Внутренне мотивированное обучение с подкреплением

Иванов Сергей Максимович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: qbrick@mail.ru

Научный руководитель *— Дьяконов Александр Геннадьевич*

В обучении с подкреплением типичной ситуацией является сильная разреженность функции награды, например, когда агент получает подтверждение успешного решения на последнем шаге эпизода и константный сигнал в остальных ситуациях. Если траектории для инициализирующей (обычно случайной) стратегии оцениваются средой одинаково, ни один алгоритм глубокого обучения с подкреплением [1] принципиально не может начать обучение.

Введение внутренней мотивации означает создание отдельного модуля в обучающей системе агента, ответственного за автоматическую генерацию «виртуальной» награды для описания универсальных вспомогательных задач. Решая их, агент может начать понимать законы мира, не используя основную сколь угодно разреженную функцию награды, которая в данной концепции называется внешней мотивацией. Главным примером универсальной задачи является исследование окружающей среды.

Любопытством называется ошибка модели мира агента. Под моделью мира понимается любая обладающая прогнозирующей способностью функция. Если в некоторой области пространства состояний модель мира делает неверный прогноз о будущем, это сигнализирует о непонимании агента законов окружения, и агент должен быть мотивирован вернуться в эту область в ходе дальнейшего обучения с целью набрать больше прецедентов для улучшения модели. Соответственно, любопытство является хорошим сигналом для внутренней мотивации.

В простейшем виде, в качестве модели мира может использоваться модель прямой динамики, которая по состоянию и действию предсказывает следующее состояние. Напрямую такой подход сопряжён с рядом проблем, хотя бы потому, что исходное пространство состояний содержит огромное количество трудно предсказуемой излишней информации.

Основной фундаментальной проблемой любопытства является так называемая проблема «шумного телевизора». В среде в силу

стохастичности функции переходов могут присутствовать источники принципиально непредсказуемой нерелевантной для агента информации (простым примером может являться гауссовский шум на сенсорах или генератор случайных текстур в видеоиграх). Непредсказуемость приводит к не уменьшающемуся со временем поощрению агентом самого себя и вызывает эффект прокрастинации, когда агент отвлекается от основной задачи на источники шума.

Для защиты от эффекта [3] модель прямой динамики $f\colon \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ строится в пространстве $\mathbb{R}^d$ представлений, полученных из некоторой функции $\varphi\colon \mathcal{S} \to \mathbb{R}^d$, называемой фильтром. Обучить фильтр φ можно, строя на его выходах модель обратной динамики $g\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{A}$, пытающейся по отфильтрованным представлениям состояния и следующего состояния предсказывать действие, которое привело к переходу между ними. По построению на выходе из такого фильтра останется информация только о тех объектах среды, с которыми агент может непосредственно провзаимодействовать, на которые может повлиять своими решениями. Итого, генерация внутренней мотивации сводится к решению двух вспомогательных задач регрессии[1]:

$$\mathbb{E}_{s,a,s'} \|g(\varphi(s), \varphi(s')) - a\|_2^2 \to \min_{g,\varphi}$$

$$\mathbb{E}_{s,a,s'} \|f(\varphi(s), a) - \varphi(s')\|_2^2 \to \min_{f}$$

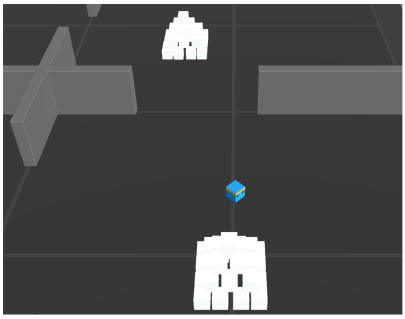
Ошибка последнего функционала используется в качестве любопытства и полагается значением сигнала внутренней мотивации агента.

В работе исследуются возможности данной концепции и её потенциала для решения задач с разреженной функцией награды. В качестве тестовой среды для экспериментов рассматривается задача Unity ML Agents: Pyramids [2], описанная на рис. 1, которую алгоритмы без модуля внутренней мотивации решить неспособны. В экспериментах удалось увидеть, что введение любопытства позволяет решить задачу для всех рассматривавшихся базовых алгоритмов обучения с подкреплением (A2C, PPO и QR-DQN) [1].

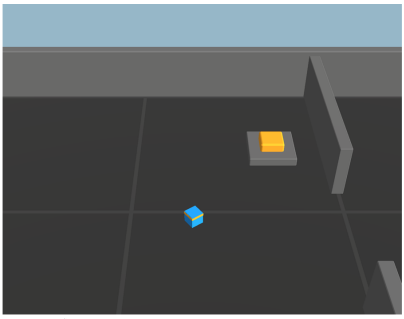
Результаты получены в рамках проекта «Центра хранения и анализа больших данных» МГУ имени М.В.Ломоносова по договору с Фондом поддержки проектов НТИ № 13/1251/2018 от 11.12.2018.

[1] для непрерывных пространств действий $\mathcal{A}$; для дискретных же пространств действий модель обратной динамики является классификатором и для неё должна оптимизироваться соответствующая функция потерь.

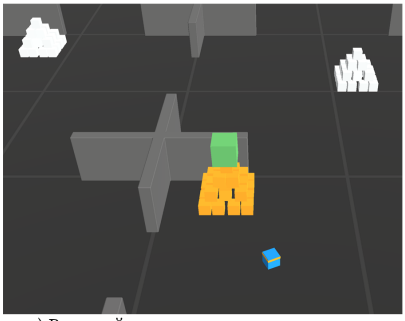
Иллюстрации



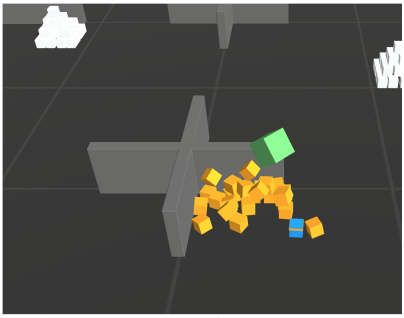
а) В начале эпизода агент располагается в случайной из девяти комнат. На вход ему поступает 172 значения сенсоров о дальности до объектов каждого типа вдоль набора направлений (ray-casts).



б) Агент должен найти случайно расположенную кнопку и нажать её (пересечься с ней). Нажатие не приводит к награде от среды.



в) В случайном месте среды появится пирамидка выделенного цвета (агент различает её от остальных пирамидок в среде).



г) Агенту нужно разбить пирамидку и добраться до большого куба, который упадёт с вершины. Только после этого агент получит награду от среды +1.

Задача Unity ML Agents: Pyramids.

Литература

1. Ivanov S., D'yakonov A. Modern Deep Reinforcement Learning Algorithms //arXiv preprint arXiv:1906.10025. – 2019.
2. Juliani A. et al. Unity: A general platform for intelligent agents //arXiv preprint arXiv:1809.02627. – 2018.
3. Pathak D. et al. Curiosity-driven exploration by self-supervised prediction //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2017. – С. 16-17.