ОТЗЫВ

официального оппонента доктора физико-математических наук, Визильтера Юрия Валентиновича на диссертацию Козлова Даниила Александровича на тему «Интеграция иерархических ансамблей и трансформерных архитектур в алгоритмы обучения с подкреплением», представленную на соискание ученой степени кандидата технических наук по специальности 1.2.1 – Искусственный интеллект и машинное обучение

1. Актуальность темы диссертации

В условиях динамичного развития робототехники и повышения сложности задач, решаемых автономными системами, управление роботами в сложных и изменяющихся средах приобретает особую научную и практическую значимость. Актуальность данного направления усиливается на фоне активного применения методов обучения с подкреплением, которые открывают новые возможности для создания адаптивных систем управления, способных эффективно функционировать в условиях ограниченной априорной информации о внешней среде.

Диссертационная работа посвящена актуальной задаче интеграции методов обучения с подкреплением с современными архитектурами искусственного интеллекта, в частности, трансформерам. Выбранное направление исследований соответствует тенденциям развития науки и техники, что подтверждается количеством современных научных публикаций и исследований в данной области. Также решается задача иерархического ансамблирования алгоритмов обучения с подкреплением. В работе представлены решения, направленные на повышение эффективности алгоритмов обучения с подкреплением при их применении в сложных трехмерных средах, что является важным вкладом в развитие методологии управления роботами.

Таким образом, исследование, ориентированное на разработку новых методов и алгоритмов обучения с подкреплением, их интеграцию с передовыми архитектурами искусственного интеллекта, представляется не только актуальным, но и значимым. Работа обладает теоретической и практической ценностью, вносит существенный вклад в решение как фундаментальных, так и прикладных задач обучения с подкреплением, подтверждая свою важность для дальнейшего развития этой области.

2. Структура и объём диссертации

Диссертация Козлова Д.А. состоит из введения, четырех глав, заключения и списка литературы.

Во введении обоснована актуальность работы, сформулированы ее цель и задачи. Обосновывается соответствие диссертации паспорту научной специальности. Приводятся перечень научных результатов, научная новизна, теоретическая и практическая значимость, положения, выносимые на защиту.

В первой главе диссертации представлены основные теоретические основы и концепции обучения с подкреплением, включая марковские процессы, стратегии и функции ценности, а также проведен обзор существующих методов данной области. Рассмотрена классификация алгоритмов на модельные и безмодельные, а также на подходы, основанные на значении и стратегии, с акцентом на безмодельные методы как наиболее подходящие для решения задач высокой вычислительной сложности. Подробно описаны принципы работы и сравнительные характеристики ключевых алгоритмов, таких как Q-learning, DQN, PPO, DDPG, Actor-Critic и другие, с анализом их преимуществ и недостатков. Также приведены примеры успешного применения алгоритмов в реальных задачах, включая проектирование сред и функции награды, что позволяет обосновать выбор подхода для решения задач автономного передвижения.

Во второй главе диссертации представлены результаты экспериментального анализа алгоритмов обучения с подкреплением, направленного на изучение их применимости для задач передвижения в трехмерном пространстве и влияния компонентов наблюдений среды на качество обучения. Описаны используемые программные инструменты и симуляторы (Unity ML-Agents, dm-control, gymnasium), что обеспечивает воспроизводимость экспериментов. Проведен сравнительный анализ алгоритмов DQN, SAC, PPO и REDQ в различных средах, включая Hopper, Walker, Ant, Half-Cheetah и Humanoid, с акцентом на производительность и выборочную эффективность. Установлено, что SAC демонстрирует наилучшие результаты в условиях сложной координации, а выбор алгоритма зависит от характеристик среды. Отдельное внимание уделено влиянию состава наблюдений среды на результаты обучения: увеличение объема информации не всегда улучшает качество решения, что подтверждено экспериментально. Разработана методика оценки вклада отдельных признаков среды в итоговую производительность алгоритма, что подчеркивает важность предварительного анализа наблюдений.

В третьей главе предложена модель интеграции алгоритмов обучения с подкреплением и кодировщика трансформера для обработки последовательностей состояний. Вместо традиционного марковского предположения модель учитывает последовательность предыдущих состояний, что достигается использованием трансформера для формирования латентных представлений среды. На основе этой модели разработан алгоритм, объединяющий трансформер и Soft Actor-Critic (SAC), где веса кодировщика обучаются сквозным образом. Для реализации введены модификации метода воспроизведения опыта, позволяющие выбирать цепочки эпизодов. Алгоритм протестирован в средах MuJoCo, демонстрируя улучшение выборочной эффективности и адаптации в большинстве случаев, хотя результаты варьируются в зависимости от среды. Средний прирост награды составил 18,5%, а 80% экспериментов показали улучшение или сопоставимость с исходным алгоритмом SAC.

В четвертой главе предложен ансамблевый метод обучения с подкреплением с иерархической структурой, включающей управляющий и управляемые алгоритмы. Управляемые алгоритмы предсказывают действия, а управляющий выбирает, какое из них выполнит агент. Все алгоритмы обучаются на каждом шаге независимо от их активации, что позволяет эффективно распространять опыт. Метод протестирован с REDQ и SAC в качестве управляемых алгоритмов и DQN как управляющего. Он показал способность улучшать среднюю суммарную награду на 2,65% и обеспечил сопоставимость или улучшение результатов по сравнению с лучшим алгоритмом ансамбля во всех средах. Такой подход способствует эффективному обучению и адаптации в разнообразных задачах.

Материалы диссертации позволяют получить общее представление о проведенном исследовании, его объеме и уровне сложности. Работа написана понятным и грамотным языком, дополнена достаточным количеством иллюстративного материала. Автореферат соответствует основным требованиям и отражает основные положения диссертации. В представленных материалах практически не встречаются орфографические и технические ошибки.

3. Научная новизна полученных результатов

Научной новизной обладают следующие, полученные автором диссертационного исследования, результаты:

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

4. Значимость полученных результатов для науки и производства

Результаты, представленные в диссертации, обладают высокой научной и практической значимостью.

Разработанные методы и алгоритмы направлены на повышение качества обучения с подкреплением при решении задач управления роботами в сложных трехмерных средах. Их научная значимость заключается в том, что они позволяют обеспечить более высокую адаптивность и автономность робототехнических систем за счет использования современных подходов, таких как интеграция алгоритмов обучения с подкреплением и трансформеров, а также иерархическое ансамблирование.

Практическая ценность диссертационной работы заключается в расширении возможностей применения алгоритмов обучения с подкреплением для робототехнических систем, действующих в условиях высокой неопределенности. Внедрение предложенных решений способствует значительному улучшению точности и эффективности управления роботами без необходимости детального моделирования среды. Это особенно актуально для таких областей, как автономные транспортные средства, роботизированные манипуляторы, системы поиска и спасения, а также управление беспилотными летательными аппаратами.

5. Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

Положения, выводы и рекомендации, представленные в диссертации, основываются на современных достижениях в области искусственного интеллекта, машинного обучения и обучения с подкреплением. Соискатель грамотно применяет математический аппарат и методы машинного обучения, включая обучение с подкреплением и трансформерные архитектуры, для решения сложных задач управления роботами. Научная обоснованность результатов определяется корректным использованием математических методов и подтверждается результатами численных экспериментов.

Диссертация опирается на широкий спектр фундаментальных и прикладных исследований. Полученные результаты согласуются с современными тенденциями в области робототехники и искусственного интеллекта и подтверждаются как теоретическим анализом, так и практической реализацией. Представленные научные достижения соответствуют п. 6 и п. 17 паспорта специальности 1.2.1 — «Искусственный интеллект и машинное обучение».

6. Замечания по работе

Имеются следующие замечания по содержанию диссертации.

1. Разработанная методика оценки влияния состава набора наблюдений на качество решений, принимаемых агентом, представляется несколько упрощенной. Формула оценки влияния канала наблюдения на среднюю величину награды, приведенная в разделе 2.4.3 диссертации, является линейной и не учитывает такие важные факторы как апостериорная информативность каждого канала и общая размерность входного вектора наблюдений. В частности, неочевидной представляется интерпретация результатов экспериментов, показывающих, что полный набор наблюдений (эксперимент 4) обеспечивает худшую результативность обучения по сравнению с частичным набором тех же наблюдений (эксперимент 6). Представляется, что введение некоторого преобразования, понижающего размерность входного вектора (выделение главных компонент, кодировщик автоэнкодера, кодировщик трансформера и т.п.), могло бы обеспечить улучшение результатов в эксперименте 4. Интересно было бы провести такие эксперименты, чтобы точнее определить природу наблюдаемого явления.

2. Разработанный алгоритм, интегрирующий кодировщик трансформера и алгоритм обучения с подкреплением Soft Actor-Critic (SAC), действительно дает существенное преимущество по сравнению с исходным алгоритмом SAC. С учетом этого, результат эксперимента в среде humanoid, в котором данный поход в принципе не сработал, кажется не вполне адекватным. Возможно, проблема заключается в сложности и высокой размерности входных данных (см. предыдущее замечание). Путем решения этой проблемы могли бы стать такие технические приемы как предобучение кодировщика трансформера методом обучения без учителя (offline SSL) или снижение размерности выходного вектора кодировщика. Желательно было бы исследовать эти или другие подходы с тем, чтобы установить причину того, что метод, показавший существенное улучшение в других средах, в данном эксперименте не сработал.

3. Чрезвычайно интересным и перспективным представляется разработанный автором алгоритм обучения с подкреплением на основе предложенного метода иерархического ансамблирования с использованием алгоритма DQN в качестве управляющего и алгоритмов SAC и REDQ в качестве управляемых. Однако кажется, что сходимость обучения для данного метода ансамблирования не гарантирована, поскольку в методе не предусмотрен способ предотвращения случаев, когда один из алгоритмов в течение долгого времени не участвует в формировании финального решения. При этом его политика будет обновляться так, как будто он эти решения принимает, что не соответствует условиям доказанных свойств каждого из методов. Возможное решение заключается в добавлении механизма принудительного выбора решения каждого из методов как финального не реже одного раза в некоторое установленное количество итераций. Это даст определенную гарантию обеспечения минимально необходимого компромисса между исследованием новых состояний и эксплуатацией уже имеющегося успеха (exploration/exploitation tradeoff). Желательно было бы такую модификацию алгоритма ансамблирования произвести, а также получить теоретические и экспериментальные оценки сходимости метода обучения с подкреплением для такой схемы ансамблирования.

Данные замечания носят рекомендательный характер и направлены на развитие темы диссертационной работы. Замечания не снижают общую высокую оценку полученных теоретических и практических результатов.

7. **Заключение**

Диссертационная работа содержит решение актуальной научной задачи в области обучения с подкреплением. Диссертация удовлетворяет требованиям «Положения о присуждении ученых степеней», соответствует паспорту специальности 1.2.1 - Искусственный интеллект и машинное обучение, а ее автор, Козлов Даниил Александрович, заслуживает присуждения ему ученой степени кандидата технических наук по заявленной специальности.

Официальный оппонент,
начальник подразделения
ФАУ «ГосНИИАС»,
д.ф.-м.н., профессор РАН Визильтер Юрий Валентинович

« 03 » ноября 2024 г.

Подпись Визильтера Ю. В. заверяю
Заместитель генерального директора
по науке, академик РАН, профессор, д.т.н. Желтов Сергей Юрьевич

« 03 » ноября 2024 г.

Визильтер Юрий Валентинович – доктор физико-математических наук по специальности 05.13.17, профессор РАН, начальник подразделения Федерального автономного учреждения "Государственный научно-исследовательский институт авиационных систем" (ФАУ "ГосНИИАС").

г. Москва, ул. Викторенко, 7, тел.: (499) 157-94-98, e-mail: viz@gosniias.ru.