Использование Python для обучения с подкреплением

Даниил Трубин

Газпром-Нефть Региональные Продажи

PiterPy 2022

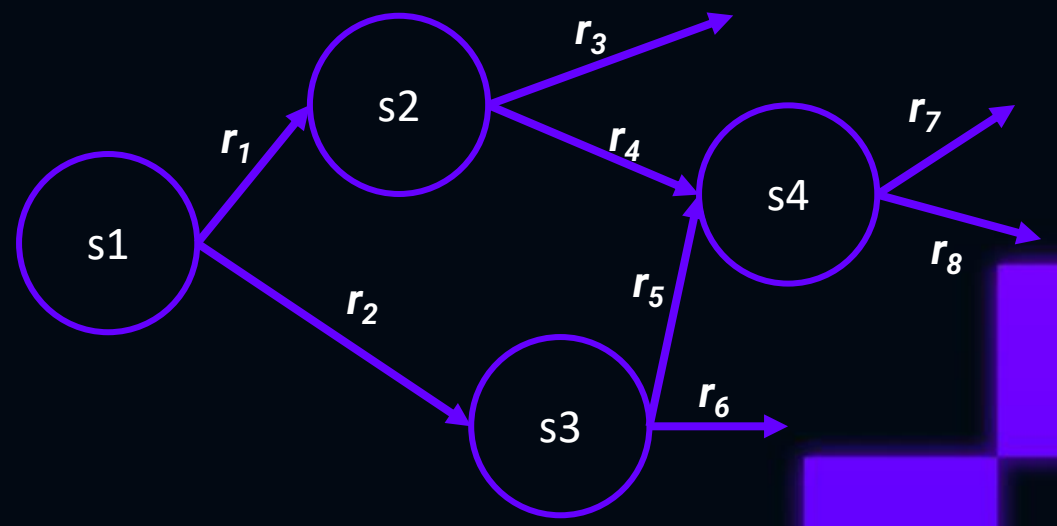
Содержание

- Что такое RL
- Как выглядит RL-проект
- Лайв-кодинг!
- Что дальше?

PiterPy 2022

Уровень 1

Уровень 2

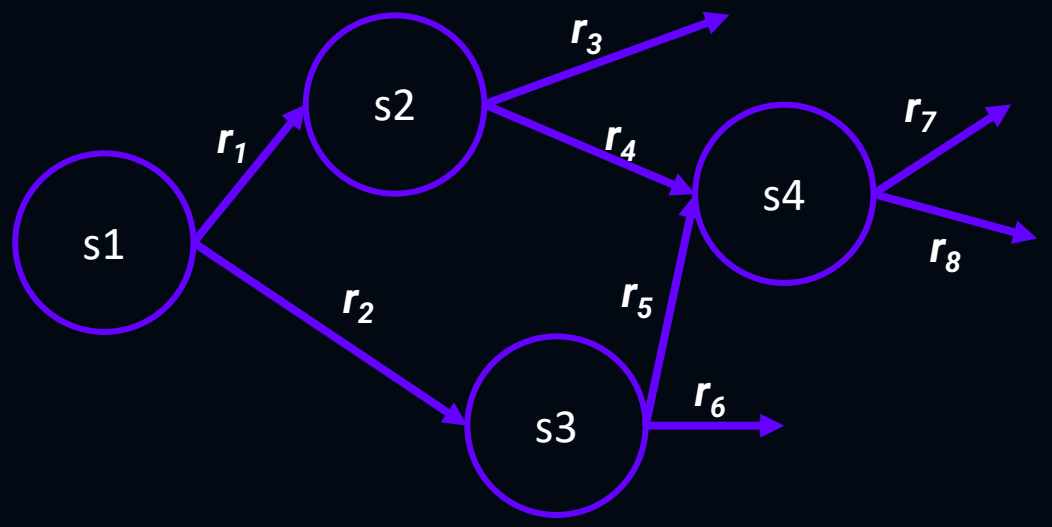


Уровень 3

$$V(s) = r + V(s`)$$

PiterPy 2022

Марковские процесс принятия решений (MDP)



Свойства:

- Текущее состояние не зависит от предыдущих, т.е зависит только от текущего
- Состояния дискретны
- Решение находится итеративно
- Может быть бесконечным (граф цикличным)

Уравнение оптимальности

$$V(s) = \sum_{i=t}^{end} r_i = r_t + \sum_{i=t+1}^{end} r_i = r_t + V(s`)$$

$$\Big/ \sum_{i=t}^{\infty} \gamma^{i-t} r_i$$

PiterPy 2022

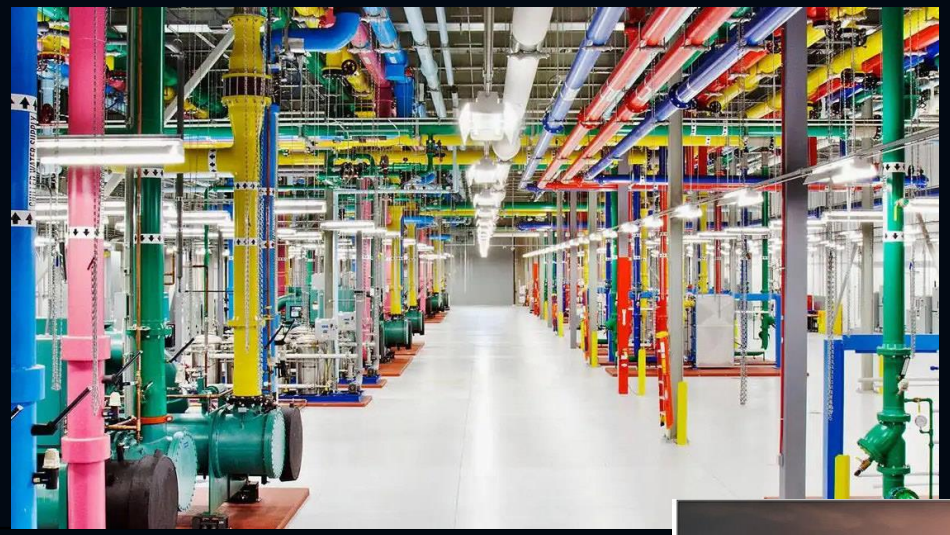
Как выглядит RL проект

Он должен быть:

- Последовательный (стратегический характер)
- Не завесить от прошлых состояний
- Обучение методом «проб и ошибок»

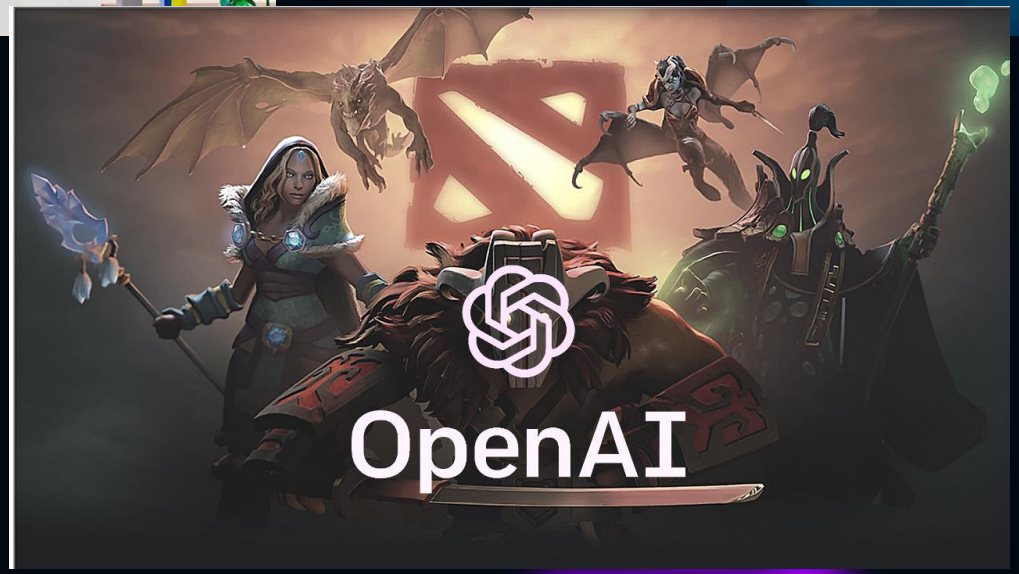
PiterPy 2022

Самые ~~модные~~ успешные кейсы RL



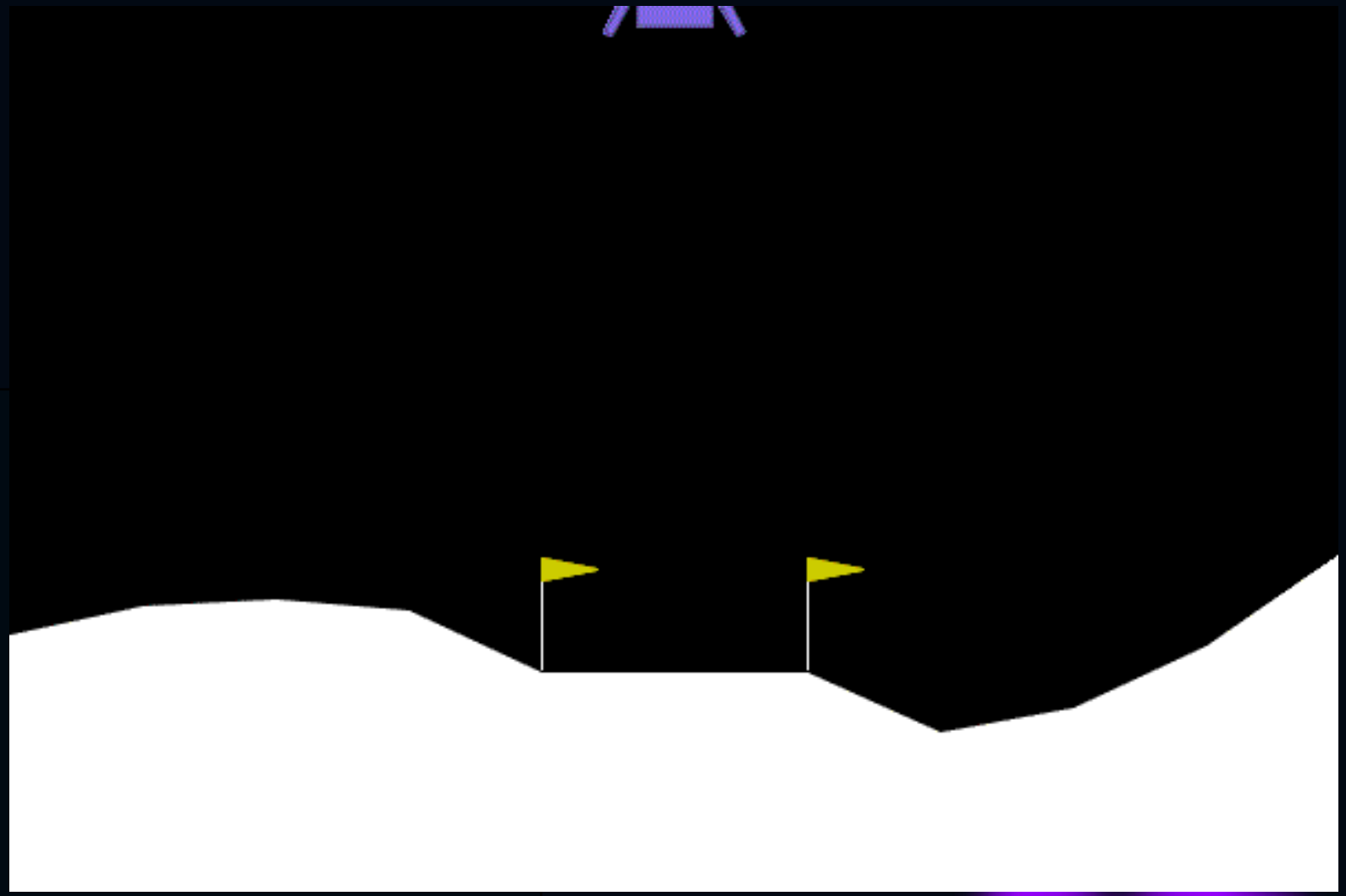
DeepMind. Охлаждение дата центров Google

DeepMind. AlphaGo Zero



OpenAI Five. Победа команды людей в Дота 2

Знакомство с проблемой: Lunar Lander

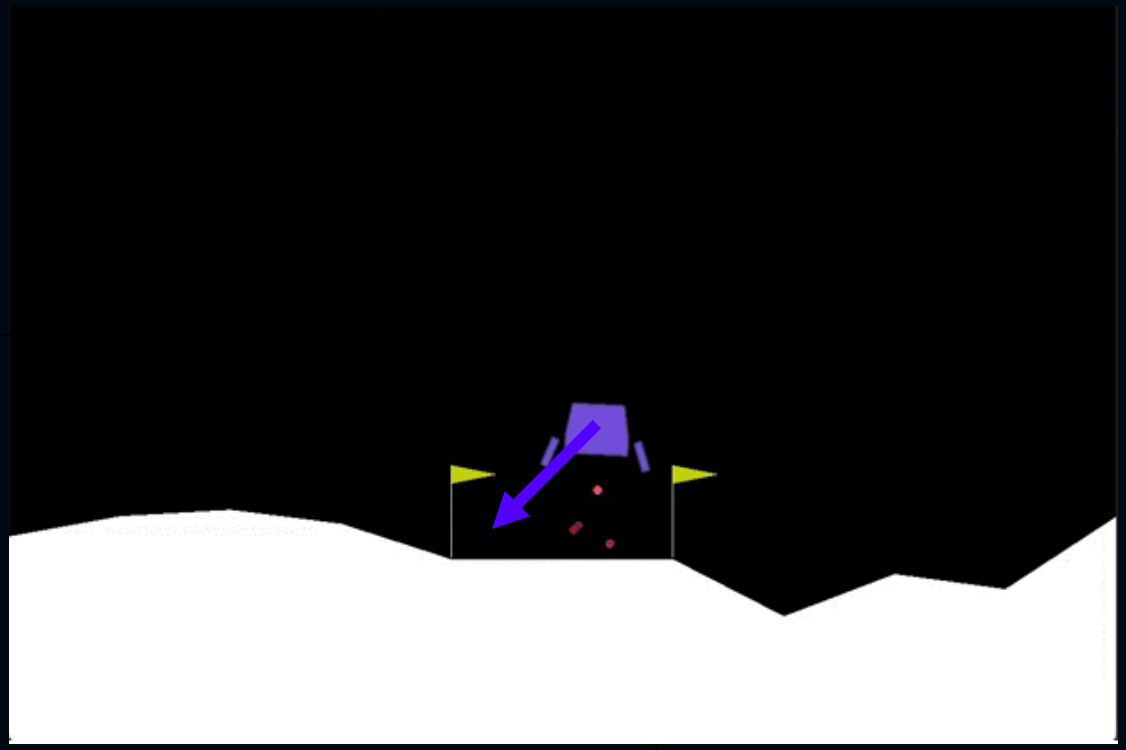


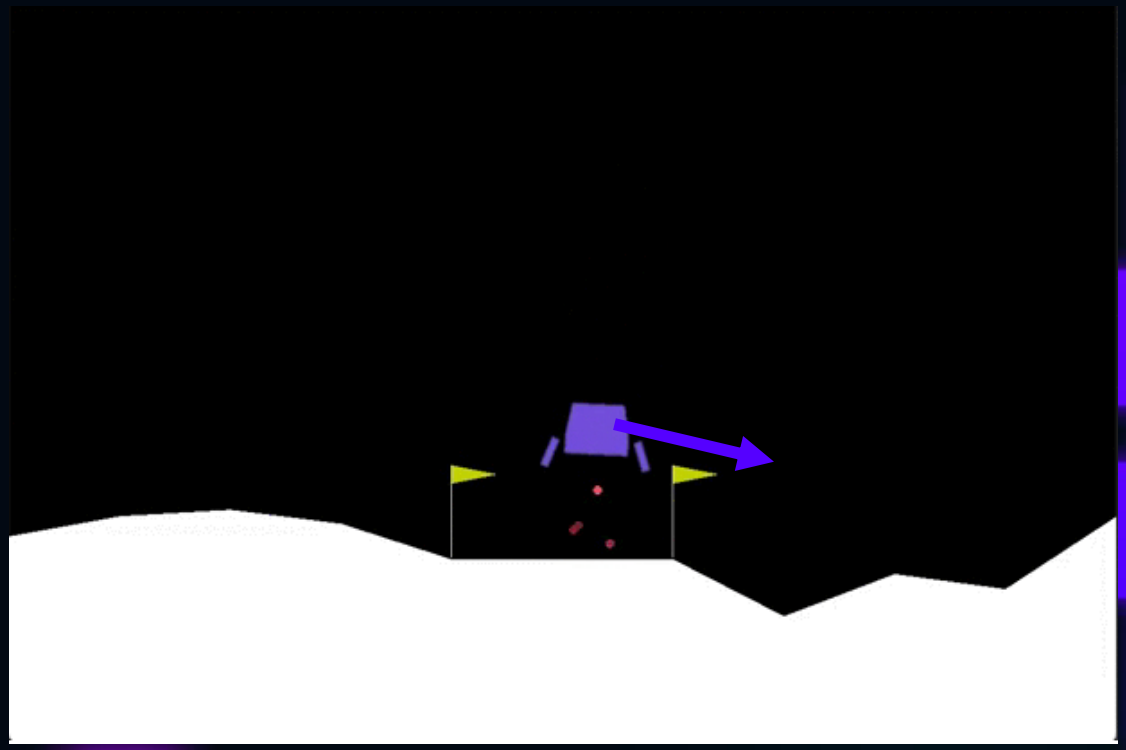
Цель: Посадить дрон на поверхность Луны между флажками

Действия:
0 – Ничего не делать
1 – Включить левый двигатель
2 – Включить основной двигатель
3 – Включить правый двигатель

PiterPy 2022

Знакомство с проблемой: Lunar Lander



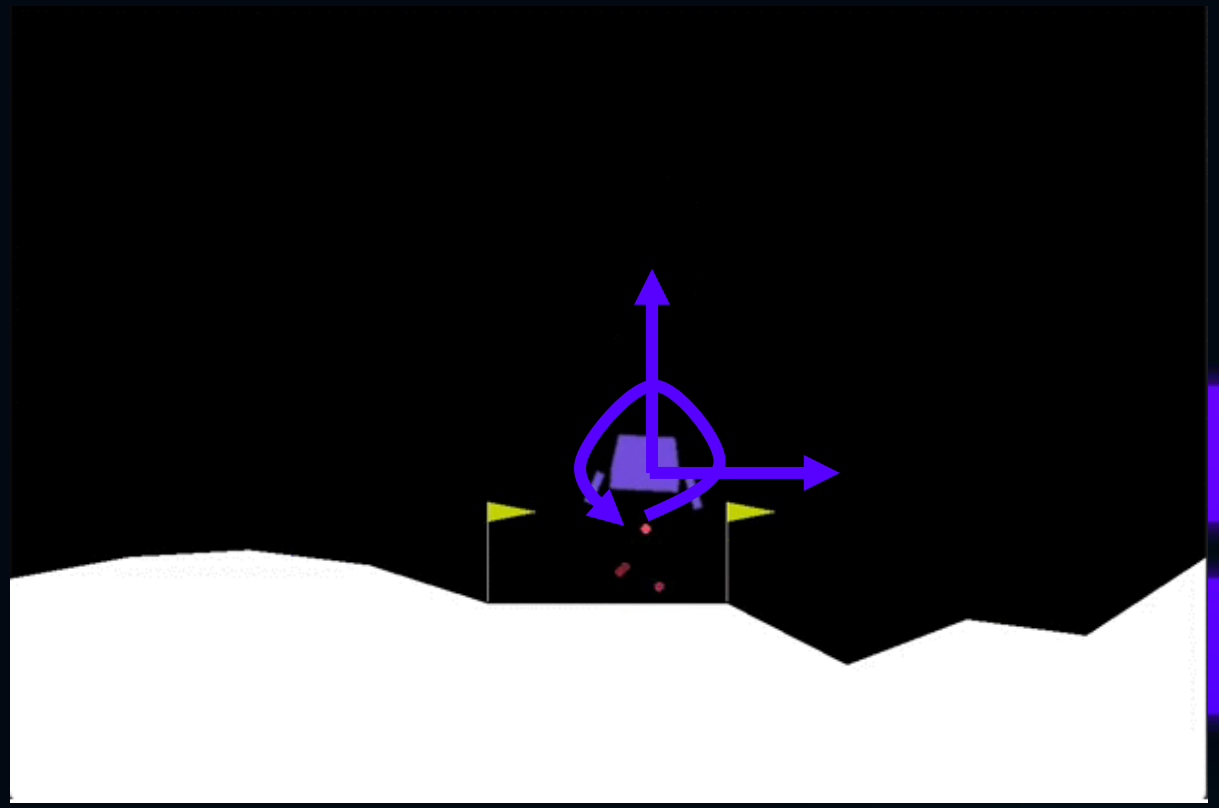


PiterPy
2022

Знакомство с проблемой: Lunar Lander

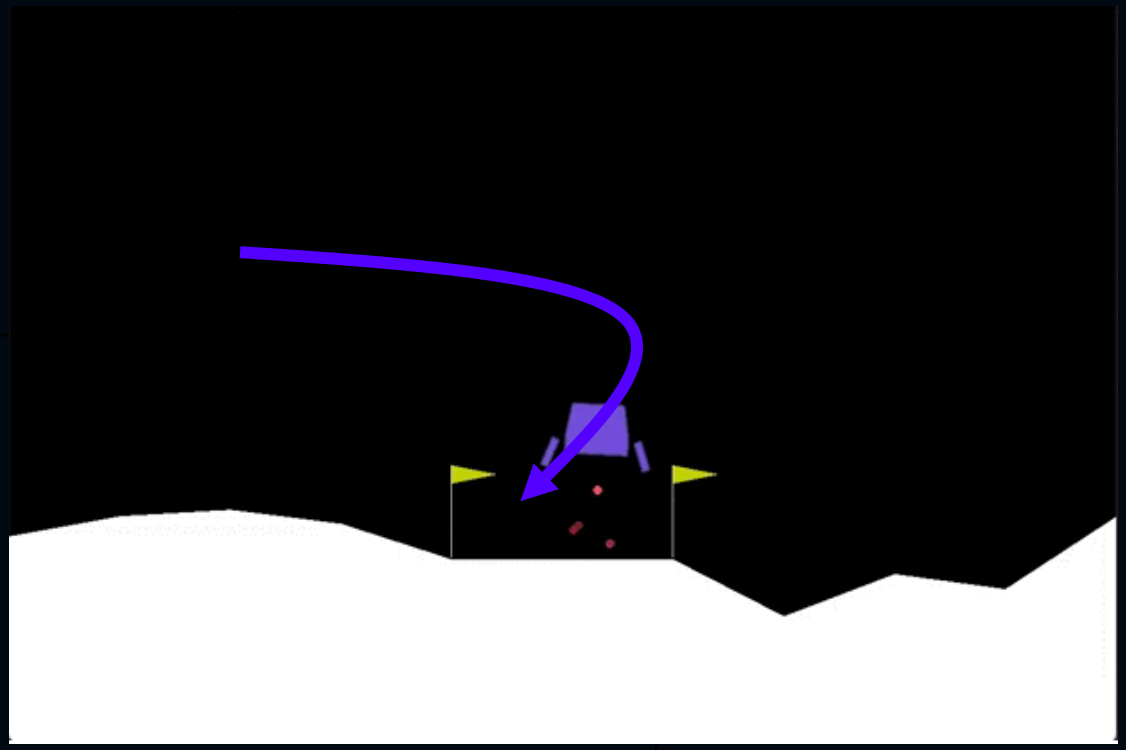
Состояние описывается вектором :

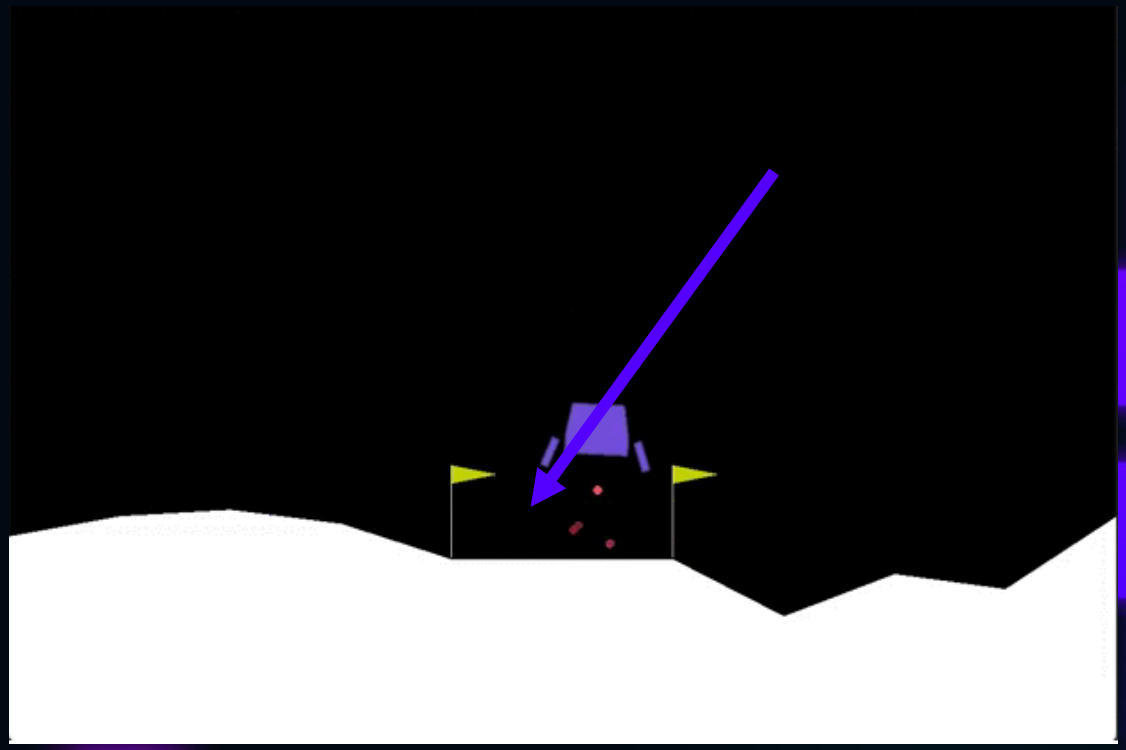
1. x – Координата
2. y – Координата
3. $\dot{x}$ – Скорость по x
4. $\dot{y}$ – Скорость по y
5. α – Угол наклона
6. $\dot{\alpha}$ – Скорость вращения
7. Касается ли правая нога земли
8. Касается ли левая нога земли



PiterPy 2022

Знакомство с проблемой: Lunar Lander

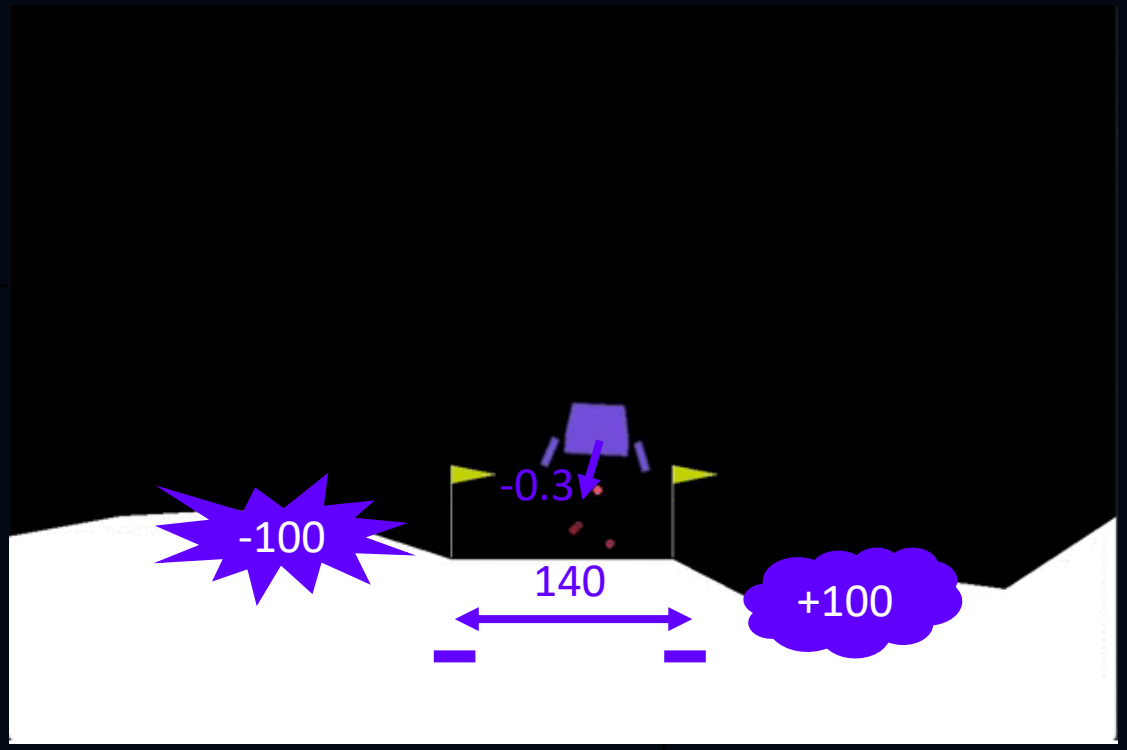




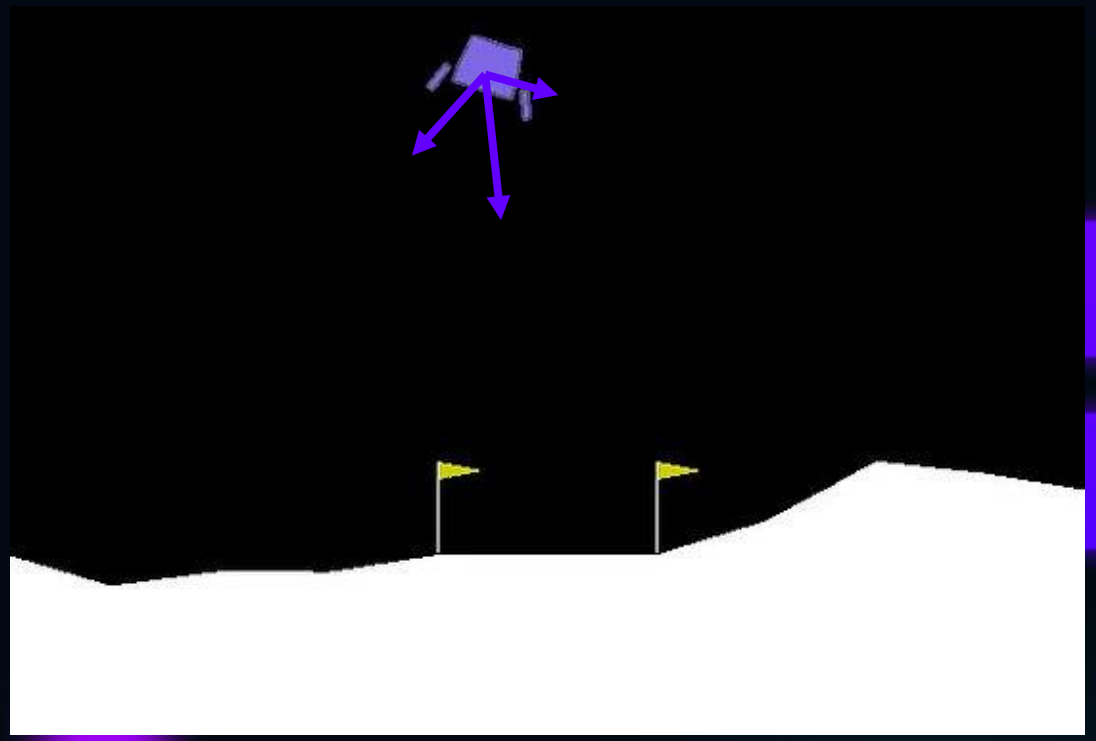
PiterPy 2022

Знакомство с проблемой: Lunar Lander

Награды



Начальные состояния



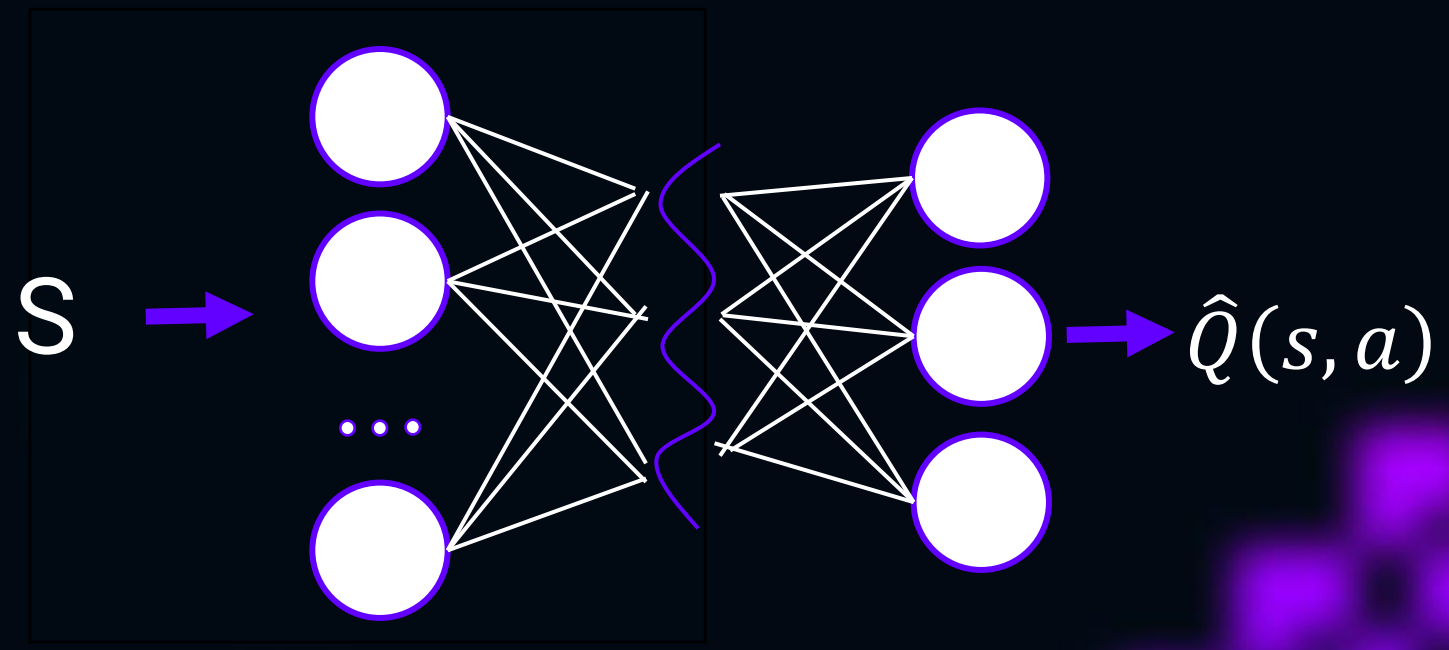
PiterPy 2022

Прикинем решение: Deep Q-Network

$$Q(s,a) = r_a + V(s`)$$
$$V(s) = \max_a Q(s,a)$$

$$\Rightarrow \quad Q(s,a) = r_a + \max_{a`} Q(s`,a`)$$

$$Q(s,a) := Q(s,a) + \alpha * (r_a + \max_{a`} Q(s`,a`) - Q(s,a))$$



Алгоритм:

1. Рассчитываем $\hat{Q}(s,a)$
2. ***<u>Эпсилон-жадно</u>*** выбираем действие a
3. Получаем новое состояние и награду
4. Рассчитываем $\hat{Q}(s`,a)$
5. Обновляем DQN

PiterPy 2022

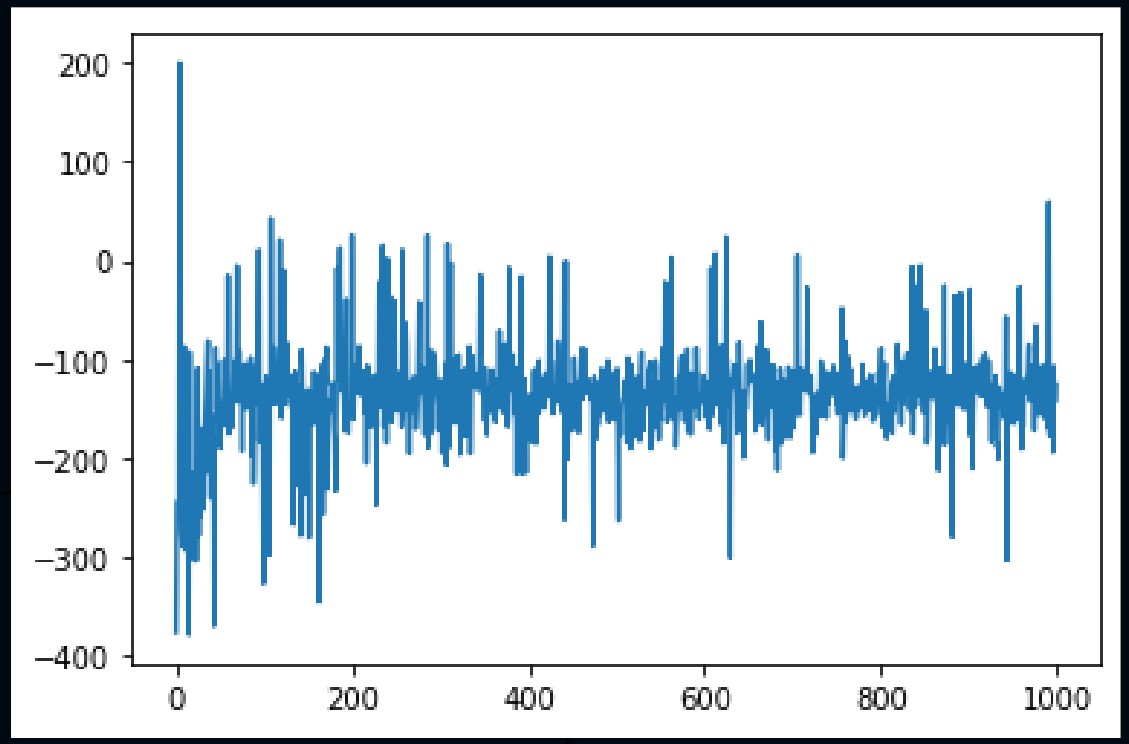
Лайв-кодинг!

<- Ссылка на гугл-коллаб

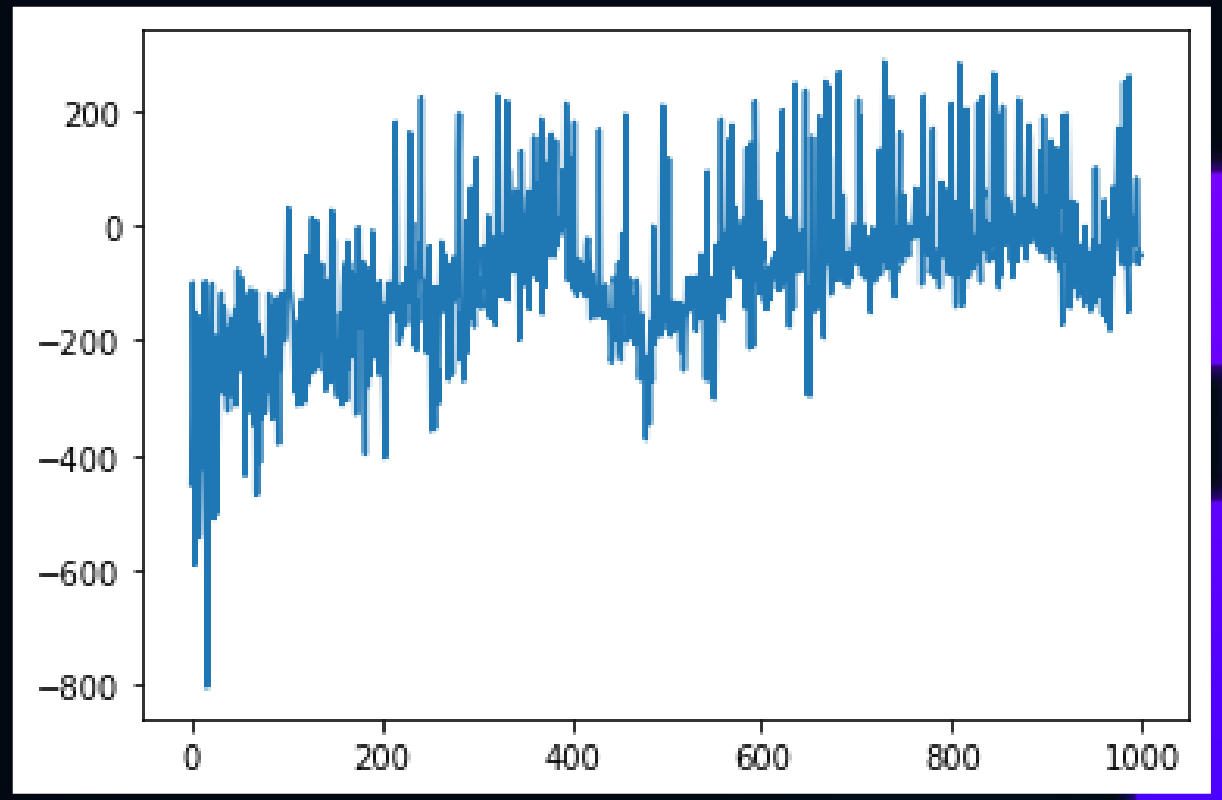
Но лучше использовать jupyter-notebook

PiterPy 2022

Что там с обучением?



Без Эпсилон-жадности



Эпсилон-жадная стратегия

PiterPy
2022

Что можно улучшить?

- Борьба с автокорреляцией – сделать хаб памяти и обучать на рандомных бачах из памяти
- Рандомное исследование мира и сбор данных в начале – первичное исследование мира, сокращение эпсилон относительно эпизодов
- Модификация наград

PiterPy 2022

Рекомендуемая литература

- Обучение с подкреплением для реальных задач. Инженерный подход. *Фил Уиндер*
- Кто такой многорукий бандит?
- Интуитивный RL (Reinforcement Learning): введение в Advantage-Actor-Critic (A2C)

PiterPy 2022

Спасибо за внимание!

PiterPy
2022