения 01-01-1984 – Текст : непосредственный;

48. МР 2.2.9.2311-07 Профилактика стрессового состояния работников при различных видах профессиональной деятельности: дата введения 18-12-2007 – Текст : непосредственный;

49. ГОСТ 12.1.038-82 Система стандартов безопасности труда.

Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов: дата введения 30-07-1982 – Текст : непосредственный;

50. НПБ 105-03 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности: дата введения 18-06-2003 – Текст : непосредственный;

51. СНиП 21-01-97* Пожарная безопасность зданий и сооружений: дата введения 13-02-1997 – Текст : непосредственный;

52. ГОСТ 12.4.021-75 Система стандартов безопасности труда. Системы вентиляционные. Общие требования: дата введения 13.11.1975 – Текст : непосредственный;

53. СП 60.13330.2020 Отопление, вентиляция и кондиционирование воздуха: дата введения 30.12.2020 – Текст : непосредственный;

54. ГОСТ Р 70280-2022 Охрана окружающей среды. Почвы. Общие требования по контролю и охране от загрязнения: дата введения 05-10-2022 – Текст : непосредственный;

55. ГОСТ Р 53692-2023 Ресурсосбережение. Обращение с отходами. Этапы технологического цикла отходов: дата введения 25-10-2023 – Текст : непосредственный;

56. Правила устройства электроустановок (ПУЭ). Седьмое издание: дата введения 08-07-2002 – Текст : непосредственный;

57. Приказ 903н Об утверждении Правил по охране труда при эксплуатации электроустановок: дата введения 15-12-2020 – Текст : непосредственный;

58. СП 12.13130.2009 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности: дата введения 25-03-2009 – Текст : непосредственный;

59. Приказ Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий: дата введения 31-12-2020 года N 2398 (с изм. от 07-10-2021) – Текст : непосредственный;

Приложение А
(справочное)

Chapter 1
Analytical review

Обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ2Л	Залогин Никита Евгеньевич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Спицын В.Г.	д.т.н.		

Консультант – лингвист ОИЯ ШОН

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ст. преподаватель	Асадуллина Л. И.			

INTRODUCTION

In recent years, robotics has been gaining prominence in various fields such as manufacturing, logistics, and medicine. Robotic manipulators, thanks to their high precision and repeatability, have become indispensable tools for performing complex industrial tasks. One of the main challenges in robotics is teaching robots new skills. Traditional robot programming methods require significant time and effort, and are also not very effective for tasks that require adaptability to changing conditions [1, 2]. To address this issue, reinforcement learning (RL) algorithms have been developed, representing a promising approach to creating agents capable of self-learning through trial and error and making decisions in complex and dynamic environments, both virtual and real.

Virtual environments like PyBullet [3] and Gazebo [4] provide safe and convenient platforms for training robots. They enable the quick and efficient creation of various scenarios, the testing of RL algorithms, and the evaluation of their effectiveness.

Reinforcement learning algorithms represent a rapidly advancing area of research in machine learning. Their application and further development could lead to the creation of more intelligent and adaptive systems capable of solving complex tasks and ensuring optimal behavior in diverse conditions. In recent years, research groups have achieved significant successes in the field of RL for robots, as evidenced by works such as TossingBot [5] and Ravens - Transporter Networks [6].

The relevance of this work is driven by the rapid advancement of RL technologies and their application in virtual environments, which opens up new possibilities for automating and optimizing complex tasks, such as robot manipulation control.

To achieve the goals of this study and test the hypotheses, various research methods were employed. Literature review was conducted to explore scholarly sources on topics related to reinforcement learning, DQN and PPO algorithms, as well as their application in virtual environments. Mechanisms and architectures of

neural networks were studied, including the adaptation of attention mechanisms to improve the PPO algorithm. Programming encompassed the development of RL models. Experiments and synthesis were utilized for hyperparameter tuning and iterative improvement of the PPO algorithm. Comparative analysis was performed to assess the performance of the original and modified versions of the algorithms based on key metrics, as well as to compare them with similar solutions.

The aim of this study is to investigate reinforcement learning algorithms and models used for creating agents in virtual environments, as well as to examine, utilize, and enhance examples of their implementations.

To achieve the stated goal, the following tasks were identified and addressed:

1. Investigate various RL algorithms used for creating agents in virtual or real environments.
2. Compile a review of current articles and research dedicated to popular RL algorithms and their modifications, as well as their application in environments with manipulators. Within this phase, an analysis of the advantages and disadvantages of various approaches, their effectiveness, and applicability in real and virtual environments will be conducted. Such a review will help identify trends and promising directions in the field of RL for robotics.
3. Conduct a comparative analysis of popular platforms for creating virtual environments, such as PyBullet, Gazebo, MuJoCo, and CoppeliaSim. The analysis will include evaluating functionality, ease of use, integration capabilities with RL algorithms, performance, and community support. Based on this analysis, the most suitable platform for conducting further experiments will be selected.
4. Conducting experiments with selected RL algorithms in a manipulator environment. A comparison of different solutions will be performed in terms of their effectiveness, stability, and learning capabilities.
5. Performing a comparative analysis of trained agents developed during the study, both among themselves and with several similar solutions.

1 ANALYTICAL REVIEW

1.1 Reinforcement Learning

Reinforcement Learning (RL) is a domain within machine learning where agents autonomously learn by interacting with their environment. RL operates on the principle of trial and error, wherein agents receive rewards for desired behaviors and penalties for undesired ones.

Some of the main advantages of reinforcement learning include:

- autonomous learning: Agents can learn without explicit programming;
- universality: RL is applicable to a wide range of tasks, including robotics control, gaming, and finance;
- efficiency: RL can find optimal or near-optimal strategies for complex tasks.

Some of the main disadvantages include:

- complexity: RL may require significant computational resources for training;
- overfitting problem: RL agents may excessively adapt to training data and perform poorly in new conditions;
- local optima problem: RL agents may get stuck in local optima and fail to reach the global optimum.

One of the popular reinforcement learning (RL) algorithms is Deep Q-Networks (DQN) (Mnih et al., 2013) [7], presented in the paper «Human-level control through deep reinforcement learning» Traditionally, RL methods have faced challenges with high-dimensional sensory inputs, similar to those encountered in real-world conditions. The DQN network introduced in the paper is an approach that utilizes deep neural networks for learning effective control policies directly based on high-dimensional sensory input data.

Recently, researchers have been drawn to the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) [8], proposed as an enhancement of the

Trust Region Policy Optimization (TRPO) algorithm (Schulman et al., 2017) [9]. PPO utilizes a computationally simpler objective function that does not directly evaluate how good the agent's policy is for achieving maximum reward. Instead, it employs a function that approximates this value. This makes the training process more stable and efficient. Despite this, the paper demonstrates that PPO is more sample efficient than TRPO in many control tasks. PPO also exhibits good empirical performance in the Arcade Learning Environment (ALE), which includes Atari games.

The PPO algorithm has been implemented by individual developers [10] with descriptions of 37 key implementation aspects for various environments and tasks:

- 13 main implementation details;
- 9 details for Atari games;
- 9 details for robotics tasks with continuous action spaces;
- 5 details for Long Short-Term Memory (LSTM) models (Hochreiter et al., 1997) [11];
- 1 detail for multi-discrete action spaces.

The authors of [12] conducted a study on 20 different key aspects of implementation performance in gradient-based policy learning algorithms, focusing on two popular methods: PPO and TRPO. In turn, the study described in [13] delved into the efficiency of 68 different key aspects of actor-critic (A2C) on-policy learning algorithms in continuous control tasks. The authors trained over 250000 agents in five different continuous control environments of increasing complexity and conducted a large-scale empirical investigation to identify which implementation details most significantly impact the performance of the trained agent.

In summary, the authors of [10, 12, 13] made the following conclusions:

- neural network architecture: the choice of architecture and activation function plays a critical role in a performance;
- code-level optimizations: non-obvious implementation details can fundamentally impact algorithm performance, altering their behavior

unpredictably;

- hyperparameters: batch size, learning rate, entropy coefficient, discount factor, and others are extremely important factors for both performance and effectiveness;
- trust region: optimizations often define the nature of the trust region used by the algorithms;
- PPO superiority over TRPO: much of PPO's advantages over TRPO (and even stochastic gradient descent) are due to code-level optimizations;
- modular design: it is highly recommended to develop RL algorithms with a modular structure to precisely evaluate the impact of each implementation detail;
- thorough evaluation: it is recommended to carefully analyze RL methods, going beyond simple performance comparisons;
- environment: training results with the same hyperparameters can vary significantly depending on the type of task.

The latest major breakthrough in the application of neural networks to reinforcement learning tasks has been the Transformer models and their attention mechanism (Ashish Vaswani et al., 2017) [14]. This work has fundamentally changed approaches to sequence data processing and has enabled remarkable achievements in natural language processing (NLP) and computer vision tasks.

Further development of Transformers occurred in the work «Decision Transformer: Reinforcement Learning via Sequence Modeling» (Lili Chen et al., 2021) [15], where researchers proposed a new architecture called Decision Transformer for RL tasks. Decision Transformer reimagines the RL task as a sequential modeling problem, predicting future actions of the agent based on its current state and the history of past actions and rewards. This approach allows for the utilization of powerful NLP methods, such as attention mechanisms and self-attention, to train efficient RL policies.

One of the competitors to Transformers is Long Short-Term Memory (LSTM). Despite LSTM significantly lagging behind Transformers in efficiency and

parallelization capabilities due to their recurrent nature, they have proven themselves in several RL tasks. For example, the AlphaStar models for the game StarCraft II [16] and OpenAI Five for Dota 2 [17] used LSTM and demonstrated outstanding results

In addition, the model Extended Long Short-Term Memory (xLSTM) (Maximilian Beck et al., 2024) [18], has achieved significant improvements compared to the original LSTM model. In NLP tasks, the new model has shown results surpassing or comparable to popular counterparts, including GPT-3 [19]. It is anticipated that xLSTM could become a worthy competitor to Transformers and other models in various domains, including RL [20].

1.2 Reinforcement Learning in Manipulator Environments

Research in the field of RL agents in virtual environments with manipulators is gaining widespread attention. This is because traditional methods of programming robots require significant time and effort, and moreover, they are not always suitable for tasks that require adaptability to changing conditions.

The authors of the paper [21] conduct research on various RL algorithms for tasks involving manipulation by a manipulator. In their study, the Twin Delayed Deep Deterministic Policy Gradient (TD3), Soft Actor Critic (SAC), and TRPO algorithms emerge as leaders. Additionally, the use of the Hindsight Experience Replay (HER) strategy to enrich the internal reward signal in off-policy algorithms (SAC and TD3) slightly reduces training stability (repeatability and sampling efficiency). However, it enables the agent to learn an effective action policy in environments with goal initialization at random coordinates. The authors also encountered challenges when transferring the agent from the simulator to reality due to differences in dynamics (the sim-to-real problem). A real robot is subject to noise and non-stationarity, but partial training on a real specific robot allows the policy to adapt to the real world.

The authors of [22] conducted a similar study of algorithms in two other

simulation-based robotics environments with vision-based control, KukaDiverseObjectEnv and RaceCarZedGymEnv, where RGB images are used as observations and the action space is continuous. In these tasks, the interpolated policy gradients (IPG) algorithm demonstrated the highest efficiency, and the use of the HER strategy, similar to the authors of the previous paper, significantly improved the results.

Author [23] investigated various approaches for training agents to lift objects, press buttons, and control multi-fingered hands. The research concluded that the original PPO effectively solves basic grasping tasks. A combined approach (state representation learning (SRL) + RL) allows to solve more complex button-pressing tasks by extracting useful knowledge from pre-training states. Imitation learning (IL) is effective for training robots in complex multi-fingered manipulations by imitating human actions.

Authors [24] propose a method of continuous curriculum learning based on Precision-Based Continuous Curriculum Learning (PCCL) to accelerate reinforcement learning (RL) in multi-goal tasks. The study revealed that PCCL enhances learning efficiency and algorithm performance in RL, especially in scenarios with sparse and binary rewards.

The authors of the paper [25] focused on the issue of safety in agent learning. Due to the challenges in modeling certain forms of behavior (such as human interaction, real-world traffic scenarios, etc.), agent training is often transferred to the real world, where safety issues are much more critical. During environment exploration, the agent may perform actions that are dangerous to others. To address this problem, the authors utilized the PPO algorithm with Lagrangian constraints. As a result, the algorithm demonstrated longer training times but ultimately achieved the same effectiveness as regular PPO, while ensuring greater safety in its actions for the surrounding environment.

The authors of [5] developed the TossingBot system based on recurrent neural networks to train a robot manipulator to grasp and throw arbitrary objects beyond the robot's reachable area. The agent simultaneously learns strategies for grasping

and throwing. A key element is the use of «Residual Physics» - a hybrid controller that applies machine learning to predict residual errors over control parameters computed using a physical model. This approach allows focusing the training on aspects of dynamics that are difficult to model analytically.

In a later work, the authors [6] developed an approach for controlling robot manipulators based on visual information, enabling the solution of tasks such as:

- building a pyramid with blocks;
- assembling sets with new, previously unseen objects;
- moving piles of small items using closed-loop feedback.

The work employs several neural network models for segmentation and object recognition, with the primary one being the Transporter Network based on attention. This model generates robot motion trajectories and predicts object displacement in space for specific robot movements. It is trained on a dataset consisting of images of objects and their corresponding displacements.

Overall, the authors of this paper anticipate that this approach will enable solving more complex tasks, such as real-time robot control at high frequencies or using tools.