УДК 004.896

Использование декомпозиции для повышения эффективности обучения с подкреплением

Андриянов В.А (Университет ИТМО)
Научный руководитель – к.т.н., доцент Ведяков А.А. (Университет ИТМО)

В работе рассматривается способ повышения эффективности в задачах обучения с подкреплением, использующий декомпозицию на обучение с подкреплением на расширенном наборе данных и дальнейший перенос результатов на модель с исходным набором данных.

Введение. В настоящий момент проблемой обучения с подкреплением для решения задач робототехники является сложная зависимость между входными данными и желаемыми действиями обучаемого агента. Зашумлённость данных, малая размерность наблюдаемых величин, изображение в качестве входных данных – это все усложняет процесс обучения. В статьях «Deep Reinforcement Learning that Matters» (2019), «Towards Characterizing Divergence in Deep Q-Learning» (2019) показывается, что прямое увеличение размерности нейронной сети может помочь решить данную проблему, но приводит к снижению скорости обучения или вовсе к не обучаемости из-за нестабильности политики во время обучения. Авторы статей «Can Increasing Input Dimensionality Improve Deep Reinforcement Learning?» (2020) и «Training Larger Networks for Deep Reinforcement Learning» (2021) добились лучших результатов в данном направлении исследований. Их основная идея – это увеличение размерности наблюдаемых величин за счёт использования Dense блоков в архитектуре сети.

Основная часть. В работе «Learning by cheating» (2019) описывается подход упрощения задачи визуального автономного вождения путем декомпозиции обучения на два этапа. Первый – обучение экспертной политики, имеющей доступ к привилегированным данным, доступным только в симуляционной среде: истинное положение автомобиля и препятствий на дороге. Второй – обучение политики, стремящейся повторить действия экспертной и имеющей доступ только к реальным данным: изображениям с камер автомобиля. В рамках области обучения с подкреплением дальнейшее развитие данный метод получил в следующих работах:

- «Learning Quadrupedal Locomotion over Challenging Terrain» (2020) – робастный контроллер для четвероного робота, оценивающий параметры контакта лап, таких как сила контакта, коэффициент трения, форма поверхности;
- «Learning Deep Sensorimotor Policies for Vision-based Autonomous Drone Racing» (2022) и «Learning Perception-Aware Agile Flight in Cluttered Environments» (2022) – навигация беспилотного летательного аппарата в сложном окружении, с использованием данных со стереокамеры или датчика глубины.

В данных работах хорошо отражена суть данного подхода – быстрое обучение политики на простых данных при помощи обучения с подкреплением и дальнейший перенос на конечную политику при помощи методов Imitation Learning, таких как Behavioral Cloning (BC), Dataset Aggregation (DAgger) «A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning» (2011). Проведены тесты алгоритма на симуляционных средах из библиотеки gym:

- для среды Mountain Car Continuous к наблюдаемым скорости и координаты добавлен угол наклона подвижной тележки;
- для Car Racing вместо цветного изображения 96x96 пикселей был выбран вектор наблюдений, состоящий из линейной и угловой скорости машины и из n целевых точек в локальной системе координат.

На этих данных была обучена экспертная политика, полученные результаты перенесены алгоритмом BC на политику, использующею исходные наблюдаемые величины. Сравнение результатов с применением и без применения описанного подхода проходило усреднено по 10 попыткам, аналогично эксперименту «Benchmarks for Spinning Up Implementations». Исходный код и результаты можно найти в репозитории github.com/Vvlad1slavV/LBC-RL.

Выводы. В работе исследовано влияние применения подхода, описанного в статье «Learning by cheating», к задачам обучения с подкреплением. Результаты показали, что использование двухэтапного обучения с видоизмененными данными позволяет повысить скорость обучения политики и качество выполнения поставленных перед агентом задач. В дальнейшем планируется апробация данного подхода на реальных практических задачах: планирования движения шагающего робота по сложным поверхностям для робота Unitree A1; операция вставки объекта для манипулятора UR5e.