УДК 004.852

Дообучение агентов для решения задач обучения с подкреплением при помощи адаптеров-трансформеров

Козин Р. А.[1]
Научный руководитель – канд. техн. наук, Щербаков О. В.
[1]Университет ИТМО
roman.a.kozin@yandex.ru

Введение

Обучение с подкреплением является одним из способов обучения моделей. При использовании такой методики обучаемая система (чаще называемая агентом) должна выучить стратегию принятия решений, опираясь на информацию, получаемую из окружающей среды. При этом правило выбора действий должно максимизировать получаемую награду. На данный момент подходы такой методики обучения можно встретить преимущественно в робототехнике и при дообучении больших языковых моделей [1]. Расширение сферы применимости обучения с подкреплением связано с развитием глубоких нейронных сетей, в частности с возникновением больших языковых и визуально-языковых моделей. Агенты, основанные на таких моделях, имеют лучшую обобщаю способность, они способны извлекать семантическую информацию из окружения [2]. На данный момент лучшие открытые агенты разработаны на основе Qwen [3], но они имеют несколько миллиардов параметров. Такой размер усложняет обучение и внедрение агентов для работы в реальном времени, требуются значительные вычислительные ресурсы. Ранее была предложена концепция обучения высокоэффективных агентов на базе маловесных визуально-языковых моделей (несколько сотен миллионов параметров) и адаптеров [4]. Такая конструкция зарекомендовала себя в бенчмарке libero, предназначенном для робототехники. Однако поведение подобных архитектур не исследовано в более динамичных средах как видеоигры, например: Minecraft. Целью данного исследования является проверка применимости адаптера-трансформера для дообучения агентов в видеоигре Minecraft. На основе полученных результатов можно выявить ключевые достоинства и ограничения по использованию адаптеров на основе трансформеров, также сформируется основа для будущих исследований и разработки более сложных архитектур.

Основная часть

В рамках данного исследования было рассмотрено несколько моделей со следующей структурой: имеется предобученная опорная сеть, которая используется для извлечения признаков из визуальных наблюдений (цветные изображения), параметры этой модели зафиксированы во время обучения. Для агрегирования информации из последовательности визуальных признаков и последовательности предыдущих действий используется адаптер. На основе собранной информации адаптер предсказывает следующее действие, которое должен совершить агент. Параметры адаптера обучаемы. Ключевая идея заключается в разработке относительно небольшой модели. В данном случае модель (опорная сеть и адаптер) ограничена размером в 500 млн параметров. В связи с этим для сравнения были выбраны две опорных сети. Первая модель это VPT, которая уже является агентом, способным играть в Minecraft [5]. Однако VPT опирается лишь на историю наблюдений. Только по наблюдениям сложно задать конкретную задачу для выполнения, внедрение адаптера позволяет добавить обуславливание в виде истории действий. Введение дополнительной информации другой модальности должно

снизить неопределённость при выборе следующего действия. Второй моделью был выбран визуальный энкодер SigLIP. Эта модель обучена на разнообразном множестве изображений и эффективно извлекает признаки из изображений [6]. В качестве адаптера было рассмотрено два варианта. Основным вариантом является адаптер на основе трансформера. Ключевая особенность трансформера заключается в обработке последовательностей [7]. Также при помощи механизма перекрёстного внимания облегчается механизм слияния модальностей [8]. В данном случае имеется две модальности: визуальные признаки и совершённые действия. Чтобы проверить работоспособность концепции на основе трансформера, для сравнения также был выбран адаптер на основе FiLM. Принцип работы такого адаптера заключается в поэлементной модуляции признаков одной модальности обуславливанием другой модальностью [9]. В рамках исследования все модели обучались предсказывать последнее действие эпизода на основе предыдущих наблюдений и действий. Датасет взят из открытого источника [5] и представляет из себя игровой процесс, записанный игроками. В датасете наблюдался дисбаланс по целевому действию. Для решения этой проблемы использовалось сэмплирование. Все модели обучались на фиксированном наборе данных на протяжении четырёх эпох. При обучении модели стремились минимизировать функцию потерь перекрёстной энтропии. Для оценки качества обучения использовались перекрёстная энтропия (действия как целочисленные классы) и средняя абсолютная ошибка (действия как двоичные векторы) на валидационной выборке данных.

Выводы

По результатам данного исследования адаптер в виде FiLM был склонен к переобучению, а адаптер на основе трансформера обучался на протяжении всех четырёх эпох. Модель с опорной сетью SigLIP2 оказалась на 10% точнее по функции потерь и на 21% точнее по средней абсолютной ошибке, чем модель с опорной сетью VPT.

Литература

1. Naeem M., Rizvi S. T. H., Coronato A. A gentle introduction to reinforcement learning and its application in different fields //IEEE access. – 2020. – Т. 8. – С. 209320-209344.
2. Zitkovich B. et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control //Conference on Robot Learning. – PMLR, 2023. – С. 2165-2183.
3. Li M. et al. Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse //Findings of the Association for Computational Linguistics: ACL 2025. – 2025. – С. 17878-17899.
4. Wang Y. et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model //arXiv preprint arXiv:2509.09372. – 2025.
5. Baker B. et al. Video pretraining (vpt): Learning to act by watching unlabeled online videos //Advances in Neural Information Processing Systems. – 2022. – Т. 35. – С. 24639-24654.
6. Tschannen M. et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features //arXiv preprint arXiv:2502.14786. – 2025.
7. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.
8. Zhao F., Zhang C., Geng B. Deep multimodal data fusion //ACM computing surveys. – 2024. – Т. 56. – №. 9. – С. 1-36.
9. Perez E. et al. Film: Visual reasoning with a general conditioning layer //Proceedings of the AAAI conference on artificial intelligence. – 2018. – Т. 32. – №. 1.