АЛГОРИТМЫ ГЛУБОКОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В КООПЕРАТИВНО-СОРЕВНОВАТЕЛЬНОЙ СРЕДЕ

Гулис А.А., студент

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Калугина М.А. – канд. физ.-мат. наук, доцент

Аннотация. В докладе приведены результаты сравнительного анализа алгоритмов глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) применительно к задаче управления автономными агентами в кооперативно-соревновательной среде. Оценка эффективности алгоритмов выполнена на базе актуального клиента многопользовательской игры Tanks Blitz в формате боя «2 на 2» с учетом жестких ограничений пропускной способности среды.

Современные методы искусственного интеллекта демонстрируют выдающиеся результаты в задачах управления автономными системами [1], однако перенос теоретических моделей в реальные приложения сопряжен с рядом инженерных ограничений. Актуальность исследования обусловлена необходимостью создания интеллектуальных агентов – автономных систем принятия решений –, способных действовать в средах с частичной наблюдаемостью и разреженными наградами, где традиционные алгоритмы с жестко заданной логикой неэффективны. В связи с этим, задачей работы является определение наиболее эффективного подхода, обеспечивающего стабильную сходимость к оптимальной политике поведения агентов в кооперативно-соревновательной среде. Особую значимость представляет проблема совместного выбора алгоритма обучения (On-policy или Off-policy) и архитектуры нейронной сети, обеспечивающей баланс между сэмпл-эффективностью и стабильностью обучения при жестком лимите вычислительных ресурсов.

В качестве испытательного полигона использовался не упрощенный симулятор, а реальный игровой клиент, взаимодействие с которым реализовано через высоконагруженный WebSocket-интерфейс. Это приводит к критическим ограничениям:

1 Низкая пропускная способность. На вычислительном кластере (24 ядра CPU, GPU RTX 4080 Super) параллельно запускается около 30 экземпляров игры, что суммарно дает генерацию лишь около 300 шагов в секунду. Это на порядки меньше, чем в стандартных RL-бенчмарках (Atari, MuJoCo).

2 Разреженная терминальная награда. Победа достигается только при полном уничтожении команды противника. Вероятность случайного выполнения этой задачи на старте обучения стремится к нулю.

3 Частичная наблюдаемость (POMDP). Агенту доступна лишь локальная информация. Механики «тумана войны» и циклов перезарядки нарушают марковское свойство среды, требуя от нейросети наличия памяти.

Для решения поставленной задачи спроектирована распределенная асинхронная архитектура. Агенты обучаются в режиме разделения параметров [2], управляясь одной нейросетью. Сбор данных организован по принципу Actor-Learner [3], что позволяет отделить генерацию опыта на CPU от обучения на GPU.

Входные данные представляют собой вектор признаков, включающий не только относительную геометрию, но и данные массива виртуальных лучевых сенсоров. Агент испускает 8 лучей (вперед, назад, влево, вправо, по диагоналям) на дистанцию 50 метров, что позволяет определять препятствия и укрытия без использования ресурсоемких сверточных сетей (CNN).

Для стабилизации обучения в условиях разреженных наград применен метод формирования вознаграждения на основе потенциалов (Potential-Based Reward Shaping, PBRS) [4]. В ходе эксперимента выявлен эффект «взлома вознаграждения» (Reward Hacking) [5]: при поощрении только за урон агенты обучались стрелять в землю рядом, избегая сложного прицеливания. Проблема была решена введением коэффициента дистанции, с которым увеличивалась награда при попадании по противнику с увеличением дистанции, что перенастраивало агентов на тактику дальнего боя.

В ходе исследования были протестированы и модифицированы четыре алгоритмических подхода.

В качестве базового решения был выбран On-policy подход. Первоначальные эксперименты в статичных дуэлях «1 на 1» показали быструю сходимость алгоритма ввиду низкой сложности задачи. Однако масштабирование до динамичного формата «2 на 2» потребовало внедрения стратегии Curriculum Learning [6]. Суть метода заключалась в поэтапном увеличении дистанции появления команд: от ближнего боя, где контакт с противником гарантирован, до дальних дистанций, требующих сложного маневрирования и прицеливания. Был использован алгоритм PPO [7] с применением строгих практик реализации (ортогональная инициализация, нормализация входов) [8].

PPO продемонстрировал высочайшую стабильность и выработал тактику точного прицеливания. Однако из-за On-policy природы алгоритм требует сброса собранных данных после каждого обновления

весов. В условиях низкой пропускной способности это привело к неприемлемому времени обучения, сделав метод экономически нецелесообразным для дальнейшего масштабирования.

Для повышения сэмпл-эффективности был осуществлен переход к Off-policy методам. Первоначальные тесты выявили проблемы: попытка применить нормализацию наград PopArt [9] в условиях разреженных наград привела к деградации градиентов из-за околонулевой дисперсии. Базовая версия Discrete SAC [10] оказалась неустойчива: из-за обилия нулевых наград дисперсия нормализатора PopArt стремилась к нулю, а энтропийный коэффициент «взрывался», превращая политику в случайный шум. Для решения проблемы был внедрен алгоритм SDSAC [11], специально разработанный для стабилизации обучения в дискретных средах. Ключевой особенностью является использование Double-Average Q-learning (целевое значение вычисляется как среднее двух Q-сетей (критиков)), что устраняет проблему недооценки (underestimation bias), характерную для стандартного минимума. Дополнительно был внедрен горизонт планирования.

Отказ от пессимистичной оценки Q-функции и увеличение горизонта планирования обеспечили плавное снижение энтропии. SDSAC успешно сформировал тактику защитного позиционирования, существенно превзойдя PPO по скорости сходимости.

Для компенсации частичной наблюдаемости был внедрен алгоритм R2D2 (Recurrent Replay Distributed DQN) [12], использующий LSTM-память и обучение на последовательностях. Наличие памяти позволило агентам эффективно использовать укрытия и контролировать дистанцию во время перезарядки. Однако рекуррентная природа привела к появлению поведенческих артефактов – цикличные микро-движениях корпуса.

Наиболее перспективным направлением стала реализация гибридного агента [13], где управление шасси и контроль орудия остаются дискретным, а наведение башни – непрерывными. Это потребовало разработки функции потерь сети стратегии (актора) с двумя независимыми регуляторами энтропии. Архитектура нейросети была перепроектирована и включала два независимых выходных слоя: первый отвечал за выбор дискретных действий, а второй для непрерывных. Дополнительно было реализовано маскирование: при повреждении модуля поворота башни выходы нейросети обнуляются, чтобы не тратить значительную долю шагов на заведомо невозможные команды и исключить некорректные градиенты.

Гибридный подход обеспечил наилучшую точность стрельбы. Для обеспечения надежности обучения дискретной части агента в гибридную модель были успешно интегрированы механизмы из ранее протестированного алгоритма SDSAC.

Проведённое исследование показало, что математический аппарат алгоритма оказывает определяющее влияние на формируемую тактику. Для сред с низкой пропускной способностью оптимальным компромиссом являются Off-policy. Полученные результаты могут быть использованы при проектировании систем управления автономными роботами.

Сравнение подходов показало, что PPO лучше сохраняет стабильность и формирует аккуратное прицеливание, но в условиях ограниченной пропускной способности оказывается слишком дорогим по времени. Off-policy алгоритмы, напротив, быстрее накапливают полезный опыт и устойчивее масштабируются при низкой скорости генерации данных, однако требуют более тщательной настройки вознаграждения и стабилизационных приёмов.

Список использованных источников:

1. Sutton, R. S. Reinforcement Learning: An Introduction / R. S. Sutton, A. G. Barto. – 2nd ed. – Cambridge, MA : The MIT Press, 2018. – 552 p.

2. Revisiting Parameter Sharing in Multi-Agent Deep Reinforcement Learning [Электронный ресурс] / J. K. Terry [et al.] // arXiv, 2020. – arXiv:2005.13625. – DOI: 10.48550/arXiv.2005.13625.

3. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures / L. Espeholt [et al.] // Proceedings of the 35th International Conference on Machine Learning (PMLR), 2018. – Vol. 80. – P. 1407–1416.

4. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping / A. Y. Ng, D. Harada, S. Russell // Proceedings of the Sixteenth International Conference on Machine Learning (ICML), 1999. – P. 278–287.

5. Concrete Problems in AI Safety [Электронный ресурс] / D. Amodei // arXiv, 2016. – DOI: 10.48550/arXiv.1606.06565.

6. Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey / S. Narvekar [et al.] // Journal of Machine Learning Research, 2020. – Vol. 21(181). – P. 1–50.

7. Proximal Policy Optimization Algorithms [Электронный ресурс] / J. Schulman [et al.] // arXiv, 2017. – arXiv:1707.06347. – DOI: 10.48550/arXiv.1707.06347.

8. What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study [Электронный ресурс] / M. Andrychowicz [et al.] // International Conference on Learning Representations (ICLR), 2021. – arXiv:2006.05990. – DOI: 10.48550/arXiv.2006.05990.

9. Multi-Task Deep Reinforcement Learning with PopArt / M. Hessel [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence, 2019. – P. 3796–3803. – DOI: 10.1609/AAAI.V33I01.33013796.

10. Soft Actor-Critic for Discrete Action Settings [Электронный ресурс] / P. Christodoulou // arXiv, 2019. – arXiv:1910.07207. – DOI: 10.48550/arXiv.1910.07207.

11. Revisiting Discrete Soft Actor-Critic [Электронный ресурс] / H. Zhou // arXiv, 2022. – DOI: 10.48550/arXiv.2209.10081.

12. Recurrent Experience Replay in Distributed Reinforcement Learning / S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, W. Dabney // International Conference on Learning Representations (ICLR) (Poster), 2019.

13. A Hybrid SAC Algorithm: Advantage-Guided Soft Actor-Critic / C. Chen, X. Deng // MSCE '25: Proceedings of the 2025 International Conference on Management Science and Computer Engineering, 2025. – P. 459–462. – DOI: 10.1145/3760023.3760100.