БОЛЬШИЕ ДАННЫЕ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ФИНАНСОВЫЕ ТЕХНОЛОГИИ И КИБЕРБЕЗОПАСНОСТЬ

БОЛЬШИЕ ВЫЗОВЫ
НАУЧНО-ТЕХНОЛОГИЧЕСКАЯ ПРОЕКТНАЯ ПРОГРАММА

Команда


Сурначёв Никита Владимирович


Ушаков Даниил Сергеевич


Пробочкин Андрей Михайлович


Фомина Полина Алексеевна

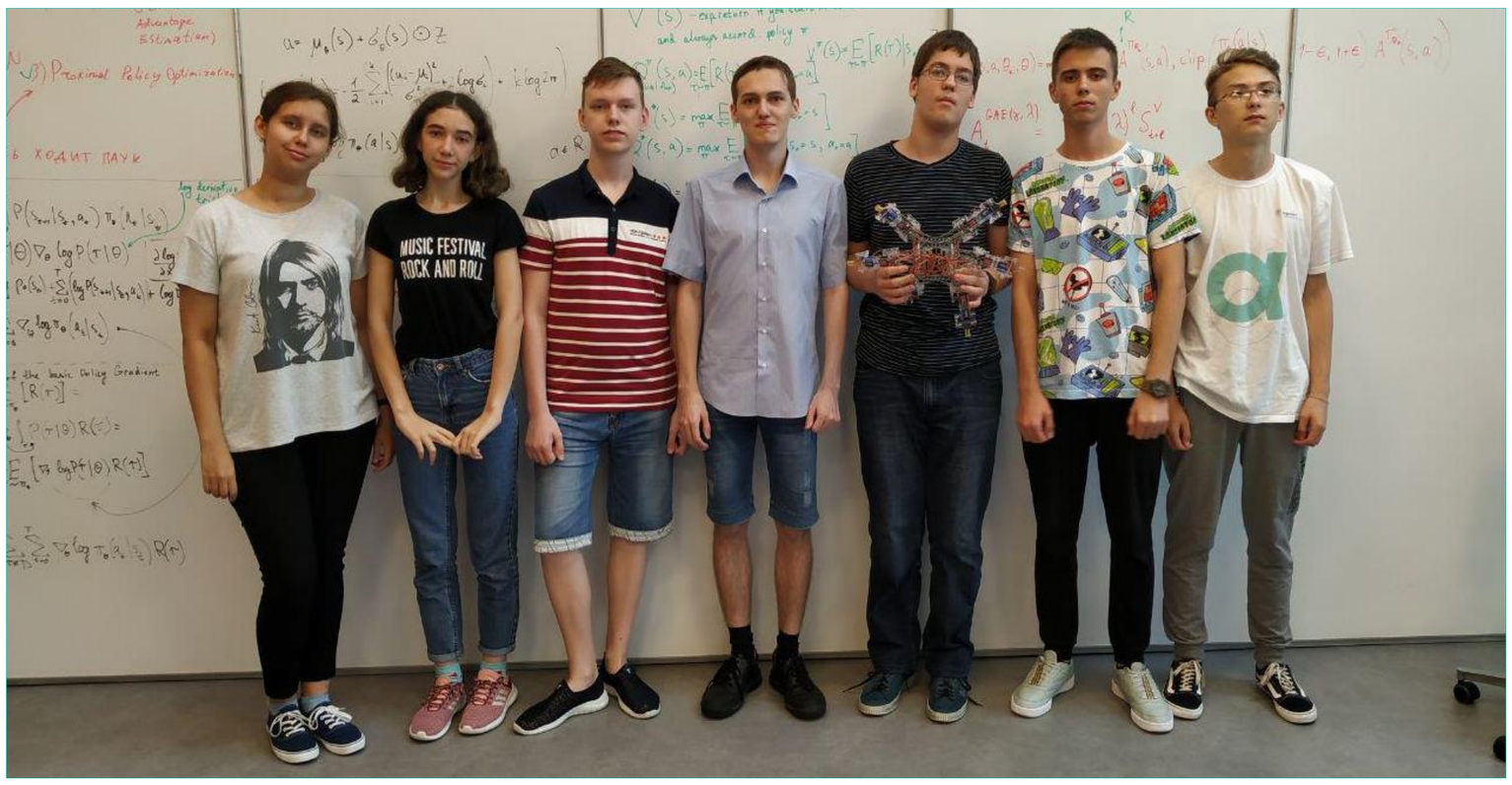

Колесникова Ксения Сергеевна

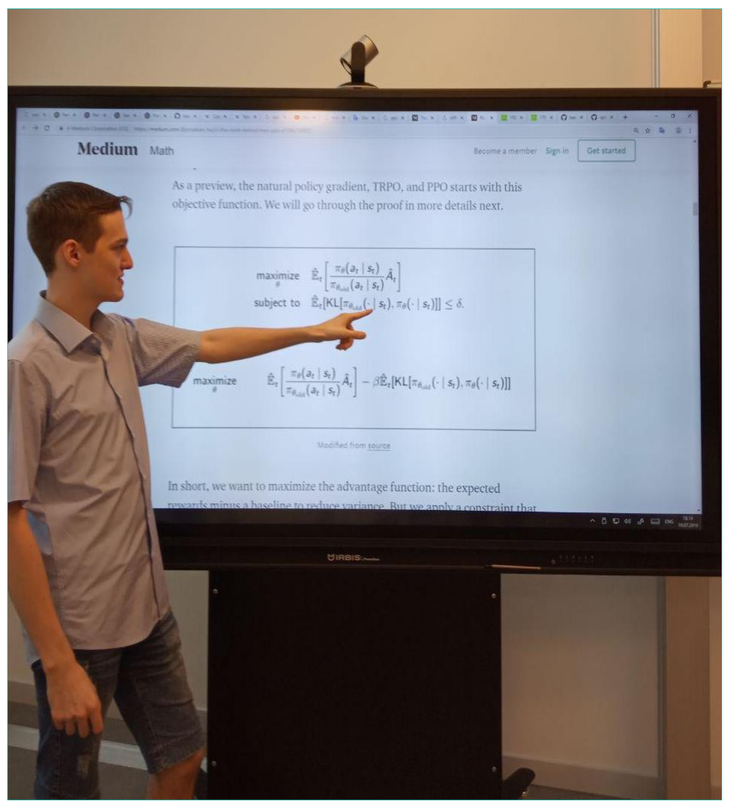

Шибаев Василий Андреевич

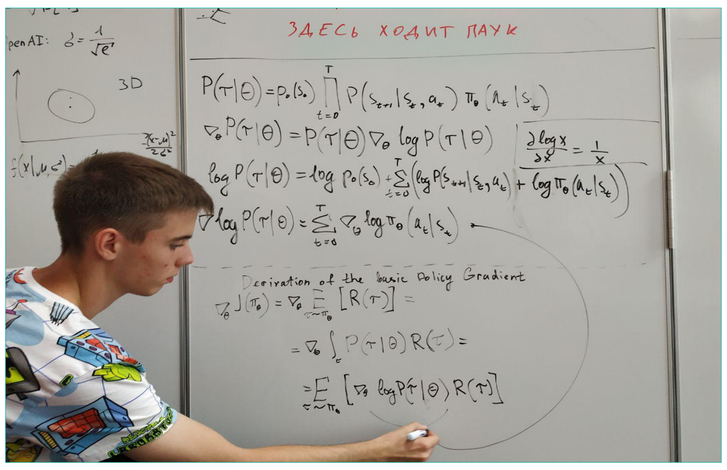
Наставник

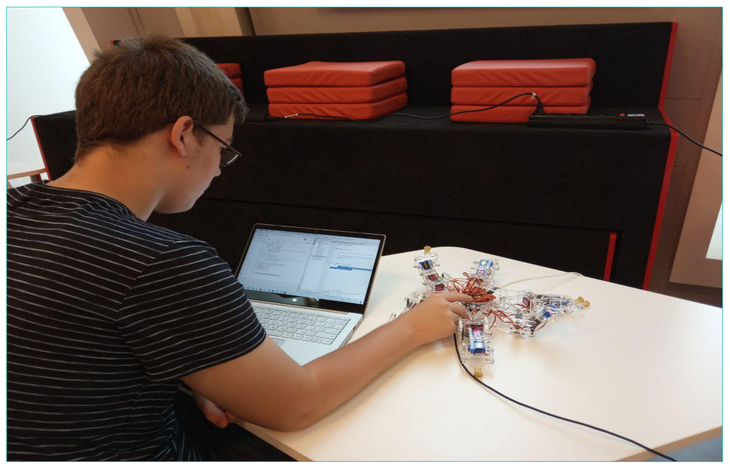

Мурашов Леонид Максимович

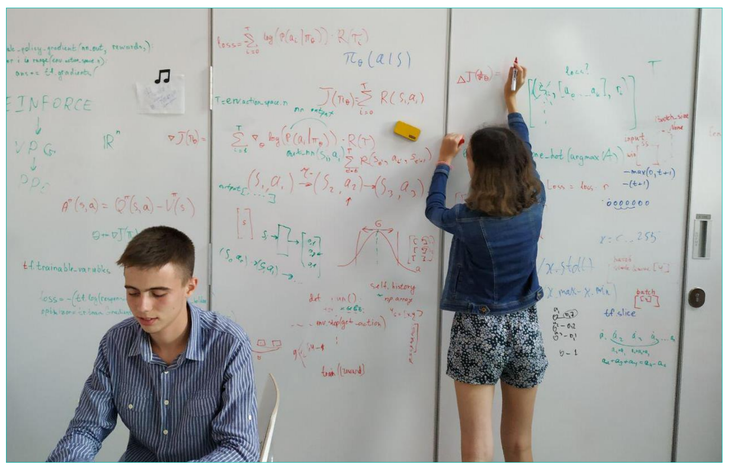
Фотоотчет

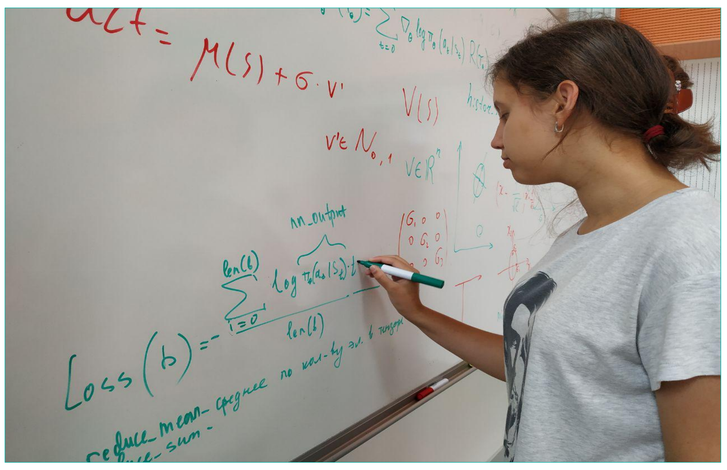












Голосование

БОЛЬШИЕ ДАННЫЕ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
ФИНАНСОВЫЕ ТЕХНОЛОГИИ И КИБЕРБЕЗОПАСНОСТЬ

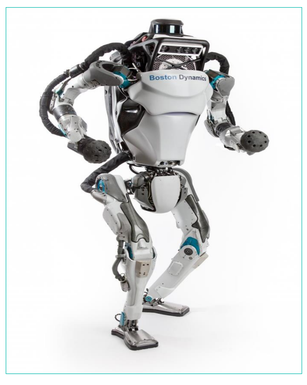
БОЛЬШИЕ ВЫЗОВЫ
НАУЧНО-ТЕХНОЛОГИЧЕСКАЯ ПРОЕКТНАЯ ПРОГРАММА

Автоматизация создания алгоритмов устойчивого движения роботов с помощью обучения с подкреплением

Проблема и актуальность

В современном мире робототехника имеет особое значение. Роботизированные системы можно встретить на каждом шагу: в медицине, сфере производства, транспортных услуг и систем, коммуникаций, социального обслуживания и т.д. Но сейчас стоит задача в создании алгоритма, который позволит роботам вести себя более естественно, приближенно к движениям живых организмов. Это позволит расширить спектр возможного применения роботизированных систем, что значительно упростит жизнь.

Актуальность: Шагающие роботы позволяют решать широкий класс задач, однако в настоящее время они используются только в экспериментальных образцах. Использование шагающего робота целесообразно, когда препятствия непреодолимы для гусеничного и колесного робота. Это могут быть предметы высотой больше чем высота шасси робота или болотистые местности, а применение шасси больших габаритов не приемлемо ввиду использования робота



на пересеченной местности или в городской обстановке, а также высокой потребляемой мощности шасси.

Проблема: На данный момент не существует general системы восприятия и адаптивного управления,

которая бы повышала производительность, надежность и универсальность поведения роботов.

Гипотеза

Reinforcement Learning позволит создать general алгоритм, который самостоятельно обучается в среде естественному хождению.

Партнеры

МФТИ

Deep Learning School

Научные статьи

[1] Playing Atari with Deep Reinforcement Learning, Mnih et al, 2013. Algorithm: DQN.

[2] Asynchronous Methods for Deep Reinforcement Learning, Mnih et al, 2016. Algorithm: A3C.

[3] High-Dimensional Continuous Control Using Generalized Advantage Estimation, Schulman et al, 2015. Algorithm: GAE.

[4] Proximal Policy Optimization Algorithms, Schulman et al, 2017. Algorithm: PPO-Clip, PPO-Penalty.

[5] Policy Gradient Methods for Reinforcement Learning with Function Approximation, Sutton et al, 2000.

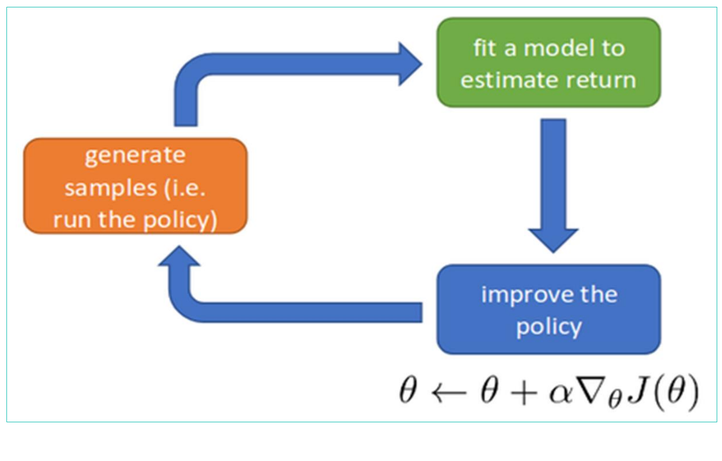
[6] A Natural Policy Gradient, Kakade, 2002

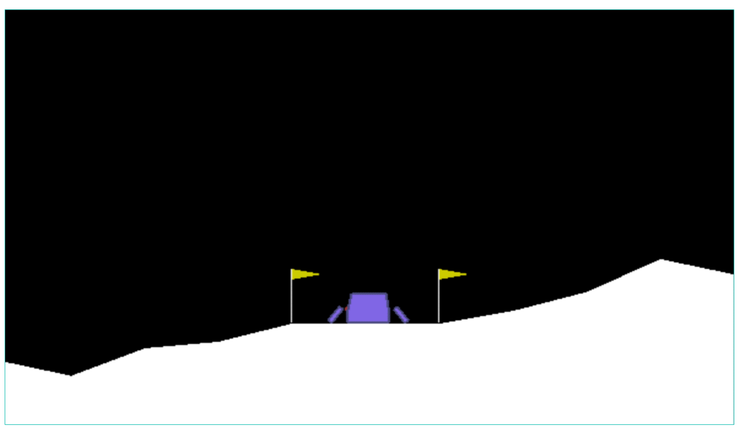
[7] Algorithms for Reinforcement Learning, Szepesvari, 2009.

Решение и результаты

Существует две категории алгоритмов[7]: одни предсказывают **вознаграждение** (**Value based**), другие – **действие агента** (**Policy based**). Алгоритмы первой категории (например, **Deep Quality Network[1], Q-table**) предсказывают награды, которые агент получит в зависимости от выполняемого действия. Благодаря этим методам мы находим наилучшее действие для каждого состояния агента в среде - действие с **наибольшей ценностью**. Это хорошо работает, когда агент может выполнить конечный набор действий (например, **игры Atary**).

Вторая категория алгоритмов (**Policy Gradient[5],[6], A2C[2],[3] и Proximal Policy Optimization[4]**) предсказывает поведение робота. Эти методы полезны в том случае, когда пространство действия непрерывно или стохастично.

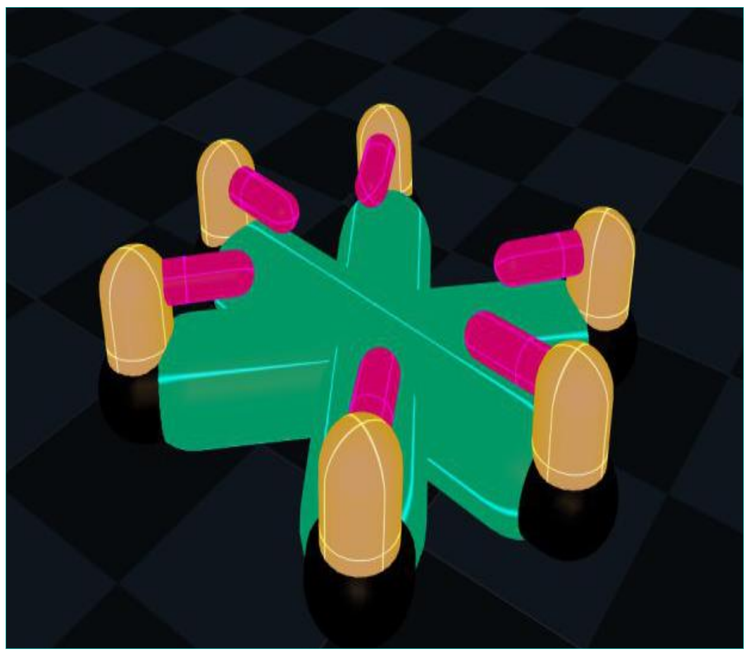




Сначала нами был написан алгоритм **Q-table**, после чего наша команда разделилась на три подгруппы, две из которых реализовывали DQN или Reward-to-go Policy Gradient, а третья занималась технической частью – создавала модель робота в **3D-симуляторе Mujoco** и разрабатывала API для связи робота и написанного алгоритма.

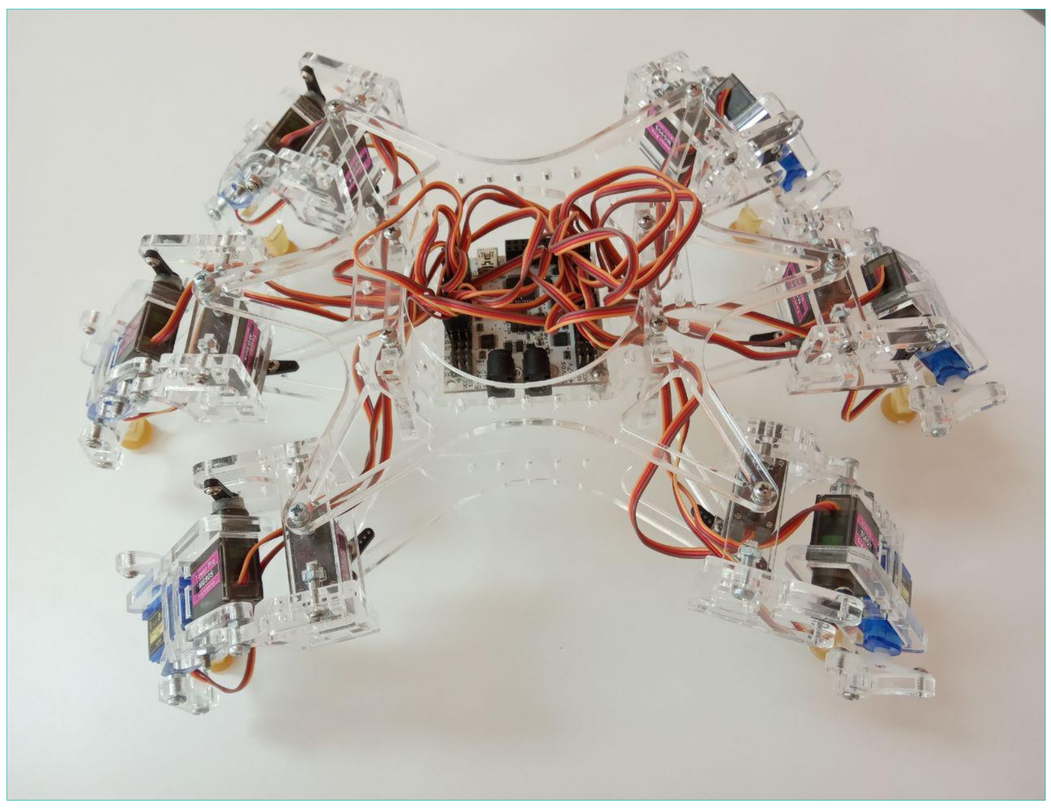
Оба алгоритма имеют недостатки. Reward-to-go Policy Gradient обращает внимание только на конечный результат, то есть на суммарную награду, которую агент получает за прохождение всей траектории, а не на сами действия, которые привели агента к этому результату. Из этого следует вывод, что данный метод **не различает плохие и хорошие действия.**

Поэтому нами было решено реализовать следующий метод, который учитывает приведённые ранее недочёты. **Метод Advantage Actor-Critic (A2C)** состоит из двух нейронных сетей. Первая (actor) предсказывает действие. Вторая (critic) анализирует результат первой сети, вычисляет ценность этого действия в данном состоянии, после чего первая сеть обновляет свои параметры в зависимости от выданных значений второй сети.

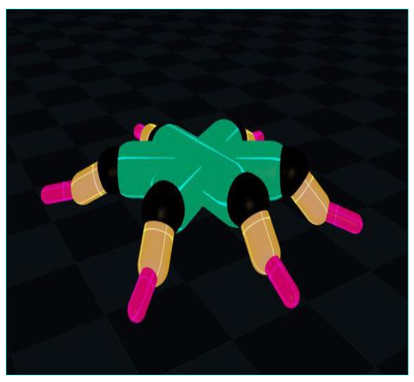


Голосование

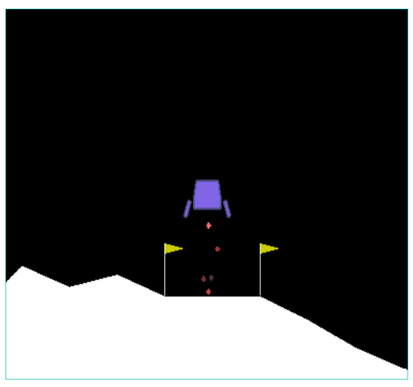
Визуализация процесса работы над проектом и результатов проекта



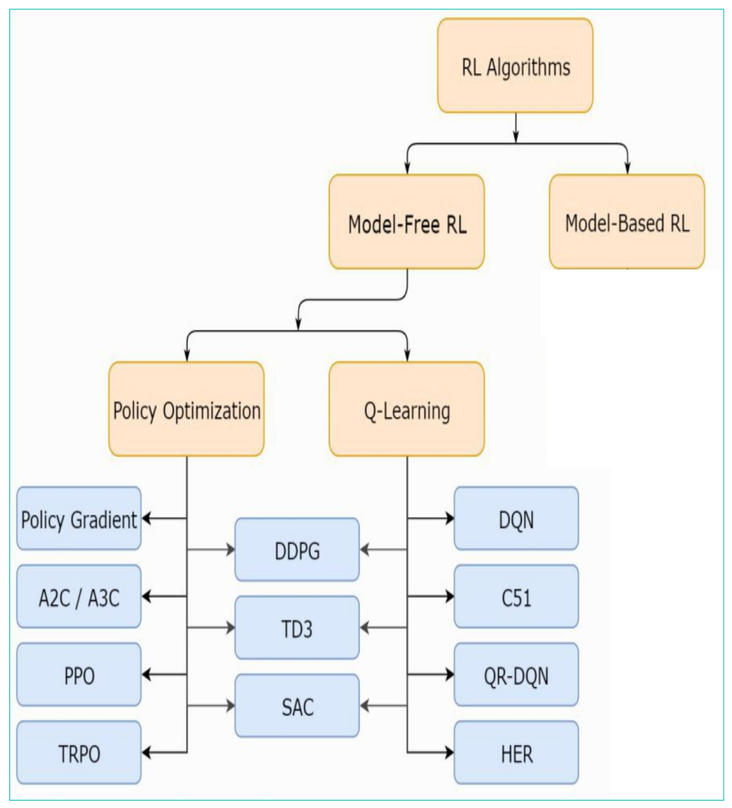
Модель нашего робота



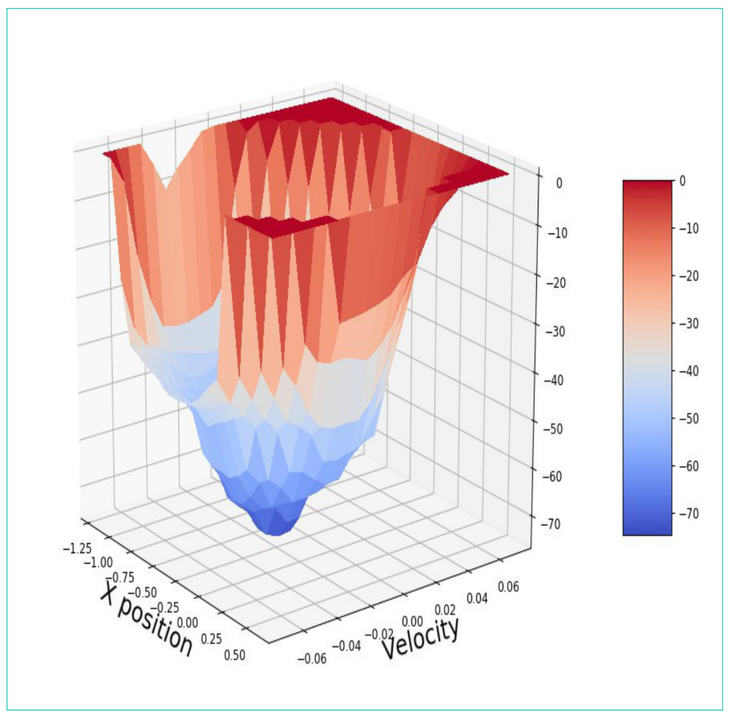
Модель робота в 3D-симуляторе



Обученный нами агент



Множество алгоритмов, потенциально решающих задачу



Обученная Q-таблица

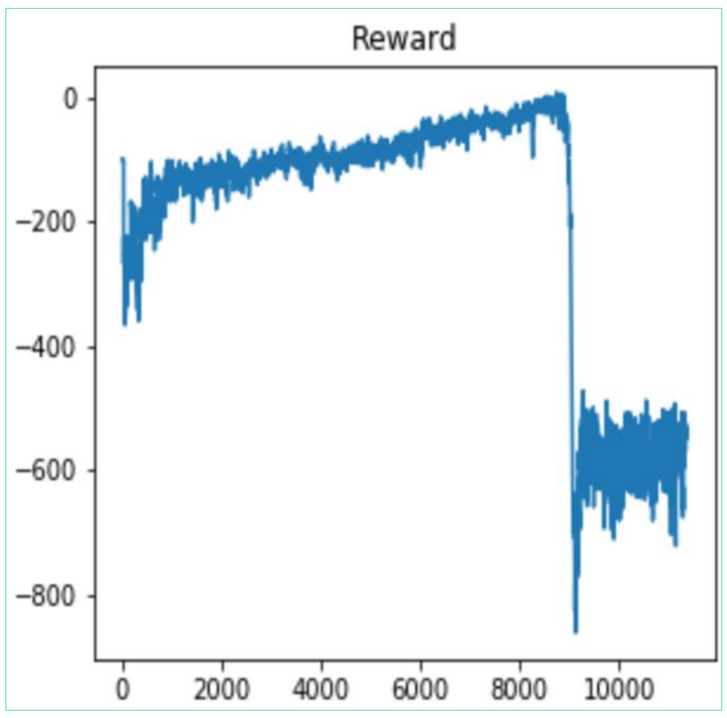


График неподходящего для нас алгоритма, не использующего Gradient Clipping

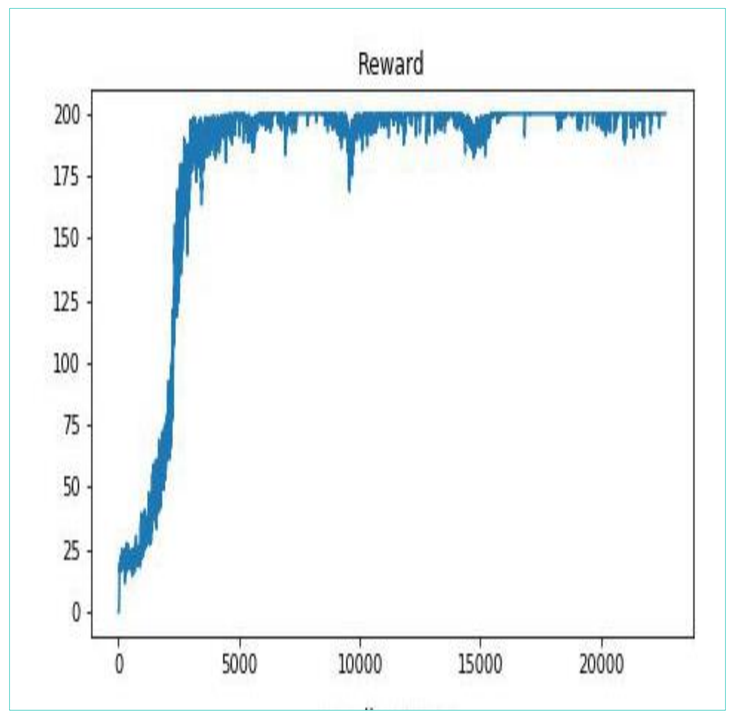


График алгоритма, на котором робот обучается ходьбе в настоящее время

Планы развития проекта

В настоящее время стремительно развивается домашняя робототехника. В наших домах становится всё больше и больше роботов, которые выполняют некоторую работу за нас. Примерами являются роботы-пылесосы, роботы-газонокосилки и даже роботы, умеющие складывать одежду. Наша работа даст толчок в развитии в данной области. Это означает, что роботы, действующие на основе алгоритма Reinforcement Learning, смогут выполнять более широкий класс задач

На данный момент робот может двигаться только по заданному вектору. Это означает, что столкнувшись с препятствием, он остановится. Поэтому, в дальнейшем мы хотим установить камеру на робота, которая бы передавала сверточной нейронной сети изображения окружающей обстановки. На их основе нейросеть сможет распознавать препятствия, что позволит нам научить робота успешно их преодолевать.

Также в естественной физической среде необходимо учитывать ситуации, когда робот может получить какие-либо повреждения, из-за которых он не сможет продолжать движение. Нашей задачей является разработка такого алгоритма, чтобы робот мог передвигаться даже с существенными неисправностями. Это означает, что он станет более приспособленным к условиям среды.

Голосование