

ПРИМЕНЕНИЕ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В МНОГОПОЛЬЗОВАТЕЛЬСКИХ СТРАТЕГИЧЕСКИХ ИГРАХ

А.С. Мисник (misnik.as@phystech.edu)

Московский физико-технический институт, Москва

В работе рассматриваются современные методы, используемые для многоагентного обучения с подкреплением на примере многопользовательских стратегических игр. Описывается модель игры, раскрывающей тонкости кооперативного многоагентного взаимодействия. Приведены результаты экспериментов в модели описанной игры, особенности каждого из рассмотренных подходов, а также предлагаемые оптимизации, способствующие ускорению и повышению качества обучения.

Ключевые слова: обучение с подкреплением, многоагентное обучение, нейросетевые методы, графовые нейронные сети, декомпозиция вознаграждений.

Введение

Обучение с подкреплением представляет собой методику в машинном обучении, в которой агент обучается взаимодействовать с окружающей средой, принимая решения для достижения определенных целей. Агенты получают сигналы вознаграждения или наказания в зависимости от результатов своих действий, и их цель заключается в максимизации итогового вознаграждения.

Многоагентное обучение с подкреплением – это раздел обучения с подкреплением, в которой несколько агентов взаимодействуют в одной среде, обучаясь принимать решения для достижения индивидуальных или коллективных целей [Stefano, 2024]. Этот раздел представляет особый интерес в практических приложениях, поскольку многие реальные задачи сводятся к совместной работе нескольких агентов в конкурентных средах.

Однако многоагентное обучение с подкреплением также связано с рядом технических сложностей – возрастает вычислительная сложность из-за экспоненциального роста пространства состояний и действий с увели-

чением числа агентов. Важными вызовами также являются координация агентов, распределение вознаграждения в командных задачах и сбалансированное сотрудничество [Foerster, 2016]. В работе рассмотрены методы и их оптимизации, которые позволяют эффективно решать проблемы координации, масштабируемости и распределения вознаграждений в много-агентных системах, а также их применение в реальных сценариях.

1. Постановка задачи

Задачу обучения с подкреплением традиционно формализуют как задачу Марковского процесса принятия решений, задаваемого кортежем $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, где:

- множество всех возможных состояний среды,
- множество допустимых действий агента,
- модель переходов, определяющая вероятность перехода в состояние s' из состояния s при выборе действия a ;
- функция вознаграждения, определяющая ожидаемое мгновенное вознаграждение, получаемое при выборе действия a в состоянии s ,
- коэффициент дисконтирования, отражающий относительную значимость будущих вознаграждений.

Цель агента: Найти оптимальную стратегию π^* , которая максимизирует математическое ожидание суммы дисконтированных вознаграждений [Саттон, 2014]:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \right]$$

В многоагентном обучении с подкреплением постановка задачи усложняется за счёт взаимодействия нескольких агентов, каждый из которых стремится максимизировать свою собственную или общую награду. Многоагентная система может быть описана кортежем $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma, n \rangle$:

- количество агентов в системе,
- множество допустимых действий i -го агента,
- функция переходов, определяющая вероятность перехода в состояние s' при совместных действиях агентов,

– функция вознаграждения i -го агента, зависящая от состояний и действий всех агентов.

Цель каждого агента заключается в поиске стратегии, максимизирующей его ожидаемую дисконтированную награду:

$$\pi_i^* = \arg \max_{\pi_i} \mathbb{E}_{\pi_1, \dots, \pi_N} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, a_{1,t}, \dots, a_{N,t}) \right],$$

где

– действие агента в момент времени.

2. Модель рассматриваемой игры

В работе для проведения экспериментов и сравнения результатов используется модель игры из публичного соревнования, приуроченного к конференции AI Journey 2024 по теме Multiagent AI.

Среда представляет из себя клеточное поле, разделенное на 9 секторов в виде квадрата 3×3 с возможными случайными смещениями. Центральный сектор недоступен для агентов, а на остальных восьми расположен агент, его пункт переработки отходов и его пункт выдачи награды. Игра происходит по шагам, на каждом из которых в среде с некоторой вероятностью появляется ресурс. Для получения награды агенту нужно дойти до ресурса, взять его и перенести в пункт выдачи награды. После этого на секторе агента формируется отход, для устранения которого агенту нужно взять его и донести до пункта переработки.

Вероятность выпадения новых ресурсов на секторе агента зависит от “экологического показателя” – количества отходов на секторе агента и на соседних по стороне секторах. Эта особенность игровой модели лежит в основе кооперативной концепции – действия каждого агента влияют на общий успех.

Метрика, которую нужно максимизировать – средняя награда среди всех пользовательских агентов за заранее известное количество шагов игры.

Важным замечанием здесь является то, что перед началом игры случайным образом определяется, какие агенты будут управляться пользовательскими стратегиями, а какие – недоступными для пользователя стратегиями (например, системой). Это является серьезной проблемой для стратегий, которые полагаются на кооперативные взаимодействия.

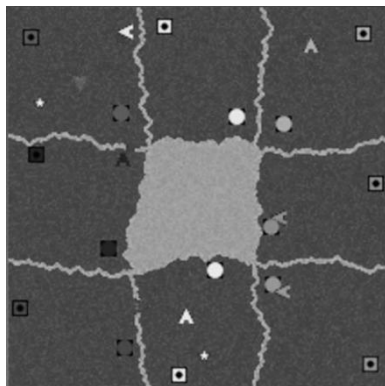


Рис. 1. Визуализация модели игры. Размер карты 210 * 210.

Стрелки – агенты, которые собирают ресурсы (звезды).

Квадраты – пункты переработки и выдачи награды

3. Нейросетевые подходы и эксперименты

В рамках исследования реализовано несколько подходов и оптимизаций к ним. При зафиксированной конфигурации среды (количество шагов, вероятность выпадений ресурсов на шагах, размер карты и прочее) предложенные методы сравнивались по следующим параметрам:

1. Награда, полученная при управлении всеми 8 агентами обученной для каждого из них стратегией.
2. Награда, полученная при управлении случайно выбранными 4 агентами, взаимодействующими с агентами, управляемыми какой-то из других стратегий (для проверки того, насколько обученные агенты зависят от кооперации).
3. Время обучения до получения стабильного уровня награды.
4. Эффективность относительно взаимодействия со средой (sample efficiency) – этот параметр не всегда коррелирует со временем обучения, поскольку эффективность использования данных может отличаться.

3.1. Метод Value Decomposition Network (VDN)

VDN [Sunehag, 2017] – алгоритм кооперативного обучения с подкреплением, основанный на разложении общей награды в сумму индивидуальных наград. Он отличается простотой реализации и может служить хорошим базовым уровнем в многих задачах многоагентного обучения. Его преимуществом относительно других многоагентных алгоритмов служит то, что стратегии агентов можно обучать многопоточно, позволяя ускорить процесс обучения на многоядерных архитектурах. Но тем не

менее, классическая реализация VDN без дополнительных оптимизаций не является наилучшим решением в поставленной задаче по нескольким причинам:

1. Поскольку агенты обучаются в предположении, что общая награда разбивается в композицию наград отдельных агентов, обученная стратегия сугубо кооперативна и, как следствие, значительно теряет качество при взаимодействии с агентами, обученными по другим стратегиям.
2. Эффективность взаимодействия со средой не оптимальна из-за параллельного обучения.

3.2. Hindsight Experience Replay (HER) при обучении VDN

Hindsight Experience Replay [Andrychowicz, 2017] – способ целеполагания в обучении с подкреплением, способствующих обучению в средах с разреженными наградами. Поскольку награда в модели рассматриваемой игры достаточно разрежена, интеграция этого метода в процесс обучения VDN способствует ускорению обучения – требуется меньше шагов симуляции, чтобы агенты обучились базовым паттернам поведения, которые смогут развивать в дальнейшем. Подобная идея была выдвинута при обучении DDPG [Zhou et al., 2023], а в данной работе она развита и адаптирована для параллельного обучения VDN.

3.3. Метод Multi-Agent Proximal Policy Optimization (MAPPO)

MAPPO – адаптация алгоритма Proximal Policy Optimization для многоагентного обучения. В сравнении с VDN, MAPPO требует больше вычислительных ресурсов для обучения - из-за отсутствия возможности параллельного обучения агентов, предоставляемой в подходе VDN. Еще одной особенностью этого подхода является высокая зависимость качества обучения от правильно подобранных гиперпараметров – например, размера примеров в батче. Но несмотря на описанные сложности, этот подход добивается наилучших результатов в сравнении с описанными ранее методами.

3.4. Оптимизация MAPPO с использованием GNN

Внедрение графовых нейронных сетей (GNN) в алгоритм MAPPO позволяет значительно улучшить координацию между агентами за счет явного моделирования их взаимодействий. В отличие от классической реализации MAPPO, где агенты обмениваются информацией через общие скрытые слои или глобальное состояние среды, GNN явно кодирует структуру взаимодействий в виде графа, где узлы соответствуют агентам, а ребра – их связям.

Кроме этого, представление взаимодействий между агентами в виде графа позволяет оценивать, насколько действия соседних агентов влияют друг на друга и насколько их наблюдения пересекаются. Эти оценки используются в процессе обучения для того, чтобы дополнительно штрафовать стратегии за чрезмерную кооперативность.

4. Сравнение подходов

Для экспериментов были зафиксированы параметры среды:

1. Поле размера $210 * 210$ разделено на сектора $70 * 70$.
2. Игровой эпизод состоит из 1000 тактов, в каждый из которых агенты совершают движение размером 7 клеток.
3. Базовая вероятность (при отсутствии мусора) появления ресурса на каждом из тактов = 0.1.

В табл. 1 приведены результаты сравнительного анализа рассмотренных методов. Время обучения приведено в условных единицах, поскольку зависит от ресурсов, использованных для обучения. Эффективность взаимодействия со средой тоже не представляется возможным оценить в абсолютных величинах, поэтому приведены относительные оценки, основанные на собранных метриках (4– наиболее эффективно, 1– наименее эффективно):

Таблица 1

Метод	Награда ¹ (модель управляет 8 агентами)	Награда ¹ (модель управляет 4 случайными агентами)	Ресурсоемкость обучения (в условных единицах)	Эффектив- ность взаи- модействия со средой
VDN	46.58	41.72	2.3 Т	1
VDN + HER	47.30	42.29	1 Т	3
MAPPO	52.43	50.89	7.8 Т	2
MAPPO + GNN	56.14	54.11	8.3 Т	4

Выводы, которые можно сделать из полученных результатов:

1. Методы с использованием MAPPO добиваются наилучших сравнительных результатов, обладают большей эффективностью взаимодействия со средой, но требуют больше ресурсов для обучения.
2. Приведенные оптимизации положительно влияют на процесс обучения – например, оптимизация VDN с помощью переопределения целей сильно улучшает ресурсоемкость.

Заключение

В данной работе рассмотрены современные подходы к многоагентному обучению с подкреплением и их применение в кооперативной среде, где успех агентов зависит как от их индивидуальных действий, так и от взаимодействия с другими участниками. Проведено сравнение нескольких методов: Value Decomposition Network (VDN), его оптимизированную

¹ Среднее по агентам количество ресурсов, утилизированных за эпизод.

версию с Hindsight Experience Replay (HER), а также Multi-Agent Proximal Policy Optimization (MAPPO) и его улучшенный вариант с использованием графовых нейронных сетей (GNN).

Реализованные оптимизации ранее не освещались в научных исследованиях и представляют альтернативный взгляд на известные методы многоагентного обучения.

В общем случае, выбор оптимального метода зависит от конкретной задачи, ограничений на вычислительные ресурсы и сложности взаимодействия со средой. Например, метод HER, приведенный в работе, подходит для задач с разреженной наградой. Метод MAPPO с приведенными оптимизациями способен достичь приемлемых результатов в задачах, где взаимодействие со средой не тратит много вычислительных ресурсов.

Применяющиеся на практике подходы не ограничиваются приведенными: зачастую наилучший результат показывают ансамбли моделей, сочетающие преимущества разных подходов.

Одной важной концептуальной проблемой предложенных методов является необъяснимость приведенных стратегий – поскольку нейронные сети хранят только переработанную информацию о состояниях среды и могут обмениваться данными только с помощью дистилляции, не способной передать структурную информацию о среде. Это отдельное направление для дальнейших исследований – применимость структурного обучения в многоагентных средах.

Список литературы

- [Саттон и др., 2020] Ричард С. Саттон, Эндрю Дж. Барто // Обучение с подкреплением (второе издание). – MIT Press, 2020. – 553 с.
- [Andrychowicz et al., 2017] Andrychowicz M. et al. Hindsight experience replay // Advances in NeurIPS. – 2017. – doi: 10.48550/arXiv.1707.01495.
- [Foerster et al., 2016] Foerster J. et al. // Learning to communicate with deep multi-agent reinforcement learning // Advances in NeurIPS. – 2016. – doi: 10.48550/arXiv.1605.06676.
- [Jiang et al., 2018] Jiang J. et al. // Graph convolutional reinforcement learning // arXiv preprint. – 2018. – doi: 10.48550/arXiv.1810.09202.
- [Munikoti et al., 2023] Munikoti S. et al. // Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications // IEEE Transactions on Neural Networks and Learning Systems. – 2023. – doi: 10.1109/TNNLS.2023.3275593.
- [Stefano et al., 2024] Stefano V. Albrecht, Filippos Christianos, Lukas Schäfer // Multi-Agent Reinforcement Learning: Foundations and Modern Approaches – MIT Press, 2024. – 395 с.
- [Sunehag et al., 2017] Sunehag P. et al. Value-decomposition networks for cooperative multi-agent learning. – 2017. – doi: 10.48550/arXiv.1706.05296.
- [Yu et al., 2022] Yu C. et al. // The surprising effectiveness of PPO in cooperative multi-agent games // Advances in NeurIPS. – 2022. – doi: 10.48550/arXiv.2203.02155
- [Zhou et al., 2023] Zhou Y. et al. // Cooperative multi-agent target searching: a deep reinforcement learning approach based on parallel hindsight experience replay // Complex Intell. Syst. – 2023.