УДК 004.852

Сравнительный анализ алгоритмов обучения с подкреплением

Козин Р. А. (ИТМО)
Научный руководитель — кандидат технических наук, Щербаков О. В.

Введение. Обучение с подкреплением является одной из методик машинного обучения. При использовании такого подхода обучаемая система (агент) должна научиться выполнять такие действия в заданном окружении, чтобы достигать наилучшего результата — максимизировать получаемую награду. Благодаря развитию глубоких нейросетевых моделей сферы применимости глубокого обучения расширилось. Сейчас алгоритмы обучения с подкреплением на основе глубокого обучения можно встретить в больших языковых моделях, робототехнике и финансах [1,2]. Целью данного исследования является сравнительный анализ распространённых алгоритмов глубокого обучения с подкреплением. Результаты этого исследования позволят не только выявить ключевые принципы и ограничения, но и сформировать базу для дальнейшего исследования и разработки более сложных алгоритмов.

Основная часть. В рамках данного исследования были рассмотрены алгоритмы глубокого обучения с подкреплением, основанные на обучении функции ценности [3] и градиента стратегии, в частности REINFORCE с базой [4] и оптимизация проксимальной стратегии [5]. Выбор данных алгоритмов был обусловлен их широким распространением и различиями в подходе к обучению агента. Основным критерием для анализа выбранных алгоритмов являлась скорость обучения. В рамках исследования оценивалось количество эпизодов, требуемое для того, чтобы агент научился выполнять поставленную задачу. Этот параметр важен для практического применения алгоритмов, потому что определяет количество требуемых ресурсов и времени для получения приемлемого уровня эффективности. Сравнение алгоритмов проводилось на экспериментальной платформе Gymnasium [6] в среде CartPole-v0. В данной среде успешное выполнение задачи признаётся, когда за последние 100 эпизодов агент в среднем набрал суммарную награду не менее 195 при максимальной 200. Данная платформа является удобным инструментом для первичной отладки на ранних этапах реализации алгоритмов и впоследствии может быть использована при разработке новых методик глубокого обучения с подкреплением.

Выводы. По результатам анализа вышеупомянутых алгоритмов метод оптимизации проксимальной политики показал лучший результат по скорости обучения агента.

Список использованных источников:

1. Li Y. Reinforcement learning applications //arXiv preprint arXiv:1908.06973. – 2019.
2. Bai Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback //arXiv preprint arXiv:2204.05862. – 2022.
3. Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with double q-learning //Proceedings of the AAAI conference on artificial intelligence. – 2016. – Т. 30. – №. 1.
4. Sutton R. S. Reinforcement learning: An introduction //A Bradford Book. – 2018.
5. Schulman J. et al. Proximal policy optimization algorithms //arXiv preprint arXiv:1707.06347. – 2017.
6. Brockman G. OpenAI Gym //arXiv preprint arXiv:1606.01540. – 2016.