Влияние состава наблюдений окружающей среды в задаче приобретения навыков передвижения в трёхмерном пространстве при использовании алгоритмов обучения с подкреплением

Д.А. Козлов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
djoade100@gmail.com

В.В. Мясников
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
vmyas@geosamara.ru

***Аннотация*—В работе исследуется влияние состава наблюдений окружающей среды на процесс обучения «двуногого» мехатронного объекта навыкам передвижения в трёхмерном пространстве. Исследования проводятся в среде игрового движка Unity с использованием пакета ML-Agents. В качестве алгоритма обучения был выбран Soft Actor Critic, как один из наиболее эффективных современных алгоритмов обучения с подкреплением (RL), показавший наибольшую эффективность на наборе аналогичных задач. Показано, что состав наблюдений может радикально менять скорость обучения и даже замедлять процесс обучения при наличии «избыточных» данных.**

Ключевые слова— SAC, Unity ML-Agents, обучение с подкреплением, симуляция, MDP, POMDP, робототехника.

1. Введение

Одной из задач робототехники является задача приобретения навыков передвижения (обучение передвижению) в трёхмерном пространстве [1]. Для её решения существует множество подходов, среди которых современным и универсальным является метод *обучения с подкреплением* (RL-reinforcement learning) [2]. Особый интерес при обучении передвижению представляют человекоподобные механизмы - «двуногие» автономные мехатронные устройства, поскольку они обладают большим потенциалом для движения по пересеченной местности [3]. С точки зрения RL-методов, обучение подобных механизмов интересно большим количеством степеней свободы (по числу подвижных частей) и, как следствие, большим количеством возможных локальных минимумов в решении: локальные минимумы соответствуют таким устойчивым положениям механизма, из которых тяжело выбраться [4].

В настоящей работе исследуется влияние объема и состава наблюдений (данных, информации) об окружающей среде (в терминах RL-алгоритмов – о состоянии) на эффективность решения искомой задачи с использованием RL-метода. В качестве целевого RL-алгоритма выбран Soft Actor Critic, как алгоритм, показавший наибольшую эффективность решения пула аналогичных задач [5]. Исследования проводятся в виртуальной среде игрового движка Unity с использованием пакета ML-Agents.

Структура работы следующая. В разделе 2 кратко описывается используемый алгоритм, в разделе 3 описывается постановка эксперимента и его результаты. В конце работы приведено заключение, обобщающее результаты эксперимента.

2. Метод и алгоритм обучения

Обучение с подкреплением – способ машинного обучения, при котором система обучается, взаимодействуя со средой. В отличие от типовых методов машинного обучения с учителем, требующих размеченных наборов данных или указания предопределенных ответов, в RL-методах необходимо указывать лишь награду, которую получает алгоритм обучения (в RL-методах именуется как *агент*) в зависимости от его действий в том или ином состоянии. Состав наблюдений, в свою очередь, описывает *состояние* системы «агент+среда» в каждый момент времени. Задача агента состоит в том, чтобы максимизировать суммарную награду, формируемую в результате всей последовательности его действий.

Задача обучения с подкреплением формализуется как Марковский процесс принятия решений. Марковский процесс принятия решений представляет собой кортеж, (S, A, R, P, ρ_0), где S — множество всех допустимых состояний, A — множество всех допустимых действий, $R: S \times A \times S \to \mathbb{R}$ — функция вознаграждения, где $r_t = R(s_t, a_t, s_{t+1})$, $P: S \times A \to \mathcal{P}(S)$ - функция вероятности перехода, где $P(s'|s,a)$ - вероятность перехода в состояние s', если вы начинаете в состоянии s и предпринимаете действия a,ρ_0 — начальное распределение состояний. Название «Марковский процесс принятия решений» относится к тому факту, что система подчиняется марковскому свойству: переходы зависят только от самого последнего состояния и действия, а не от предыдущей истории. Под термином (*стохастическая*) *политика* понимают стратегию поведения агента, то есть правило выбора действия в определенном состоянии: $\pi(a|s)$. На практике политику можно реализовать как таблично, так и в виде параметризируемого отображения.

Soft Actor Critic [5] (SAC) — это современный RL-алгоритм, предложенный Google и UC Berkley и позиционированный авторами для использования в

задачах робототехники. Алгоритм агрегирует подходы стохастической оптимизации политики с градиентными методами (англ.: Policy Gradient Algorithms) [6].

Центральной особенностью SAC является регуляризация энтропии. Политика настраивается так, чтобы максимизировать компромисс между ожидаемой доходностью и энтропией, характеризующей меру случайности в политике. Такой подход тесно связан с идеей поиска компромисса между разведкой и эксплуатацией [7], позволяет предотвратить преждевременную сходимость политики к плохому локальному оптимуму.

3. Эксперименты и результаты

А. Постановка экспериментов

В качестве среды симуляции для экспериментов был использован игровой движок Unity совместно с пакетом ML-Agents [8]. Для проведения экспериментов была создана собственная модель подвижного устройства (агент), названная SimplestBipedal: агент в данной среде имеет две конечности, каждая из которых подвижна в двух суставах. При этом в верхнем суставе присутствует подвижность в двух плоскостях, а в нижнем – лишь в одной. Такая модель подразумевает высокую сложность задачи обучения. Задачей агента являлось достижение «цели» - предопределенного места в 3D среде.

Наблюдения, передаваемые агенту в каждом эксперименте, выбирались в соответствии с таблицей 1. Помимо указанной в ней информации агенту всегда передавалась информация о направлении к цели. *Награда* вычислялась как величина, пропорциональная скорости движения агента и направлению движения к цели.

Таблица I. Состав экспериментов

Передаваемая агенту информация	**Номер эксперимента**								
	0	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Перемещение частей в глобальных координатах	+	+	+	+	+	+		+	+
Поворот частей в глобальных координатах	+	+	+	+	+		+	+	+
Перемещение частей в локальных координатах			+	+	+	+	+		
Поворот частей в локальных координатах	+		+	+	+		+	+	
Скорость частей в глобальных координатах	+			+	+	+	+		
Скорость частей в локальных координатах				+	+			+	
Положения и углы суставов					+				+
Сила прикладываемая в суставах					+				+

Б. Результат эксперимента

Результаты кратко представлены в таблице 2. В таблице представлены усредненные значения награды к шагу номер 10 млн. Максимальная награда достигнута для пятого эксперимента, она составила 6000.

Из представленных результатов видно, что информация о положении, угле суставов, а также о силе, приложенной к ним, является (с точки зрения достигаемого результата и скорости обучения) избыточной и оказывающей негативное влияние на качество решения задачи, хотя, очевидно, полезна для решения задачи обучения.

Таблица II. Результаты экспериментов

	Номер эксперимента								
	0	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Награда (в тысячах)	5	3	3	5,5	4	6	3	5,7	2

4. Заключение

По результатам проведенных исследований становится понятно, что полезные с содержательной точки зрения наблюдения могут оказывать негативное влияние на процесс обучения с использованием RL-методов. Таким образом, необходимым этапом решения задачи обучения с подкреплением в реальной среде (где «стоимость» процесса обучения и самого агента несоизмеримо выше, чем в среде имитационной) становится анализ доступного состава наблюдений и отбор тех, которые целесообразно использовать для RL-обучения.

Остался открытым вопрос, является ли «негативное» влияние наблюдений объективной или субъективной характеристикой, то есть оказывает ли влияние на этот выбор используемый RL-метод.

Благодарности

Работа выполнена при поддержке Российского научного фонда (проект № 21-11-00321, https://rscf.ru/en/project/21-11-00321/).

Литература

[1] Peters, J. Towards Motor Skill Learning for Robotics / J. Peters, K. Mülling, J. Kober, D. Nguyen-Tuong, O. Kroemer. – 2009. – P. 469-482.

[2] Haarnoja, T. Learning to Walk via Deep Reinforcement Learning / T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, S. Levine // ArXiv: 1812.11103, 2019.

[3] Atlas™ // Boston Dynamics [Electronic resource]. – Mode of access: https://www.bostondynamics.com/atlas (17.04.2022).

[4] Kozlov, D. Comparison of Reinforcement Learning Algorithms for Motion Control of an Autonomous Robot in Gazebo Simulator / D. Kozlov // International Conference on Information Technology and Nanotechnology (ITNT). – 2021. – P. 1-5. DOI: 10.1109/ITNT52450.2021.9649145.

[5] Haarnoja, T. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor / T. Haarnoja, A. Zhou, P. Abbeel, S. Levine // ArXiv: 1801.01290, 2018.

[6] Silver, D. Deterministic Policy Gradient Algorithms / D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller. – 2019. – P. 9.

[7] Rocca, J. The exploration-exploitation trade-off: intuitions and strategies / J. Rocca // Medium, 2021 [Electronic resource]. – Mode of access: https://towardsdatascience.com/the-exploration-exploitation-dilemma-f5622fbe1e82 (17.04.2022).

[8] Juliani, A. Unity: A General Platform for Intelligent Agents / A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, D. Lange // ArXiv: 1809.02627, 2020.