УДК 004.032.26

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА PPO ДЛЯ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В СРЕДАХ С ДИСКРЕТНЫМ И НЕПРЕРЫВНЫМ ПРОСТРАНСТВОМ ДЕЙСТВИЙ

© Баранов И.С., Савельев Д.А.

Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Российская Федерация

e-mail: i1baranov9@gmail.com

На данный момент обучение с подкреплением является одним из лучших методов для обучения агентов в неизвестной ему среде (в том числе автопилотов [1] и роботов [2]). Тестирование таких агентов в реальной жизни является достаточно дорогостоящим, из-за чего игровые среды стали площадкой для изучения алгоритмов обучения с подкреплением. Кроме того, данные методы нашли применение и в игровой индустрии.

В данной работе представлен один из новейших способов тестирования игр – тестирование при помощи обучения с подкреплением. Данный способ является универсальным и недорогим, поэтому в будущем он может стать важным инструментом для тестирования. При этом исследования проводились в игре жанра Match-3, который является одним из самых популярных жанров игр в мире.

Для обучения тестирующего агента был выбран алгоритм PPO [3], оно проводилось в визуальной среде Match-3 от Unity Machine Learning Agents (Unity ML-Agents) [4]. На рисунке 1 представлено изображение среды для обучения, цель агента – собрать в ряд или столбец от 3 до 5 одинаковых объектов.

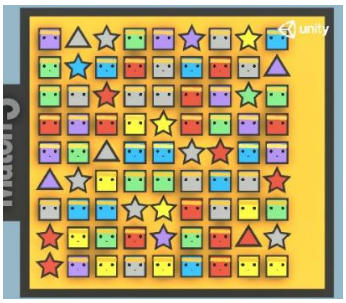

Рис. 1. Match-3 среда от Unity ML-Agents

Как правило, для обучения в визуальной среде применяется сверточная нейронная сеть. В данной работе была исследована зависимость награды агента от количества сверточных слоев в такой сети. Структура базовой нейронной сети для данного исследования показана на рисунке 2 [5].

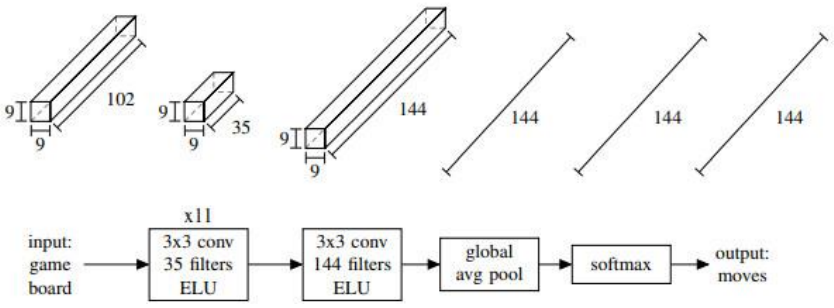

Рис. 2. Структура базовой нейронной сети

Было уменьшено количество сверточных слоев в базовой нейронной сети [5], а затем постепенно увеличивалось. Исследование показало, что при увеличении количества сверточных слоев до определенного момента увеличивается награда, получаемая агентом за эпизод. На рисунке 3 представлены графики зависимости награды агента за эпизод от шага обучения, при этом график красного цвета показывает зависимость при 3 сверточных слоях, а график розового цвета – при 10.

Рис. 3. Зависимость награды агента от шага обучения

Было выявлено, что при увеличении количества сверточных слоев обучение проходит медленнее, но в итоге награда, получаемая агентом, становится больше. Так, увеличение количества сверточных слоев до 10 позволяет увеличить начальную награду на 10 %.

Библиографический список

1. Анцыферов С.С. Проблемы искусственного интеллекта // Проблемы искусственного интеллекта. 2015. № 0(1). С. 5–12.

2. Anderson E. Playing smart artificial intelligence in computer games / E.F Anderson // Proceedings of zfxCON03 Conference on Game Development. 2003. URL: https://core.ac.uk/download/pdf/9599273.pdf (дата обращения: 10.01.2021).

3. Proximal Policy Optimization Algorithms / J. Schulman [et al.] // arxiv.org. 2017. URL: https://arxiv.org/pdf/1707.06347v2.pdf (дата обращения: 25.04.2021).

4. Github Unity Technologies Web Site. Access mode: https://github.com/Unity-Technologies/ml-agents/tree/master/gym-unity (дата обращения: 25.04.2021).

5. Gudmundsson S. Human-Like Playtesting with Deep Learning // IEEE Conference on Computational Intelligence and Games. 2018. URL: https://www.researchgate.net/publication/328307928_Human-Like_Playtesting_with_Deep_Learning (дата обращения: 25.04.2021).