УДК 519.7

Управление походкой двуногого робота с использованием обучения с подкреплением

И. А. Шаргин

Московский физико-технический институт(государственный университет)

141701, Московская область, г. Долгопрудный, Институтский переулок, д.9.

E-mail: shargin.ia@phystech.edu

Ключевые слова: двуногие роботы, обучение с подкреплением, глубокое обучение

Аннотация: Целью данной работы является обучить модель гуманоидного робота SAHR (Starkit autonomous humanoid robot) устойчиво ходить по ровной поверхности в симуляторе, используя методы машинного обучения. Модель робота была загружена в симулятор Isaac Gym и обучена ходьбе.

1. Введение

Человечество приближается к моменту, когда роботы будут помогать людям не только на заводах, но и в быту. Одним из требований к роботу, который смог бы быть универсальным помощником является умение эффективно перемещаться в пространстве, созданном для людей. Это делает двуногих роботов хорошими кандидатами, так как человек обустраивал среду таким образом, что она больше всего подходит для таких же как он. Для получения устойчивой походки инженеру необходимо учесть непредсказуемость среды. Обучение с подкреплением позволяет роботу через взаимодействие со средой научиться справляться с ее непостоянством и непредсказуемостью, получая таким образом устойчивую походку, неплохо сохраняющую свою эффективность даже при использовании в условиях, отличающихся от тех, в которых был обучен агент. Большинство современных методов, эксплуатирующих обучение с подкреплением, предполагают обучение модели робота в симуляции с дальнейшим переносом этой модели на реального робота.

2. Описание используемых алгоритмов обучения и сред симуляции

Isaac Gym – высокопроизводительная платформа симуляции, созданная специально для решения задач в области робототехники через обучение с подкреплением. Эта платформа позволяет производить сбор данных агентом из среды и обновление политики от начала до конца на видеокарте, что позволяет

сильно ускорить процесс обучения. Legged gym [1] представляет из себя фреймворк для обучения ходьбе моделей роботов с ногами в симуляторе Isaac Gym.

3. Метод

В работе использован алгоритм обучения с подкреплением PPO. Сеть «агент» представляет из себя полносвязную сеть с тремя скрытыми слоями, на вход сетки подаются наблюдения(Таблица 1), на выходе – 14 действий – целевых положений сервомоторов. Эти 14 позиций затем передаються на ПД контроллеры сервомоторов. Сеть «критик» похожа по структуре на сеть агента, за исключением выхода – на выходе одно число – аппроксимация value function. В симуляции параллельно запускаются n моделей, случайным образом выбираются команды на желаемую линейную и угловую скорости из диапазона допустимых команд. Каждый из роботов продолжает идти пока либо не упадет, либо не пройдет время $max\_episode\_length$. При падении или истечении времени для данного робота заканчивается эпизод, происходит ресет робота и случайно выбираются новые значения команд. Команды не меняются в течении одного эпизода. Как только со всех роботов суммарно собрали данных размером с batch size $B = num\_steps \cdot num\_robots$ (что означает каждый робот совершил необходимое количество num_steps), проводится обновление стратегии агента методом обратного распространения ошибки.

Таблица 1. Наблюдения агента.

Наблюдение	размерность
команда на линейную скорость x	1
команда на линейную скорость y	1
команда на угловую скорость	1
угол поворота моторов	14
угловая скорость моторов	14
ускорение корпуса xyz	3
угловое ускорение корпуса xyz	3
проекции g на собственные оси робота	3
предыдущие действия агента	14

3.1. Функция награды

Во время обучения за каждое совершенное действие агент получает награду и изменяет свою политику управления для максимизации награды. Награда агента за шаг в симуляции представляет собой взвешенную сумму наград:

Положительные награды (линейная скорость торса (x, y) и угловая скорость торса (z)) побуждают агента следовать командной скорости. Остальные награды корректируют движение агента, штрафуя его за прыжки, избыточные нагрузки на моторы и резкие движения.

Линейная скорость торса (x, y) – tracking linear velocity – минимизирует разницу между линейными скоростями торса агента, управляющей и действительной. Коэффициент σ предназначен для корректировки значения

Таблица 2. Слагаемые функции награды

Награда	Формула
tracking linear velocity	$\exp(-\|\|lin\_vel - desired\_lin\_vel\|\|^2/\sigma)$
tracking angular velocity	$\exp(-\|\|ang\_vel - desired\_ang\_vel\|\|^2/\sigma)$
joint torques	$\|\|torques\|\|^2$
linear velocity z	$\upsilon^2_{base,z}$
action rate	$\|\|a_{last} - a\|\|^2$
dof acceleration	$\|\|\dot{q}_{last} - \dot{q}\|\|^2$

ошибки, при его увеличении агент способен сильнее отклоняться от целевой скорости.

Угловая скорость торса (z) – tracking angular velocity – аналогична предыдущей награде, но используется для поворота агента вокруг вертикальной оси.

Линейная скорость торса (z) – linear velocity z – штрафует агента за перемещение торса по вертикальной оси, используется при обучении с высокими управляющими скоростями для предотвращения прыжков.

Крутящий момент моторов – torques – минимизирует суммарный крутящий момент моторов, делая походку энергоэффективной.

Ускорение моторов – dof acceleration – штрафует агента за резкие изменения скорости моторов.

Скорость изменения действий – action rate – делает движения агента более плавными, исключает «лишние» действия.

4. Результаты обучения

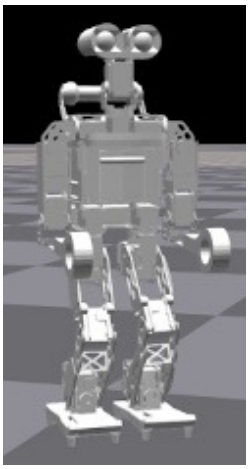

Рис. 1. Кадр походки

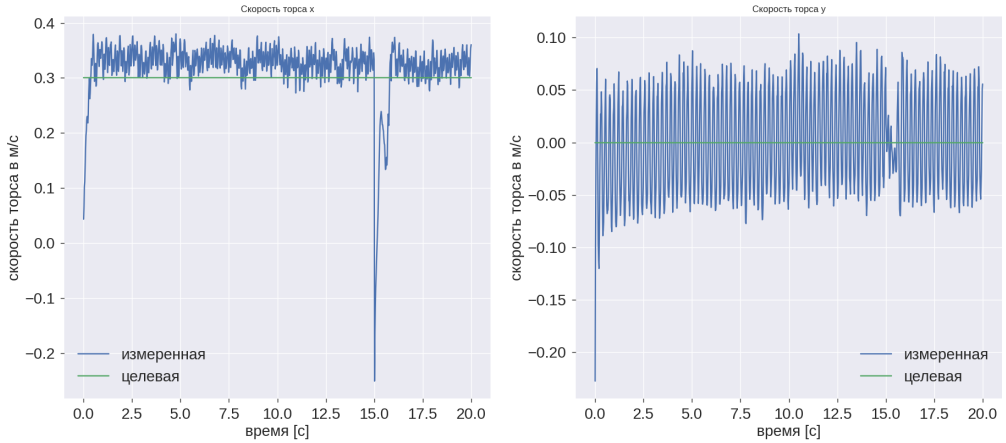

Рис. 2. Следование командной скорости после обучения(слева скорость по оси x, справа – по оси y)

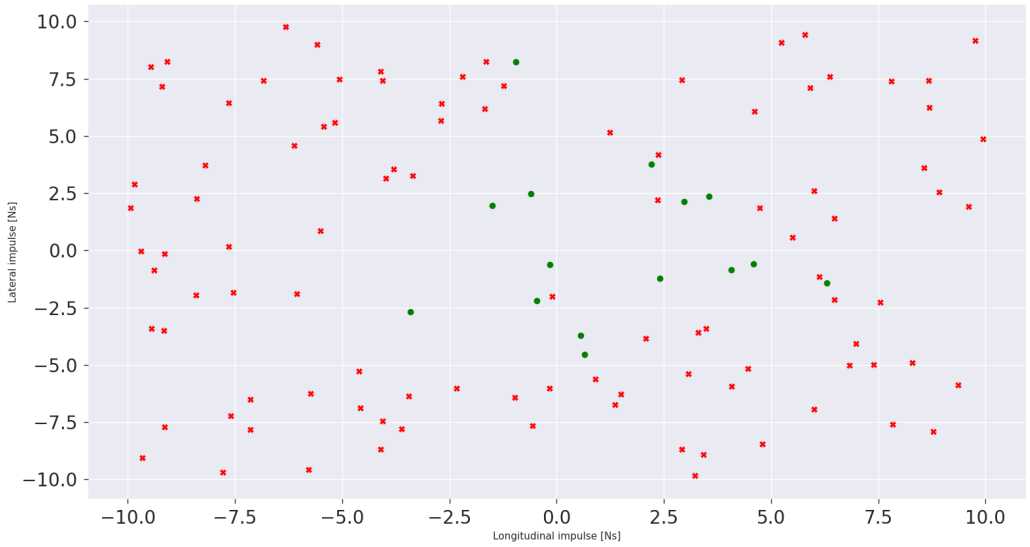

Рис. 3. Устойчивость ко внешним толчкам, красные кресты – робот упал, зеленые точки – устоял

5. Заключение

С разработанными способами оценки качества походки(соответствие командной скорости, энергоэффективность, тест на устойчивость и др.) были подобраны оптимальные параметры обучения. В дальнейшем планируется перенос на реального робота обученных в данной работе нейронных сетей. Обучение с подкреплением является многообещающим подходом к управлению походкой двуногих роботов.

Список литературы

1. N. Rudin, D. Hoeller, P. Reist, M. Hutter, Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. 2021 hrefhttps://arxiv.org/abs/2109.11978 abs/2109.11978
2. V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Mackli, D. Hoeller, N. Rudin, A. Allshire, A. Handa, G. State. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning. 2021. abs/2109.11978
3. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. Proximal Policy Optimization Algorithms. 2017. abs/1707.06347