ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

Мисник Андрей Сергеевич

ПРИМЕНЕНИЕ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В МНОГОПОЛЬЗОВАТЕЛЬСКИХ СТРАТЕГИЧЕСКИХ ИГРАХ

Специальность 09.04.01 —
«Информатика и вычислительная техника»

Черновик статьи на КИИ

Научный руководитель:
д.ф.-м.н. Виноградов Дмитрий Вячеславович

Москва 2025

Замечания

1. Для докладов на КИИ предполагается оформление в word-формате, но по договоренности с научным руководителем первоначально собираем наработки в tex-формате. Кроме того, возможно будем подаваться на более релевантную конференцию.
2. Планы на ближайшее время: нужно довести эксперименты с ВКФ в среде игры, и дооформить главу с экспериментами с помощью полученных результатов.
3. Как обсуждалось ранее, среди нейросетевых подходов имеет смысл попробовать более новые подходы - планируются эксперименты с GNN [1].
4. Стоит добавить литературы - например, идея формирования знаний из LLM, и подумать про то, как соотнести эту идею с Behaviour Cloning.

Аннотация

В работе рассматриваются современные методы, используемые для обучения с подкреплением на примере многопользовательских стратегических игр. Помимо фундаментальных нейросетевых методов, требующих тяжелых вычислительных ресурсов, в работе предлагается развитие алгоритма ВКФ для кооперативного взаимодействия агентов, основанное на получении информации с использованием MCTS и возможности передачи базовых знаний (подсказок) из LLM. Основные преимущества предлагаемого подхода заключаются в эффективности обучения и интерпретируемости полученных результатов, что может быть полезно в некоторых сферах решаемых задач.

Содержание

Стр.

Введение 5

Глава 1. Постановка задачи 6

Глава 2. Модель рассматриваемой игры 7

Глава 3. Нейросетевые подходы 8

3.1 VDN 8

3.2 MAPPO 8

3.3 Сравнение нейросетевых подходов 9

Глава 4. Структурное обучение и ВКФ 10

Глава 5. Преимущества и тенденции развития ВКФ 12

5.1 Передача знаний с использованием LLM 12

5.2 Мышление модели в терминах подзадач 12

Заключение 14

Список литературы 15

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.

К сожалению, большинство современных моделей машинного обучения хранит переработанную информацию об объектах и результаты обучения отдельно от исходных признаков, теряя структурную взаимосвязь между ними. Кроме того, дистилляция, как единственная общепринятая методика передачи знаний между моделями, не способна передать структурные особенности рассматриваемой среды.

Решением этих проблем занимается структурное обучение с подкреплением. Оно распространяет концепции обучения с подкреплением на сложные задачи, требующие работы с богатыми структурами данных. Использование структурного обучения способно подтолкнуть агента учитывать контекст и скрытые зависимости, а также сформировать человекочитаемый результат обучения - некоторый набор правил, который определяет оптимальное действие в зависимости от некоторых условий. Это может быть полезно, например, для ручного анализа экспертами в предметной области исходной задачи.

1. Постановка задачи

Задачу обучения с подкреплением традиционно формализуют как задачу Марковского процесса принятия решений, задаваемого кортежем (S, A, T, R, γ), где:
S — множество всех возможных состояний среды
A — множество допустимых действий агента
$T : S \times A \times S \to [0, 1]$ — модель переходов, определяющая вероятность перехода в состояние $s' \in S$ из состояния $s \in S$ при выборе действия $a \in A$:

$$T(s, a, s') = \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$$

- $R : S \times A \to \mathbb{R}$ — функция вознаграждения, определяющая ожидаемое мгновенное вознаграждение, получаемое при выборе действия a в состоянии s: R(s, a) = E[$r_{t+1} \mid s_t = s, a_t = a$]
$-\gamma \in [0, 1]$ — коэффициент дисконтирования, отражающий относительную значимость будущих вознаграждений.

Цель агента: Найти оптимальную стратегию $\pi^* : S \to A$, которая максимизирует математическое ожидание суммы дисконтированных вознаграждений:

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

В структурном обучении с подкреплением вводится набор бинарных признаков состояний и фиксируется отображение F : S $\to \mathcal{P}(\{0, 1\})$ такое, что $\forall s \in S$, $F(s) = \{f_1, f_2, \ldots, f_n\}$, где $f_i \in \{0, 1\}$.

Целью агента в структурном обучении является порождение множества гипотез H $(\forall h_i \in H : h_i = \{F_i, A_i\}, F_i \subset \{0, 1\}^n)$, где каждая гипотеза имеет свойство $\forall s_t \in S : F(s_t) \in F_h \Rightarrow A_h \sim \pi(\cdot|s_t)$. В конечном итоге каждая из полученных гипотез описывает "правило", объединяющее некоторые классы состояний.

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.

Количество отходов на всем поле влияет на вероятность выпадения новых наград для всех агентов. Этот факт зарождает в модели игры "кооперативную" концепцию - действия каждого отдельного агента влияют на общий успех.

Метрика, которую в рамках модели нужно максимизировать - средняя награда среди всех агентов, управляемых пользователем за некоторое конечное количество шагов игры.

Важным замечанием здесь является то, что перед началом игры система определяет, какие агенты будут управляться пользовательскими стратегиями, а какие - недоступными для пользователя стратегиями. Это является серьезной проблемой для стратегий, которые полагаются на кооперативные взаимодействия.



Рис. 2.1 — Схематичное представление модели игры

3. Нейросетевые подходы

В качестве базовых подходов для сравнения на моделях многопользовательских игр были выбраны VDN [2] и MAPPO [3].

3.1 VDN

VDN - алгоритм кооперативного обучения с подкреплением, основанный на разложении общей награды в сумму индивидуальных наград.
Основной его проблемой в модели рассматриваемой игры является то, что сформированная им стратегия сугубо кооперативна и её можно обучить только в предположении, что все агенты управляются пользователем. Если итоговые стратегии случайным образом смешиваются с другими стратегиями, итоговое качество значительно падает.

3.2 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. Тем не менее, результаты, полученные этим подходом, меньше зависят от кооперативности и по метрикам обгоняют подход VDN.

3.3 Сравнение нейросетевых подходов

С зафиксированными в публичной конфигурации среды параметрами получены следующие результаты:

	Score (8 agents)	**Score (4 agents vs 4 foreign agents)**
VDN	42.53	37.11
MAPPO	46.94	42.03

Из этой таблицы видно, что MAPPO добивается наилучших результатов среди рассмотренных подходов.

4. Структурное обучение и ВКФ

Приведенные нейросетевые подходы объединяет одна концептуальная проблема, упомянутая во введении - в результате обучения теряется структурная взаимосвязь между состояниями игры, что не позволяет интерпретировать сгенерированные стратегии.

В статье [4] предлагается инновационный подход к структурному обучению с подкреплением, результатом обучения которого являются правила, которые при наличии у состояния определенного набора свойств утверждают наилучшее действие для этого состояния.

Между состояниями вводится функция сходства, основанная на их признаковом представлении: $sim(S_1, S_2) = F(S_1)\&F(S_2)$. Гипотезы формируются для наборов сходных параметров между похожими состояниями, в которых предпринимается одно и то же действие. Кроме того, гипотезы должны быть сформированы так, чтобы у них не было "отрицательных примеров" - состояний, обладающих этими же свойствами, но имеющих другое оптимальное действие.

Алгоритм порождения гипотез устроен следующим образом:

1. Каждая гипотеза-кандитат порождается независимо с использованием цепи Маркова
2. Для каждого из затрагиваемых в гипотезе состояний проверяется оптимальность предпринятого действия (используя результат симуляции MCTS для этого состояния)
3. Гипотеза подтверждается, если действие оптимально для всех затрагиваемых состояний

Data: множество обучающих (+)- и (-)-примеров; число N порождаемых ВКФ-гипотез
Result: случайная выборка S ВКФ-гипотез
$O :=$ (+)-примеры, $F :=$ признаки; $I \subseteq O \times F$ формальный контекст для (+)-примеров;
$C :=$ (-)-примеры; $S := \emptyset$; $i := 0$;
while $(i < N)$ **do**
  породить кандидата $\langle A, B \rangle$ с помощью цепи Маркова;
  $hasObstacle :=$ **false**;
  for $(c \in C)$ **do**
    if $(B \subseteq c')$ **then**
      $hasObstacle :=$ **true**;
    end
  end
  if $(hasObstacle =$ ***false***$)$ **then**
    $S := S \cup \{\langle A, B \rangle\}$;
    $i := i + 1$;
  end
end

Рис. 4.1 — Процедура индуктивного обобщения ВКФ для порождения гипотез

5. Преимущества и тенденции развития ВКФ

5.1 Передача знаний с использованием LLM

Для многих моделей игр, в том числе многопользовательских, существует большое количество доступных ресурсов - советы, инструкции, черты оптимальных стратегий, и прочее. Подобные данные могут быть агрегированы и обработаны с помощью различных LLM - в качестве результата обработки можно получить набор некоторых правил, описывающих семантику действий в определенных ситуациях. Эти правила имеют такой же вид, что и гипотезы, порождаемые ВКФ. В таком случае, будучи приведенными в определенный формат, они могут быть переданы алгоритму ВКФ в качестве "подсказок".

В этой идее важно осознавать, что сформированные правила подвержены галлюцинациям - проблему следования таким правилам решает механизм "отвержения", концептуально похожий на "забывание" (forget gate) в архитектуре LSTM.

5.2 Мышление модели в терминах подзадач

На небольших ограничениях исходной модели игры (среда меньшего размера, сильно огранниченное количество эпизодов) можно алгоритмически (с использованием обхода графов игры) получить гарантированно оптимальные стратегии.

Анализируя эти стратегии, можно прийти к выводу, что оптимальная стратегия для агента разбивается в последовательность подзадач, которые он друг за другом выполняет.

К примеру, такой подзадачей может быть "дойти кратчайшим путем от текущей клетки до клетки с ресурсом и забрать ресурс". Эта

подзадача имеет очевидное решение, которое может быть легко получено без обучения:

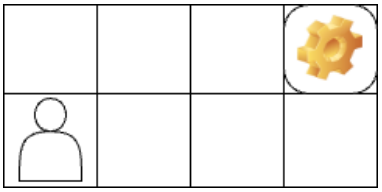

Рис. 5.1 —
Пример подзадачи, которая разбивается в последовательность (UP R R R PICK)

В таком случае, можно поменять подход к обучению - свести его к нахождению последовательности "подзадач", каждую из которых решать независимо и оптимально. Оптимальная стратегия будет выглядеть следующим образом:

- Если активных подзадач нет, возьми наиболее приоритетную, выбранную с помощью ВКФ, и выполни первое её действие.
- Если есть активная подзадача, сделай действие, оптимально приближающее к завершению подзадачи.

Заключение

В данной работе представлены различные методы для решения задач в рамках кооперативного взаимодействия. Актуальным в научной сфере нейросетевым методам противопоставляется подход ВКФ, способный масштабироваться на среды со сложной архитектурой. Предлагаемая идея с обучением "подзадач" в отличие от переходов позволяет перейти на более высокий уровень абстракции, недоступный при обучении нейросетей. Кроме того, рассмотренный подход с формированием базовых инструкций на основе результата работы LLM предоставляет альтернативу существующему методу дистилляции моделей.

Помимо прочего, вычисления, требуемые при обучении предлагаемого подхода, не требуют масштабных вычислительных мощностей и отличаются своей легковесностью.

Список литературы

1. Graph convolutional reinforcement learning / Jiechuan Jiang, Chen Dun, Tiejun Huang [и др.] // arXiv preprint arXiv:1810.09202. 2018.

2. Value-decomposition networks for cooperative multi-agent learning / Peter Sunehag, Guy Lever, Audrunas Gruslys [и др.] // arXiv preprint arXiv:1706.05296. 2017.

3. The surprising effectiveness of ppo in cooperative multi-agent games / Chao Yu, Akash Velu, Eugene Vinitsky [и др.] // Advances in neural information processing systems. 2022. Т. 35. С. 24611–24624.

4. Виноградов Дмитрий. Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток // ФИЦ ИУ РАН. 2018.