# Имитационное обучение с наблюдений с использованием независящей от действий функции переходов

***Цыпин Артём Андреевич***

*Студент*

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

*E-mail:* tsypinartem9@gmail.com

***Научный руководитель*** *— Майсурадзе Арчил Ивериевич*

В последнее время обучение с подкреплением было успешно применено для решения ряда сложных задач. Процесс обучения подразумевает взаимодействие агента со средой, в течение которого агент получает от среды отклик, который также называется наградой. Агент обучается максимизировать полученную награду. Однако, проектирование подобной награды в большинстве нетривиальных случаев является крайне сложной задачей.

В связи с этим возникает идея использовать экспертные демонстрации для ускорения и стабилизации процесса обучения. Классическая постановка задачи имитационного обучения подразумевает наличие информации как о состояниях, в которых побывал эксперт, так и о действиях, которые он совершил. Это ограничивает область применения подобных методов, так как существует огромное количество экспертных данных, не содержащих информации о действиях агента.

Существует множество подходов к решению данной проблемы. В [1] рассматривается распределение $p^s_\pi(s, s')$ переходов агента, действующего согласно политике $\pi$. Предлагается использовать состязательное обучение для обучения политики, имеющей то же распределение переходов, что и политика эксперта. Политика, обучающаяся копировать эксперта, является генератором, в то время как дискриминатор $D$ обучается присваивать значения, близкие к 1, переходам агента и значения, близкие к 0, переходам эксперта. Для обучении политики используется алгоритм обучения с подкреплением с наградой $-(\mathbb{E}_\tau[\log(D(s, s'))])$.

Данный подход, однако, обладает рядом недостатков, присущих большинству методов, использующих состязательное обучение: нестабильность при обучении и необходимость в очень большом объеме данных. Чтобы частично решить данные проблемы в настоящей работе предлагается обучать дискриминативную модель $p(s' \mid s)$ на траекториях эксперта:

$$p(s' \mid s) = \frac{exp(\varphi_1(s)^T \varphi_2(s')))}{\sum_{s'' \in S} exp(\varphi_1(s)^T \varphi_2(s''))} \approx \frac{exp(\varphi_1(s)^T \varphi_2(s')))}{\sum_{s'' \in S_{expert}} exp(\varphi_1(s)^T \varphi_2(s''))} \tag{1}$$

В (1) $s'$ – следующее после $s$ состояние, $S$ – множество всех состояний, $S_{expert}$ – множество состояний, которые встречаются в экспертных демонстрациях, $\varphi_1, \varphi_2$ – нейросети. Такую модель можно использовать в качестве награды для агента:

$$R = \sum_{t=0}^{T} r_t = \sum_{t=0}^{T} \log\left(p(s_{t+1} \mid s_t)\right) = \log \prod_{t=0}^{T} p(s_{t+1} \mid s_t). \tag{2}$$

Обучать $\varphi_1, \varphi_2$ предлагается с помощью функции потерь, рассмотренной в [2–3]:

$$L = -\sum_{i=1}^{B} \log \frac{exp(\frac{\varphi_1(s_i)^T \varphi_2(s_{i+1})}{\tau})}{\sum_{k=1}^{B} exp(\frac{\varphi_1(s_i)^T \varphi_2(s_{k+1})}{\tau})} \tag{3}$$

В (3) $\tau$ – температура. Проведенные в средах Acrobot, CartPole и MountainCar эксперименты показывают, что:

- Дискриминативная модель $p(s' \mid s)$ стабильно обучается на всех опробованных средах. Качество модели незначительно различается при обучении на наборе данных из 100, 1000 и 10000 экспертных траекторий, то есть модель не требует огромного количества данных.
- Предложенная награда позволяет агентам достигать качества эксперта за приемлемое количество взаимодействий со средой. В некоторых средах агент с подобной наградой обучается быстрее RL агента с наградой из среды. Более того, награда позволяет успешно обучать агентов в «hard-exploration» средах.

**Литература**

1. Torabi F., Warnell G., Stone P. Generative adversarial imitation from observation //arXiv preprint arXiv:1807.06158. – 2018.
2. Oord A., Li Y., Vinyals O. Representation learning with contrastive predictive coding //arXiv preprint arXiv:1807.03748. – 2018.
3. Chen T. et al. A simple framework for contrastive learning of visual representations //International conference on machine learning. – PMLR, 2020. – С. 1597-1607.