По графику награды можно заметить, что примерно после 40 попытки агент понимает, как нужно задействовать моторы коптера, чтобы взлететь, о чем свидетельствует уровень награды, который в последующие эпизоды колеблется около максимального значения.

Таким образом, в данной статье разобрано одно из применений RL для задач коптеров. Подобных задач, стоящих на стыке техники, транспорта и ОсП, много: беспилотные автомобили, шагающие роботы, интеллектуальные манипуляторы и т.д. [7]. Поэтому, дальнейшие исследования в рамках этих направлений могут привести к созданию умных машин, способных обучаться, ставить и решать задачи самостоятельно. Это позволит автоматизировать промышленность, городской транспорт, а также расширит технические возможности для космических исследований.

1. Хамзатов М. Особенности развития беспилотной авиации в современных условиях [Электронный ресурс]. – режим доступа: URL: <http://uav.ru/articles/hmm.pdf> (дата обращения: 07.07.2020)
2. Квадрокоптер для любительской съемки [Электронный ресурс]. – режим доступа: URL: <http://infocopter.ru/kvadrokoptery-dlya-lyubitelskoj-semki-obzor-modelej/> (дата обращения: 10.06.2020)
3. Сюрмаков Р. Применение машинного обучения и Data Science в промышленности [Электронный ресурс]. – режим доступа: URL: <https://vc.ru/ml/79368-primenenie-mashinnogo-obucheniya-i-data-science-v-promyshlennosti> (дата обращения: 07.07.2020)
4. Никитин П.В., Баркова А.С. Математические методы для определения оценки резерва в страховании. Часть 1. Метод цепной лестницы // Вестник евразийской науки. 2019. Т. 11. № 6. С. 43.
5. Курама В., Алла С. Обучение с подкреплением на языке Python [Электронный ресурс]. – режим доступа: URL: <https://habr.com/ru/company/piter/blog/434738/> (дата обращения: 26.06.2020)
6. Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra. Continuous control with deep reinforcement learning. URL: <https://arxiv.org/abs/1509.02971> (дата обращения 25.06.2020)
7. Обучение с подкреплением / Р.С. Саттон, Э.Г. Барто; пер. с англ. – М.: БИНОМ. Лаборатория знаний, 2011. – 399 с.

Моисеев И.Ю.

Тестирование метода совместного использования параметров в мультиагентном глубоком обучении с подкреплением

*ФГБОУ ВО «Финансовый университет при Правительстве Российской Федерации»
(Россия, Москва)*

doi:10.18411/sciencepublic-12-07-2020-03

idsp: sciencepublic-12-07-2020-03

*Научный руководитель
Никитин П.В.*

Введение

Мультиагентное обучение с подкреплением является активно развивающейся ветвью машинного обучения, и имеет множество прикладных областей, от имитации рынка до системы управления дроном. [1-2]

Большинство моделей обучают агентов по независимым политикам. В результате каждый агент может действовать независимо от других агентов и центрального контроллера. Обучение подобных агентов сталкивается с проблемами из-за нестационарности среды.

Нестационарная среда – среда, которая изменяется в процессе обучения. Нестационарность сильно усложняет обучение, ведь знания агента о среде со временем устаревают.

В работе [3] был предложен способ обучения агентов методом совместного использования параметров. При этом методе агенты имеют часть общих параметров – вплоть до того, что они могут обучаться одной политике. Это помогает агентам получить информацию друг о друге быстрее и тем самым ускорить процесс обучения. Это подтверждает теорема, доказанная в работе [3]. В данной работе рассматривается реализация обучения модели с применением независимых агентов и агентов, применяющих совместное использование параметров. Сравнивается эффективность рассмотренных методов.

Реализация модели и сравнение результатов.

Реализуем модель языке Python с использованием библиотек Ray и TensorFlow. Обучение будем проводить в игре Multiwalker.

Multiwalker представляет собой простую 2D игру, в которой агенты – 3 робота, держат на своей спине блок. Задача роботов – пройти как можно большее расстояние не уронив блок. Каждый агент управляет двумя ногами, имеющими по 2 точки сцепки каждая – все пространство действий содержит в себе возможные движения ног и представляет собой 4-х элементный вектор. Пространство обозрения это 32-х элементный вектор. Роботы получают информацию с помощью симулируемого лидара. Восприятие среды у агентов ограничено, они видят лишь часть среды перед собой. Каждый агент получает небольшую награду за прохождение участка пути (1 единица за 1 условное деление пути). Штраф дается за движение в обратном направлении (по такой же цене) а также за падение блока (100 единиц). Одна игровая сессия включает в себя 500 шагов.

Рисунок – игра Multiwalker

В качестве метода поиска оптимальной стратегии используем Proximal Policy Optimization (PPO). [4] Данный метод относится к методам Policy Optimization – Actor critic.

В методе PPO на вход нейросети подаются наблюдения агентов. На выходе же оказываются действия агентов и оценка качества этих действий. Причем оценка является относительной – это критерий того, насколько ситуация станет лучше после совершенного действия по сравнению с тем что есть сейчас. Эта оценка в дальнейшем используется для корректировки весов нейросети. Для недопущения резких скачков во время обучения максимальное значение модуля оценки ограничивают небольшим числом. На PPO в настоящее время обучают большинство агентов в мире обучения с подкреплением.

Рисунок – схема метода PPO

Нейросеть представляет собой полносвязную сеть с двумя скрытыми и одним выходным слоями. На вход подается тензор содержащий 32-х элементные векторы – наблюдения агента. На первом слое находится 400 нейронов, на втором – 300. Оба слоя содержат функцию активации Relu – отсечение отрицательной части. На выходе сети слой без функции активации. Нейросеть выдает тензоры, содержащие 4-х элементные векторы.

Рисунок – схема нейронной сети

При обучении использовались следующие гиперпараметры:

- Число эпох: 400;
- Размер обучающего пакета: 500;
- Размер тестового пакета: 50;
- Мини-пакеты для стохастического градиентного спуска: 100.

Результаты обучения представим в виде графика.

Рисунок – результаты обучения

Из графиков видно, что уже после 220 эпох метод с применением совместного использования параметров показывает лучший результат, чем метод с независимыми агентами. Для наглядности также приведен график случайной стратегии. Наибольшей награды в -93,8 единиц достигает метод PPO с совместным использованием параметров на 395 эпохе.

Вывод

В данной работе рассмотрена реализация двух моделей мультиагентного глубокого обучения с подкреплением. Представлены результаты обучения агентов. Подтверждены теоретические результаты о более быстром ходе обучения модели с совместным использованием параметром в сравнении с модели с независимыми агентами.

1. Gong, C., Wu, D., Gong, N., Qi, R. Multi-agent mixed complementary simulation of natural gas upstream market liberalization in China // Energy. 2020. №200,117535.
2. Hammami, S.E., Afifi, H., Mouncla, H., Kamel, A. Drone-Assisted Cellular Networks: A Multi-Agent Reinforcement Learning Approach // IEEE International Conference on Communications. 2019-May. №8762079.
3. Justin K Terry, Nathaniel Grammel, Ananth Hari, Luis Santos, Benjamin Black, Dinesh Manocha Parameter Sharing is Surprisingly Useful for Multi-Agent Deep Reinforcement Learning // arXiv:2005.13625v1 [cs.LG]. 27 May 2020.
4. Proximal Policy Optimization // OpenAI Spinning Up URL: <https://spinningup.openai.com/en/latest/algorithms/ppo.html> (дата обращения: 10.07.2020).