8. Анализ современного состояния рынка труда и производительности труда Кыргызской Республики. Доклад. Бишкек 2019.
9. Конвенция о защите физических лиц при автоматизированной обработке персональных данных (СЕД № 108) Страсбург 28.01.1981
10. Определение Судебной коллегии по гражданским делам Верховного Суда Российской Федерации от 14.07.2020 N 58-КГ20-2
11. Государственный портал электронных услуг Кыргызской Республики (https://portal.tunduk.kg/)

Рецензент: доктор юридических наук, профессор Колсариева Н.Ш.

УДК 004.89
DOI 10.58649/1694-9099-2026-2-284-290

Ян Хаоцзе, Бексултанов Ж.Т., Ван Цун
Жусуп Баласагын атындагы КУУ
Ян Хаоцзе, Бексултанов Ж.Т., Ван Цун
КНУ имени Жусупа Баласагына
Yang Haojie, Beksultanov Zh.T., Wang Cong
KNU Jusup Balasagyn

КҮЧӨТҮЛГӨН ОКУТУУГА НЕГИЗДЕЛГЕН «ЖЫЛАНЧА» ОЮН АГЕНТИН ДОЛБООРЛОО ЖАНА АЛГОРИТМДЕРДИ САЛЫШТЫРМАЛУУ ТАЛДОО
ПРОЕКТИРОВАНИЕ ИГРОВОГО АГЕНТА «ЗМЕЙКА» НА ОСНОВЕ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ И СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ
DESIGN OF A SNAKE GAME AGENT BASED ON REINFORCEMENT LEARNING AND COMPARATIVE STUDY OF ALGORITHMS

Кыскача мүнөздөмө: Күчөтүлгөн окутуу ырааттуу чечим кабыл алуу жана интеллектуалдык башкаруу тармагында, өзгөчө аракеттердин дискреттик мейкиндиги бар оюн чөйрөлөрүндө олуттуу артыкчылыктарды көрсөтүп келет. Бул макалада Gym интерфейсинин негизинде «Жыланча» оюн чөйрөсү түзүлүп, Stable-Baselines3 жана PyTorch фреймворкторун колдонуу аркылуу күчөтүлгөн окутуунун төрт алгоритми - DQN, A2C, PPO жана MPPOнун жакшыртылган версиясы ишке ашырылган. Андан соң бирдей эксперименталдык шарттарда алгоритмдердин салыштырма талдоосу жүргүзүлгөн. Эксперименталдык платформа Intel i5-10300H процессорунан, 24 ГБ оперативдүү эс тутумдан жана NVIDIA GTX1650 видеокартасынан (4 ГБ видео эс тутуму менен) турган. Ар бир алгоритм 10 000 000 окутуу кадамынан өткөн. Дискреттик башкаруу тапшырмаларында төрт алгоритмдин иштеши оюндун узактыгынын өзгөрүшүн, сыйлыкка ылайык келүүчү конвергенция ийри сызыктарын, акыркы упайларды жана окутуунун туруктуулугун талдоо аркылуу бааланган. Эксперименталдык жыйынтыктар MPPO алгоритми конвергенция ылдамдыгы, акыркы упай жана туруктуулук көрсөткүчтөрү боюнча башка алгоритмдерден артыкчылыкка ээ экенин көрсөттү. Жалпы аткаруу рейтинги төмөнкүдөй аныкталган: MPPO > A2C > PPO > DQN. Алынган жыйынтыктар орто масштабдагы дискреттик чечим кабыл алуу чөйрөлөрүндө күчөтүлгөн окутуу алгоритмдерин тандоо жана түзүмдүк жактан өркүндөтүү үчүн эксперименталдык негиз болуп кызмат кылат.

Аннотация: Обучение с подкреплением продемонстрировало значительные преимущества в задачах последовательного принятия решений и интеллектуального управления, особенно в игровых средах с дискретным пространством действий. В данной работе на основе интерфейса Gym создана игровая среда «Змейка» и с использованием фреймворков Stable-Baselines3 и PyTorch реализованы четыре алгоритма обучения с подкреплением: DQN, A2C, PPO и улучшенная версия MPPO. Затем

проведено сравнение алгоритмов в единых экспериментальных условиях. Экспериментальная платформа включала процессор Intel i5-10300H, 24 ГБ оперативной памяти и видеокарту NVIDIA GTX1650 с 4 ГБ видеопамяти. Каждый алгоритм прошёл 10 000 000 шагов обучения. Производительность четырёх алгоритмов в задачах дискретного управления оценивалась путём анализа изменения длительности игры, кривых сходимости вознаграждения, итоговых результатов и стабильности обучения. Экспериментальные результаты показывают, что MPPO превосходит другие алгоритмы по скорости сходимости, итоговому результату и стабильности. Общий рейтинг производительности выглядит следующим образом: MPPO > A2C > PPO > DQN. Полученные результаты служат экспериментальным основанием для выбора и структурного улучшения алгоритмов обучения с подкреплением в средах принятия дискретных решений среднего масштаба.

Abstract: Reinforcement learning has demonstrated significant advantages in sequential decision-making and intelligent control, particularly in game environments with discrete action spaces. In this study, a Snake game environment was developed based on the Gym interface, and four reinforcement learning algorithms-DQN, A2C, PPO, and an improved version of MPPO-were implemented using the Stable-Baselines3 and PyTorch frameworks. Subsequently, a comparative analysis of the algorithms was conducted under unified experimental conditions. The experimental platform consisted of an Intel i5-10300H processor, 24 GB of RAM, and an NVIDIA GTX1650 graphics card with 4 GB of VRAM. Each algorithm underwent 10 000 000 training steps. The performance of the four algorithms in discrete control tasks was evaluated by analyzing game duration variations, reward convergence curves, final scores, and training stability. Experimental results indicate that MPPO outperforms the other algorithms in terms of convergence speed, final score, and stability. The overall performance ranking was determined as follows: MPPO > A2C > PPO > DQN. The obtained results provide an experimental basis for the selection and structural improvement of reinforcement learning algorithms in medium-scale discrete decision-making environments.

Негизги сөздөр: күчөтүлгөн окутуу; чектелген Марковдук чечим кабыл алуу процесси; аракеттерди тандоону чектөө; жакынкы саясатты оптималдаштыруу; дискреттик башкаруу.

Ключевые слова: обучение с подкреплением; Марковский процесс принятия решений с ограничениями; маскирование действий; проксимальная оптимизация политики; дискретное управление.

Keywords: reinforcement learning; constrained Markov decision process; action masking; proximal policy optimization; discrete control.

В последние годы глубокое обучение с подкреплением достигло значительных успехов в решении сложных задач принятия решений. С появлением глубоких Q-сетей (DQN), обеспечивших уровень управления, сопоставимый с человеческим, сочетание глубоких нейронных сетей и обучения с подкреплением стало одним из основных направлений исследований. В дальнейшем методы градиента политики и фреймворк «актор - критик» постоянно совершенствовались, демонстрируя более высокую стабильность обучения как в задачах непрерывного управления, так и в задачах дискретного принятия решений.

Игровые среды благодаря чётким правилам, хорошо определённым метрикам оценки и возможности повторения экспериментов широко используются для проверки алгоритмов и проведения сравнительного анализа. Несмотря на простоту правил игры «Змейка», её пространство состояний непрерывно изменяется по мере роста змейки, процесс принятия решений характеризуется долгосрочными зависимостями, а распределение вознаграждений отличается значительной разреженностью. Поэтому данная задача сочетает в себе сложность исследования и проблему стабильности политики, представляя собой типичную задачу дискретного управления для оценки эффективности алгоритмов обучения с подкреплением.

При параллельном развитии методов оценки функции ценности и методов градиента политики различные алгоритмы по-прежнему демонстрируют различия в производительности в средах средней

сложности [2]. Для систематического анализа характеристик производительности разных алгоритмов в среде дискретного принятия решений в данной работе проводится экспериментальное сравнение DQN (Deep Q-learning), A2C (Advantage Actor-Critic), PPO (Proximal Policy Optimization) и улучшенного алгоритма MPPO (Multi-Path Policy Optimization) в рамках единой структуры обучения и при одинаковых гиперпараметрах. Также анализируются их сходимость и обобщающая способность.

Основная работа в рамках данной статьи включает:

1. построение масштабируемой игровой среды «Змейка» на основе стандартного интерфейса Gym, а также проектирование векторизованного представления состояний и механизма вознаграждения;
2. реализацию алгоритмов DQN, A2C, PPO и многопутевого алгоритма MPPO с использованием библиотеки Stable-Baselines3;
3. проведение сравнительных экспериментов при едином аппаратном обеспечении и одинаковом масштабе обучения, а также анализ различий в производительности алгоритмов по многомерным показателям;
4. объяснение причин различий в производительности с точки зрения способности к исследованию, стабильности обновления и разнообразия политики.

Таблица 1. Аппаратное и программное обеспечение экспериментальной платформы

CPU	Intel i5-10300H
Память	24GB
GPU	NVIDIA GTX1650 (4GB VRAM)
Фреймворк	PyTorch
Библиотека обучения с подкреплением	Stable-Baselines3
Инструмент визуализации	TensorBoard

Обзор алгоритмов обучения с подкреплением

1. Алгоритм *DQN*. DQN относится к методам аппроксимации функции ценности, оценивающему функцию ценности действия Q(*s*,*a*) с помощью нейронной сети. Его основная формула обновления:

$$Q(s,a) = r + \gamma a' max Q'(s',a')$$

Данное уравнение, которое называют уравнением Беллмана, утверждает что значение Q для определенной пары состояние-действие должно быть наградой r, полученной при переходе в новое состояние (путем выполнения этого действия), добавленной к значению Q наилучшего действия в следующем состоянии с каким-то дисконтом [1].

Он вводит механизм воспроизведения опыта для устранения корреляции между выборками и смягчает колебания обучения с помощью целевой сети. В дискретных пространствах действий DQN имеет преимущества простоты реализации и высокой утилизации выборок. Однако в задачах принятия долгосрочных решений переоценка Q-значений и ошибки аппроксимации функции могут привести к нестабильности обучения, особенно в средах с разреженными вознаграждениями, где эффективность исследования относительно ограничена.

2. Алгоритм *A2C*. A2C (Advantage Actor-Critic) - синхронный метод «Актор-Критик». Он выводит распределение действий через политическую сеть и оценивает функцию ценности состояния через сеть ценности. Его обновление опирается на функцию преимущества:

$$A(s,a) = Q(s,a) - V(s)$$

для снижения дисперсии градиента. A2C имеет простую структуру и стабильные обновления, обычно демонстрируя хорошие характеристики сходимости в задачах малого и среднего масштаба [4].

3. Алгоритм *PPO*. PPO (Proximal Policy Optimization) вводит функцию цели с обрезкой в рамках метода градиента политики.

$$L_{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right]$$

Этот механизм ограничивает величину обновления политики, предотвращая снижение производительности из-за слишком сильного изменения политики. PPO отлично работает в областях непрерывного управления, но его преимущества в низкоразмерных дискретных пространствах зависят от сложности задачи [3].

4. Алгоритм *MPPO.* Для повышения способности к исследованию в данной работе разработана структура с несколькими ветвями политики на основе PPO. MPPO использует многоголовую структуру сети Actor, где каждая ветвь политики независимо выводит распределения вероятностей действий и совместно использует единую сеть Critic для оценки ценности.

На этапе обучения несколько политик параллельно выбирают траектории в среде; на этапе обновления градиенты от разных политик объединяются. Этот механизм повышает разнообразие выборок, снижает риск попадания единой политики в локальный оптимум, тем самым улучшая эффективность исследования и долгосрочную производительность.

Проектирование среды и настройка эксперимента. Построение среды игры «Змейка». Среда реализована в соответствии со спецификацией интерфейса Gym и включает функции reset, step и render. Пространство состояний использует векторизованное кодирование и в основном включает следующие параметры:

– координаты головы змейки;

– координаты еды;

– относительное положение тела змейки;

– текущее направление движения.

Пространство действий представлено четырёхмерным дискретным набором действий: вверх, вниз, влево, вправо**.**

Функция вознаграждения разработана следующим образом:

– за поедание еды начисляется вознаграждение **(**+10**)**;

– при столкновении назначается штраф **(**–10);

– за каждый выполненный шаг применяется штраф **(**–0,01).

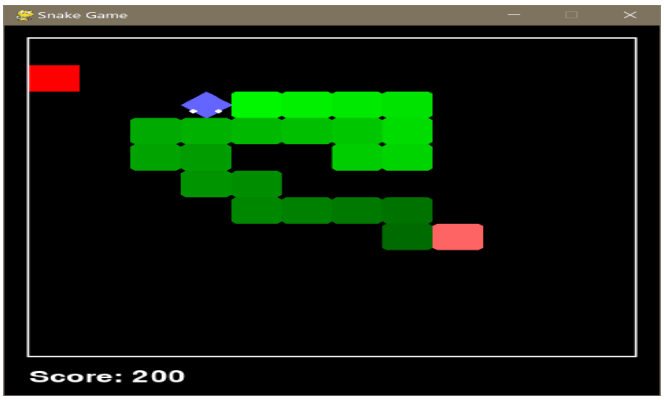


Рисунок 1. Игра «Змейка»

Кроме того, введен простой механизм маскировки действий, предотвращающий выполнение агентом прямых обратных действий, что снижает долю неэффективных исследований. Эта среда может рассматриваться как ограниченный процесс Марковского принятия решений [5; 6].

Метрики оценки. В данной работе используются следующие показатели для комплексной оценки:

1. Длина эпизода;

2. Кривая сходимости среднего вознаграждения;

3. Средний балл на финальном этапе тестирования.

Результаты и анализ.

1. Анализ изменения длительности игры. По данным записей обучения, длительность игры для алгоритма DQN увеличивается относительно медленно; значительное улучшение наблюдается только после середины периода обучения. Алгоритмы A2C и PPO демонстрируют стабильную восходящую динамику, причём у A2C рост выражен сильнее.

MPPO быстро достигает высокой длительности игры уже на ранних этапах обучения, после чего наблюдаются небольшое снижение и последующая стабилизация. Это указывает на то, что многополитическая структура обладает высокой исследовательской способностью на начальных этапах, однако на поздней фазе объединения политик проходит определённый процесс сходимости и корректировки.

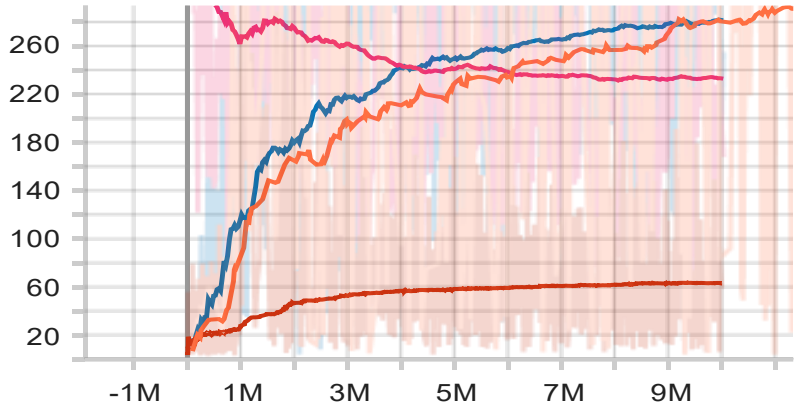


Рисунок 2. Длительность игры за эпизод в зависимости от времени при разных алгоритмах

2. Анализ тенденции сходимости вознаграждения. По кривым вознаграждения можно отметить, что DQN обучается медленно на начальных этапах, демонстрирует поэтапное улучшение в середине периода обучения, однако характеризуется значительными общими колебаниями. A2C показывает стабильный рост с умеренными колебаниями. PPO увеличивает показатели медленно, но в целом остаётся стабильным. MPPO быстро улучшает результаты на ранних этапах обучения и в итоге достигает наивысшего стабильного уровня вознаграждения.

Высокий прирост вознаграждения у MPPO на ранних этапах свидетельствует о том, что разнообразие политик эффективно повышает исследовательскую способность алгоритма.

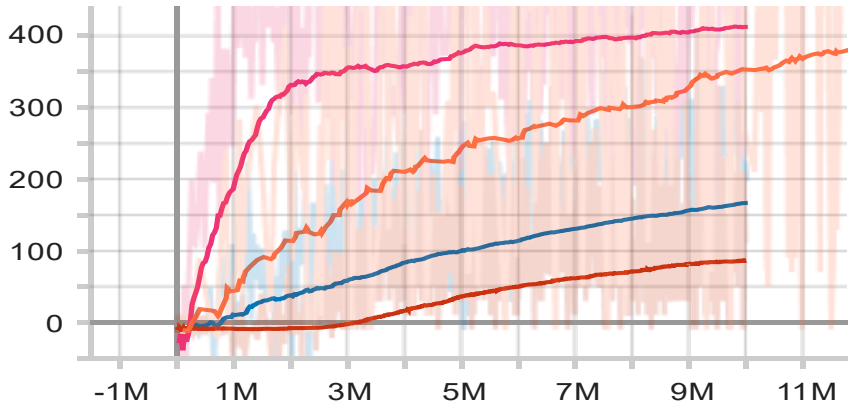


Рисунок 3. Очки игры за эпизод в зависимости от времени при разных алгоритмах

3. Сравнение игровых баллов. После 10 000 000 шагов обучения все четыре алгоритма были независимо протестированы и оценены. Получены следующие результаты: MPPO - 212 очков; A2C - 186 очков; PPO - 97 очков; DQN - 48 очков.

MPPO значительно превзошёл остальные алгоритмы, продемонстрировав более высокую способность к долгосрочному набору очков. A2C показал стабильную общую производительность и

превысил результаты PPO и DQN. DQN получил самый низкий балл и продемонстрировал значительные колебания во время тестирования.

Обсуждение. Экспериментальные результаты показывают, что различия в производительности алгоритмов обучения с подкреплением в задачах дискретного управления тесно связаны с механизмами их обновления.

DQN как метод функции ценности склонен к накоплению ошибок аппроксимации Q-значений в средах, где пространство состояний динамически расширяется по мере роста тела змейки. Кроме того, относительно разреженное распределение вознаграждений ограничивает его эффективность при долгосрочном принятии решений.

В отличие от DQN, A2C и PPO снижают влияние ошибок оценки ценности на обновление политики за счёт прямой оптимизации распределения политики и демонстрируют более стабильную общую производительность. В рамках данного экспериментального масштаба A2C превосходит PPO. Это, вероятно, связано с меньшей размерностью пространства действий и ограничением амплитуды обновления механизмом отсечения PPO, вследствие чего преимущества PPO реализуются не полностью.

MPPO достигает лучших результатов главным образом потому, что несколько ветвей политики повышают разнообразие исследования и увеличивают покрытие выборок. При этом совместно используемая сеть Critic обеспечивает стабильность оценки ценности и снижает вероятность попадания в локальные оптимумы.

В целом повышение разнообразия политики может эффективно улучшить производительность обучения с подкреплением в задачах дискретного принятия решений среднего масштаба. Однако данные выводы требуют дальнейшей проверки в средах большего масштаба и при различных случайных начальных значениях.

Заключение. Данная работа посвящена задаче дискретного управления в игре «Змейка». В исследовании построена экспериментальная среда обучения с подкреплением на основе интерфейса Gym. При единой структуре сети, одинаковых гиперпараметрах и масштабе обучения, составляющем 10 000 000 шагов, проведено систематическое сравнение четырёх алгоритмов: DQN, A2C, PPO и улучшенного MPPO.

На основе анализа таких показателей, как изменение длительности игры, тенденция сходимости вознаграждения и итоговый тестовый балл, проведена комплексная оценка производительности алгоритмов в среде дискретного принятия решений среднего масштаба.

Результаты эксперимента показывают, что методы градиента политики в целом превосходят методы функции ценности. При этом MPPO демонстрирует наилучшую производительность, за ним следует A2C, тогда как PPO и DQN показывают относительно более слабые результаты. Итоговый рейтинг производительности выглядит следующим образом: MPPO > A2C > PPO > DQN**.**

Полученные результаты свидетельствуют о том, что в задачах с динамически изменяющимся пространством состояний и долгосрочными зависимостями методы функции ценности, основанные исключительно на оценке Q-значений, подвержены влиянию ошибок аппроксимации и смещению переоценки. Фреймворк «актор - критик» за счёт прямой оптимизации распределения политики в определённой степени повышает стабильность обучения и эффективность сходимости.

MPPO за счёт дополнительного введения структуры с несколькими ветвями политики повышает эффективность исследования на ранних этапах обучения и сохраняет высокую стабильность, а также способность к набору очков на поздних этапах. Это достигается благодаря увеличению разнообразия политики и расширению покрытия выборок, что подтверждает эффективность многополитического механизма в задачах дискретного управления.

Таким образом, существует тесная связь между проектированием структуры алгоритма и характеристиками задачи. В средах дискретного управления среднего масштаба введение разнообразия политики и рационального механизма оценки ценности может значительно повысить производительность обучения с подкреплением.

В будущих исследованиях целесообразно дополнительно проверить и расширить полученные выводы с помощью статистического анализа при разных случайных начальных значениях, более сложных структурах кодирования состояний и средах большего масштаба. Это позволит обеспечить более универсальную теоретическую и экспериментальную основу для применения алгоритмов обучения с подкреплением в сложных задачах дискретного принятия решений.

Список использованной литературы

1. De La Fuente N., Guerra D.A.V. A comparative study of deep reinforcement learning models: DQN vs PPO vs A2C // arXiv preprint. – 2024. – arXiv:2407.14151.
2. Ghasemi M., Moosavi A.H., Ebrahimi D. A comprehensive survey of reinforcement learning: From algorithms to practical challenges // arXiv preprint. – 2024. – arXiv:2411.18892.
3. Terven J. Deep reinforcement learning: A chronological overview and methods // AI. – 2025. – Vol. 6, No. 3. – Article 46.
4. Cui Z., Deng K., Zhang H. et al. Deep reinforcement learning-based multi-agent system with advanced actor–critic framework for complex environment // Mathematics. – 2025. – Vol. 13, No. 5. – P. 754.
5. Jiang S., Salagame A., Ramezani A. et al. Hierarchical RL-guided large-scale navigation of a Snake robot // Proceedings of the 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). – IEEE, 2024. – P. 1347–1352.
6. Zhang D., Ju R., Cao Z. Reinforcement learning-based motion control for snake robots in complex environments // Robotica. – 2024. – Vol. 42, No. 4. – P. 947–961.

Рецензент: кандидат технических наук, и.о.доцента Керимов У.Т.