УДК 004.85
DOI: 10.26102/2310-6018/2025.50.3.029

Гибридная система обучения агентов с использованием A2C и эволюционных стратегий

А.П. Корчагин✉

Воронежский государственный университет, Воронеж, Российская Федерация

Резюме. Актуальность исследования обусловлена необходимостью повышения эффективности обучения агентов в условиях частичной наблюдаемости и ограниченного взаимодействия, характерных для многих реальных задач в мультиагентных системах. В связи с этим данная статья направлена на разработку и анализ гибридного подхода к обучению агентов, сочетающего преимущества градиентных и эволюционных методов. Ведущим методом исследования является модифицированный алгоритм Advantage Actor-Critic (A2C), дополненный элементами эволюционного обучения – кроссовером и мутацией параметров нейросети. Такой подход позволяет комплексно рассмотреть проблему адаптации агентов в условиях ограниченного обзора и кооперативного взаимодействия. В статье представлены результаты экспериментов в среде с двумя кооперативными агентами, задачей которых является извлечение и доставка ресурсов. Показано, что гибридная методика обучения обеспечивает значительный рост эффективности поведения агентов по сравнению с чисто градиентными подходами. Динамика среднего вознаграждения свидетельствует об устойчивости метода и его потенциале в более сложных сценариях многоагентного взаимодействия. Материалы статьи представляют практическую ценность для специалистов в области обучения с подкреплением, разработки мультиагентных систем и построения адаптивных кооперативных стратегий в условиях ограниченной информации.

Ключевые слова: обучение с подкреплением, эволюционные алгоритмы, многоагентная система, A2C, LSTM, кооперативное обучение.

Для цитирования: Корчагин А.П. Гибридная система обучения агентов с использованием A2C и эволюционных стратегий. *Моделирование, оптимизация и информационные технологии.* 2025;13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1991 DOI: 10.26102/2310-6018/2025.50.3.029

Hybrid agent training system using A2C and evolutionary strategies

A.P. Korchagin✉

Voronezh State University, Voronezh, the Russian Federation

Abstract. The relevance of the study is due to the need to increase the efficiency of agent training under conditions of partial observability and limited interaction, which are typical for many real-world tasks in multiagent systems. In this regard, the present article is aimed at the development and analysis of a hybrid approach to agent training that combines the advantages of gradient-based and evolutionary methods. The main method of the study is a modified Advantage Actor-Critic (A2C) algorithm, supplemented with elements of evolutionary learning — crossover and mutation of neural network parameters. This approach allows for a comprehensive consideration of the problem of agent adaptation in conditions of limited observation and cooperative interaction. The article presents the results of experiments in an environment with two cooperative agents tasked with extracting and delivering resources. It is shown that the hybrid training method provides a significant increase in the effectiveness of agent behavior compared to purely gradient-based approaches. The dynamics of the average reward confirm the stability of the method and its potential for more complex multiagent interaction scenarios. The materials of the article have practical value for specialists in the fields of reinforcement learning,

multi-agent system development, and the design of adaptive cooperative strategies under limited information.

Keywords: reinforcement learning, evolutionary algorithms, multiagent system, A2C, LSTM, cooperative learning.

For citation: Korchagin A.P. Hybrid agent training system using A2C and evolutionary strategies. *Modeling, Optimization and Information Technology*. 2025;13(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1991 DOI: 10.26102/2310-6018/2025.50.3.029

Введение

Современные задачи в области робототехники, интеллектуальных систем и автоматизации все чаще требуют взаимодействия нескольких агентов в общем пространстве. Такие мультиагентные системы находят применение в самых разных сферах – от оптимизации перемещения транспорта [1], распределения парковочных мест [2], разработки системы зарядки дронов для доставки товаров [3] до защиты инфраструктуры от дронов [4].

Одной из ключевых проблем при построении подобных систем является организация эффективного и устойчивого обучения агентов в условиях частичного наблюдения, ограниченной коммуникации и высокой динамики среды. В таких условиях каждый агент имеет доступ только к локальной информации и должен принимать решения, часто не располагая полной картиной происходящего.

Методы обучения с подкреплением (Reinforcement Learning, RL), такие как Deep Q-Network (DQN)[5], Proximal Policy Optimization (PPO) и Advantage Actor-Critic (A2C), успешно применяются в задачах с одним агентом и полной наблюдаемостью [6, 7]. Однако при масштабировании на многоагентные сценарии они сталкиваются с рядом проблем:

– рост нестабильности при обучении из-за одновременного обновления стратегий всех агентов;

– проблемы нестационарности среды (каждый агент влияет на среду для других агентов);

– ухудшение качества политики из-за ограниченного поля зрения и отсутствия глобальной координации.

Одним из перспективных направлений решения этих проблем является использование эволюционных стратегий (Evolution Strategies, ES). В отличие от градиентных методов, эволюционные подходы не требуют вычисления градиента функции вознаграждения, что делает их устойчивыми к шуму и локальным минимумам [8]. Кроме того, они позволяют реализовать глобальный поиск в пространстве параметров и обеспечить разнообразие поведения агентов за счет мутаций и кроссоверов.

В данной работе предлагается гибридный подход к обучению агентов в частично наблюдаемой мультиагентной среде, сочетающий:

– градиентное обучение на основе алгоритма A2C для локальной адаптации агента;

– эволюционные стратегии (кроссовер и мутация параметров нейросети) для глобального поиска и устойчивости обучения.

Такое объединение позволяет достичь баланса между точной настройкой поведения и поиском новых стратегий. В отличие от чисто градиентных методов, предложенный подход демонстрирует более стабильную динамику обучения, лучшую переносимость к другим задачам и повышенную кооперативность агентов, что критически важно для применения в реальных системах с ограниченной информацией.

Материалы и методы

Среда, используемая в данной работе, представляет собой квадратную сетку размером 15×15, на которой случайным образом размещаются ресурсы, агенты и цель. Агенты (всего двое) должны собирать ресурсы и доставлять их к цели. Каждый агент видит лишь часть окружающего пространства (локальная зона размером 5×5) и не имеет информации о действиях второго.

Агент получает:

– +0,5 за сбор ресурса;

– +1 за его успешную доставку до цели;

– −0,01 за каждый шаг.

Особенности среды:

– ограниченное наблюдение: агенты не видят друг друга и не имеют общей памяти;

– распределенное управление: каждый агент действует независимо, принимая решения на основе личного опыта;

– неполная обратная связь: награды выдаются не за каждое действие, а при достижении промежуточных целей (сбор ресурса, доставка к цели);

– стохастичность: расположение объектов каждый раз различное.

Таким образом, обучение требует учета последовательности состояний, способности планировать и эффективно использовать частичную информацию. Это делает использование рекуррентных сетей и продвинутых стратегий обновления параметров особенно актуальными.

Среда моделируется как частично наблюдаемый марковский процесс принятия решений (POMDP), определяемый:

$$\mathcal{M} = (S, A, P, R, \Omega, O, \gamma),$$

где S – множество состояний среды, $A = \{0,1,2,3\}$ – дискретное множество действий (вверх, вниз, влево, вправо), $P(s'|s,a)$ – функция перехода, $R(s,a)$ – функция вознаграждения (зависит от сбора ресурса и доставки его к цели), Ω – множество наблюдаемых агентами состояний, $O(o|s)$ вероятность наблюдения o в состоянии s, γ – коэффициент дисконтирования (в коде 0,99).

Каждое наблюдение агента i, в момент времени t, $o_t^i \in \mathbb{R}^{30}$ состоит из:

– 5 признаков: координат самого агента (x_{agent}, y_{agent}), координат цели (x_{goal}, y_{goal}) и булевого идентификатора наличия ресурса;

– 25 признаков: бинарная карта локальных (в пределах манхэтенского расстояния ≤ 2) ресурсов 5×5

$$o_t^i = \left[x_t^i, y_t^i, x_{goal}, y_{goal}, carry_i, flatten(R^{5\times5})\right] \in \mathbb{R}^{30}.$$

Политика каждого агента реализована нейросетевой моделью π_θ, параметризованной вектором θ, целью обучения агента является максимизация ожидаемого суммарного дисконтирования за эпизод

$$J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_t^T \gamma^t r_t],$$

где T – горизонт эпизода, в данном случае равный 200 шагам.

Политика каждого агента реализована нейросетевой моделью π_θ, параметризованной вектором θ, которая отображает наблюдение в распределение на действиях. Данная модель использует LSTM-блок для обеспечения учета временной зависимости наблюдений.

$$\pi_\theta: \Omega \times \mathbb{R}^H \to \Delta(A),$$

где $\Delta(A)$ – пространство вероятностных распределений на множестве действий, $H = 128$ – размер скрытого состояния LSTM (ячейки памяти). Входом сети является наблюдение $o_t^i \in \mathbb{R}^{30}$. [9]

Архитектура нейросетевой модели агентов представлена на Рисунке 1.

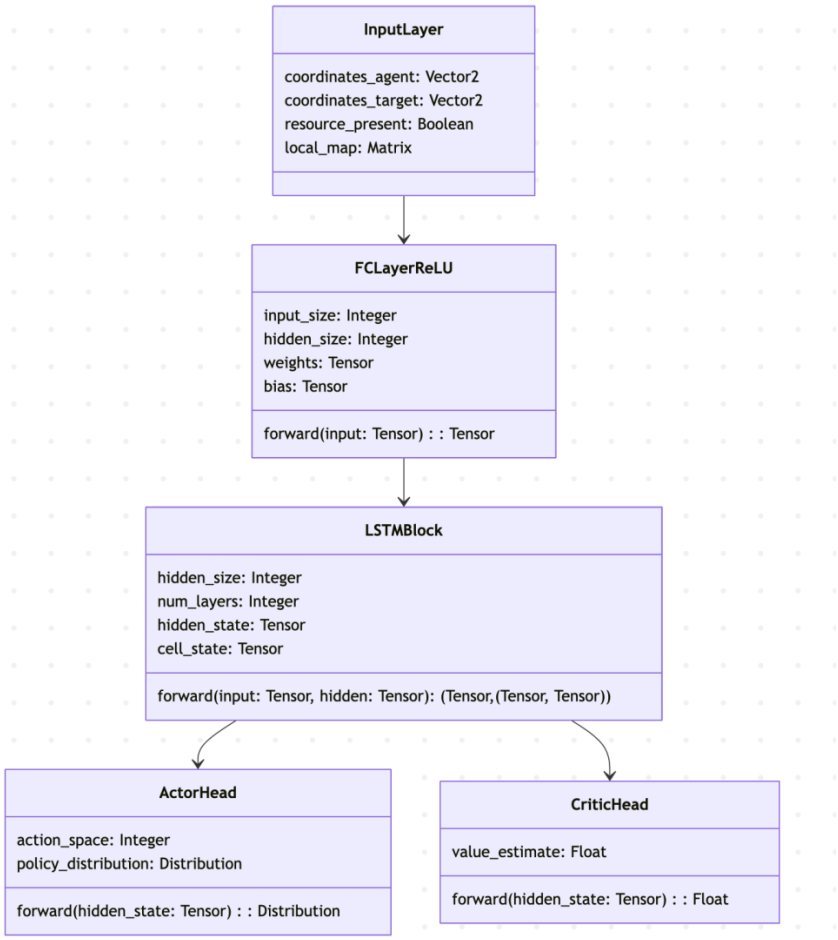

Рисунок 1 – Архитектура нейросетевой модели агентов
Figure 1 – Architecture of the neural network model of agents

Входной слой (InputLayer) включает координаты агента, координаты цели, флаг наличия ресурса и локальную карту ресурсов; затем данные обрабатываются скрытым слоем (FCReluLayer), который представляет собой полносвязный слой с функцией активации ReLU, затем LSTM-блок (LSTMBlock) используется для последовательной обработки для учета временной зависимости, затем обрабатываются раздельно Actor (ActorHead) и Critic (CriticHead), которые представляют собой две «головы», первая отвечает за политику (распределение действий, их оценка), вторая – за оценку ценности состояния.

Таким образом, сеть выдает оценку ценности состояния – $V(o_t) \in \mathbb{R}$, логиты действий $z_t \in \mathbb{R}^4$ и скрытое состояние $h_t, c_t \in \mathbb{R}^{1\times1\times\mathrm{H}}$.

В качестве функции потерь используется общая функция потерь, вычисляемая на основе функций потерь для Actor и Critic:

$$L_{total} = L_{actor} + \lambda_v \cdot L_{critic} - \lambda_e \cdot \mathcal{H}(\pi),$$

где L_{actor} – функция потерь для Actor, L_{critic} – функция потерь для Critic, $\mathcal{H}(\pi)$ – энтропия распределения действий, стимулирующая исследование, λ_v, λ_e – коэффициенты баланса, в коде они будут равны 0,5 и 0,01 соответственно.

Для Actor и Critic используется следующая функция потерь:

$$L_v = -log\ \pi_\theta(a_t|o_t) \cdot A_t,$$
$$L_{critic} = (R_t - V(o_t))^2,$$
$$A_t = R_t - V(o_t),$$
$$R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k},$$

где $\pi(a_t|o_t)$ – это вероятность выбора действия a_t при наблюдении o_t, получаемая непосредственно из оценки Actor для каждого действия, A_t – оценка преимущества, R_t – суммарное дисконтированное вознаграждение.

Энтропия распределения действий вычисляется следующим образом:

$$\mathcal{H}(\pi) = -\sum_{a \in A} \pi(a|o_t) \cdot log\ \pi(a|o_t).$$

Обновление параметров данной модели осуществляется методом стохастического градиентного спуска в модификации Adam [10].

В дополнение к градиентному обучению стратегии каждого отдельного агента в рамках метода A2C (Advantage Actor-Critic), предложенная система использует механизм эволюционного обновления популяции, обеспечивающий разнообразие стратегий и устойчивость к локальным минимумам. Такая гибридная схема сочетает в себе локальное обучение по градиенту и глобальный поиск в пространстве параметров, реализуемый через популяционную эволюцию.

Каждые N эпох (в данной N = 1000) выполняется операция замены текущей популяции агентов, состоящей из K (в данной реализации K = 8) экземпляров, на новую, сформированную на основе отбора, скрещивания (кроссовера) и мутации весов нейронных сетей. Основная метаэвристика повторяет идею генетических алгоритмов, но применяется непосредственно к параметрам нейросетевых моделей θ_i.

После завершения очередного этапа обучения (A2C) каждый агент получает интегральную оценку своей стратегии в виде суммарной награды за эпизоды (фитнес-функция):

$$F_i = \sum_{e=1}^{E} \sum_{t=0}^{T} r_t^{(i,e)},$$

где E – количество сыгранных эпизодов (в данной реализации 1), $r_t^{(i,e)}$ – вознаграждение, полученное агентом i в эпизоде e на шаге t, F_i – фитнес-оценка i-го агента.

После вычисления фитнеса производится отбор лучших представителей (в данной реализации – двух лучших, TOP_K = 2). Их веса сохраняются на диск в виде файлов agent0_best.pt и agent1_best.pt, а также используются как родительские особи для формирования следующей популяции. Новые агенты получают параметры по следующей схеме: лучшие (TOP_K) копируются напрямую, оставшиеся (K – TOP_K) формируются через кроссовер и мутацию.

Кроссовер осуществляется побитово (поэлементно) между двумя родителями $\theta^{(p1)}$ и $\theta^{(p2)}$:

$$\theta_j^{(child)} = \begin{cases} \theta_j^{(p1)}, \text{если } \xi_j < 0{,}5 \\ \theta_j^{(p2)}, \text{иначе} \end{cases}, \forall j,$$

где $\xi_j \sim \mathcal{U}(0,1)$ – случайная величина из равномерного распределения. Такая процедура обеспечивает разнообразие и комбинирование признаков обеих стратегий.

После скрещивания применяется мутация – добавление гауссовского шума к параметрам потомка:

$$\theta_j^{(child)} \leftarrow \theta_j^{(child)} + \varepsilon_j, j \sim \mathcal{N}(0, \sigma^2),$$

где нормальное распределение с математическим ожиданием 0 и дисперсией σ^2, σ – дисперсия мутации (в данной реализации равная 0,05).

В результате формируется новая популяция из K агентов, каждый из которых инициализируется независимо на основе сохраненных или сгенерированных весов. На следующем этапе обучения они участвуют в новых эпизодах взаимодействия, вновь обучаясь по градиенту, после чего эволюционный цикл повторяется. Таким образом, происходит непрерывный процесс отбора, обучения, скрещивания и обновления стратегий.

Подобная схема обладает рядом преимуществ:

– позволяет эффективно избегать переобучения за счет стохастической мутации;
– поддерживает постоянное разнообразие популяции, важное для исследования;
– обеспечивает устойчивость к шуму и нестабильности среды;
– способствует естественной адаптации стратегий за счет конкурентного отбора.

Таким образом, эволюционная составляющая вносит в процесс обучения дополнительный уровень метаобучения, способствующий устойчивой оптимизации при наличии стохастичности, частичного наблюдения и сложного поведения агентов в среде.

Результаты и обсуждение

Для оценки эффективности разработанного алгоритма обучения двух агентов в среде сбора ресурсов была проведена серия экспериментов, в которых сравнивались две стратегии оптимизации поведения: одна с применением периодической эволюционной селекции и мутаций (гибридная стратегия), другая – с применением только стохастического градиентного спуска по методу A2C (Advantage Actor-Critic) без механизма наследования и видоизменения генотипов (нейросетевых весов).

Обе стратегии были протестированы в одинаковой среде, где агенты должны были взаимодействовать и соревноваться за ресурсы, доставляя их к целевым координатам. Основным метриками качества служили средние значения наград по эпохам для каждого агента (Agent 0 и Agent 1), что отражает успешность выполнения задач и степень адаптации агентов к динамике среды и сопернику.

Обе стратегии обучались в течение 20000 эпох. Графики с метриками их обучения представлены на Рисунке 2 (A2C стратегия) и Рисунке 3 (гибридная стратегия).

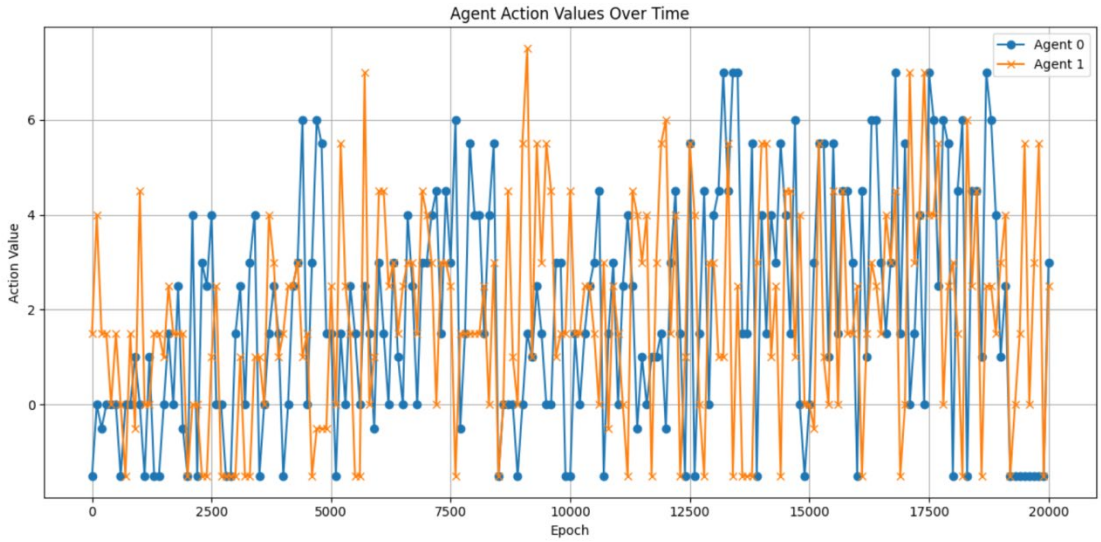

Рисунок 2 – График метрик обучения при стратегии A2C
Figure 2 – Training metrics chart under the A2C strategy

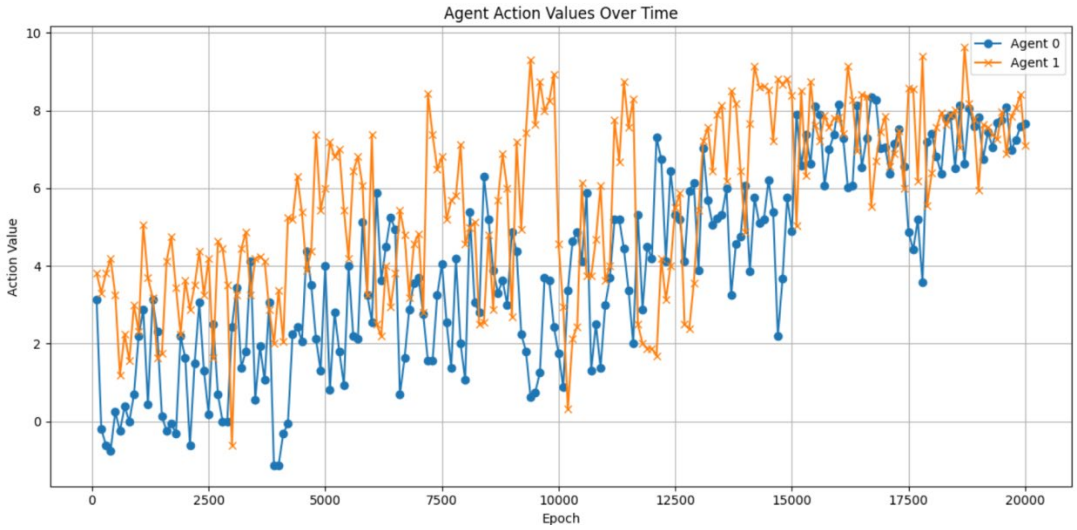

Рисунок 3 – График метрик обучения при гибридной стратегии
Figure 3 – Training metrics chart under the hybrid strategy

Как видно из данных графиков в эксперименте с обучением при чистой стратегии A2C вплоть до 10000-й эпохи агенты демонстрируют очень низкие награды, в основном колеблющиеся около −1,5 или 0. Лишь после 15000-й эпохи можно отметить небольшое улучшение: у Agent 1 награды начинают выходить на уровень 2,0–4,0, тогда как Agent 0 остается преимущественно в диапазоне от −1,5 до 1,5. Даже в поздних эпохах (20000–30000) обучение остается неустойчивым: у Agent 1 средние награды колеблются в диапазоне 1,5–5,5, но демонстрируют высокую дисперсию; Agent 0 поднимается лишь эпизодически до уровня 3,0–4,5 и не показывает устойчивого роста. Отсутствие механизмов селекции и мутаций приводит к стагнации в пространстве параметров.

В эксперименте с гибридной стратегией начальные эпохи показывают нестабильную динамику наград, особенно у Agent 0, что можно объяснить случайным начальным распределением весов и слабой приспособленностью к среде. Однако начиная с ~1000-й эпохи, наблюдается уверенное улучшение производительности обоих агентов. К 5000-й эпохе Agent 1 стабильно достигает средней награды в районе 6,0–7,5, а Agent 0 – в диапазоне 2,0–5,0. С 10000-й до 20000-й эпохи агенты демонстрируют высокую стабильность в наградах, часто превышающих значения 7,0, достигая пиковых значений в районе 8,5–9,3.

Такой рост можно напрямую связать с периодическим применением эволюционной селекции, при которой сохраняются только лучшие особи популяции и производится комбинирование их параметров с внесением случайных мутаций. Это позволяет избежать застоя в локальных минимумах и сохранить высокий уровень диверсификации решений в популяции агентов. В частности, именно благодаря мутациям, распределенным по нормальному закону, реализуется регулярный дрейф в пространстве параметров, стимулирующий исследование новых траекторий поведения.

Таким образом, включение эволюционного компонента в обучение двухагентных систем позволяет существенно ускорить сходимость и повысить итоговую адаптивность агентов к целевой задаче. Механизмы кроссовера и мутаций в нейросетевых параметрах обеспечивают дополнительное глобальное исследование пространства политик, что позволяет выходить за пределы локальных оптимумов, присущих градиентным методам оптимизации.

Заключение

В данной работе был реализован и исследован гибридный подход к обучению многоагентной системы в дискретной среде с ограничениями на локальное восприятие и конкуренцию за ресурсы. Комбинация градиентного метода Advantage Actor-Critic (A2C) с периодическим применением эволюционных механизмов отбора, кроссовера и мутации показала свою эффективность при обучении агентов с частично наблюдаемым состоянием среды.

Сравнительный анализ двух стратегий – обучения только по градиенту и обучения с добавлением эволюционного этапа – выявил критические различия в динамике сходимости и финальной результативности. Обучение с эволюцией продемонстрировало существенно более высокую устойчивость, скорость адаптации и итоговую производительность, особенно в условиях конкуренции между агентами. Отсутствие эволюционного давления в чистом A2C привело к стагнации и медленной сходимости, подтверждая необходимость диверсификации поведения в многозадачных и многопользовательских средах.

Проведенные эксперименты подтвердили, что включение стохастических мутаций и отбора особей на основе награды позволяет существенно повысить эффективность обучения, особенно в контексте совместной и конкурентной оптимизации. Данный результат может быть полезен при проектировании систем кооперативного или соревновательного взаимодействия в робототехнике, распределенном управлении и интеллектуальных многоагентных платформах.

В будущем планируется расширить метод за счет:

– использования более сложных сред с динамикой и препятствиями;

– добавления коммуникативных модулей;

– интеграции механизмов обучения с подкреплением с использованием памяти (replay buffer) и off-policy техник.

Предложенный подход может служить основой для разработки более сложных моделей адаптивного поведения в распределенных средах, включая как симуляцию агентных систем, так и прикладные задачи управления в реальном времени.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Yadav A., Kumar A., Choudhary Ch. Integrated Swarm Intelligence Framework for Dynamic Traffic Optimization in Delhi: A Three-Layer PSO-Fuzzy-MAS Approach. *International Scientific Journal of Engineering and Management.* 2025;04(05). https://doi.org/10.55041/ISJEM03921
2. Icarte-Ahumada G., He Zh., Godoy V., García F., Oyarzún M. A Multi-Agent System for Parking Allocation: An Approach to Allocate Parking Spaces. *Electronics*. 2025;14(5). https://doi.org/10.3390/electronics14050840
3. Dey S., Munsi A., Pradhan S., Aditya K. Bidirectional Wireless System for Drone to Drone Opportunity Charging in a Multi Agent System. In: *2023 International Conference on Control, Communication and Computing (ICCC), 19–21 May 2023, Thiruvananthapuram, India.* IEEE; 2023. P. 1–5. https://doi.org/10.1109/ICCC57789.2023.10164995
4. Souli N., Kolios P., Ellinas G. Multi-Agent System for Rogue Drone Interception. *IEEE Robotics and Automation Letters*. 2023;8(4):2221–2228. https://doi.org/10.1109/LRA.2023.3245412
5. Sanghi N. Deep Q-Learning (DQN). In: *Deep Reinforcement Learning with Python: RLHF for Chatbots and Large Language Models.* Berkeley: Apress; 2024. P. 225–271. https://doi.org/10.1007/979-8-8688-0273-7_6

6. Jeungthanasirigool W., Sirimaskasem Th., Boonraksa T., Boonraksa P. Comparison of PPO-DRL and A2C-DRL Algorithms for MPPT in Photovoltaic Systems via Buck-Boost Converter. *International Journal of Innovative Research and Scientific Studies*. 2025;8(3):2438–2453. https://doi.org/10.53894/ijirss.v8i3.7022
7. Bel Rio A., Jimenez D., Serrano J. Comparative Analysis of A3C and PPO Algorithms in Reinforcement Learning: A Survey on General Environments. *IEEE Access*. 2024;12:146795–146806. https://doi.org/10.1109/ACCESS.2024.3472473
8. Chen T.-Yo., Chen W.-N., Hao J.-K., Wang Ya., Zhang J. Multi-Agent Evolution Strategy with Cooperative and Cumulative Step Adaptation for Black-Box Distributed Optimization. *IEEE Transactions on Evolutionary Computation*. 2025. https://doi.org/10.1109/TEVC.2025.3525713
9. Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Computation*. 1997;9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
10. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 07–09 May 2015, San Diego, CA, USA.* 2015. URL: https://arxiv.org/abs/1412.6980

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Корчагин Алексей Павлович, аспирант, Воронежский государственный университет, Воронеж, Российская Федерация.
e-mail: aleksei.korchagin200@mail.ru

Aleksei P. Korchagin, Postgraduate, Voronezh State University, Voronezh, the Russian Federation.

Статья поступила в редакцию 15.06.2025; одобрена после рецензирования 18.07.2025; принята к публикации 30.07.2025.

The article was submitted 15.06.2025; approved after reviewing 18.07.2025; accepted for publication 30.07.2025.