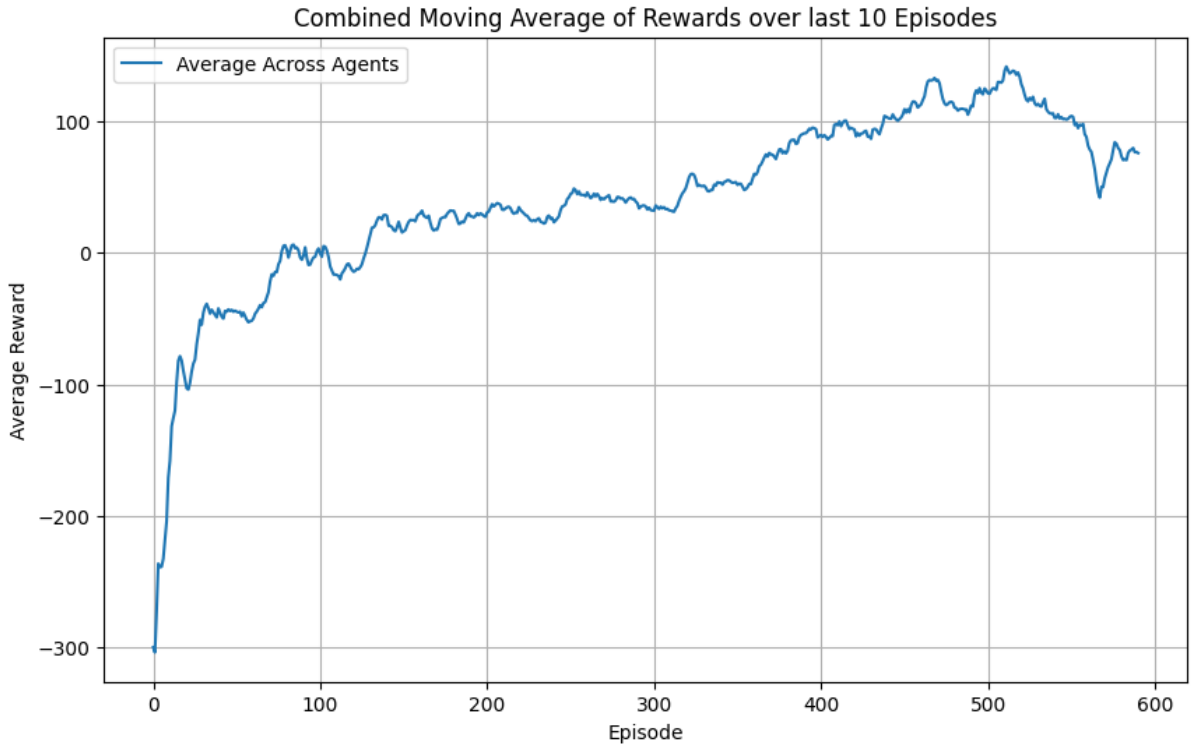
Домашняя работа #7 RL

Введение

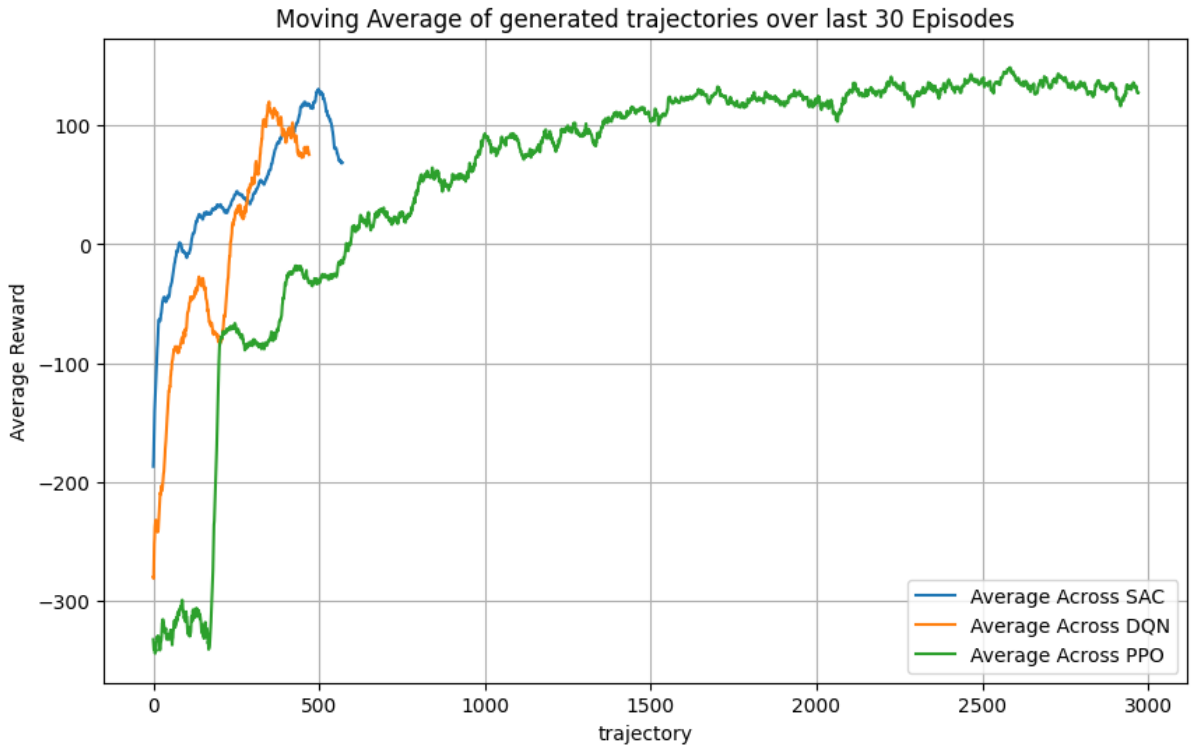
Первое задание

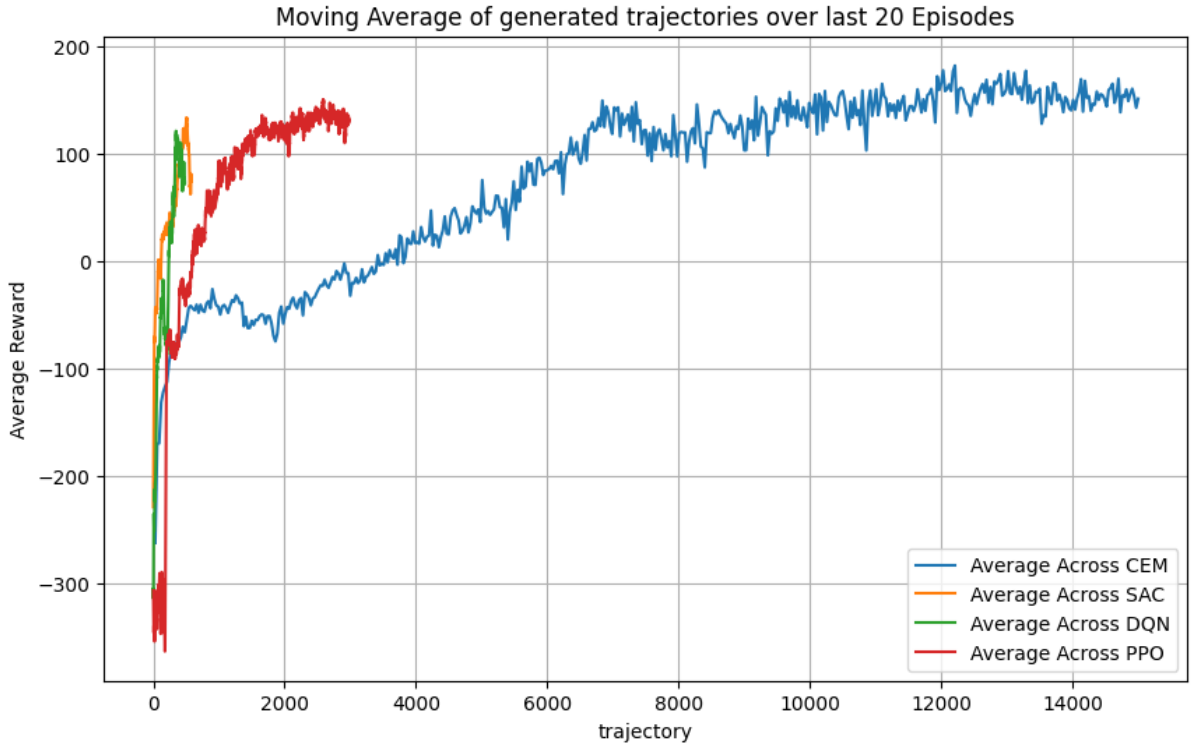
Обучение

Для этого задания я решил выбрать Lunar-Lander-v2 continious. Я сделал алгоритм многомерным, и протестировал разные гиперпараметры. В конечном итоге выбрал следующие гиперпараметры:
hidden_dim=128, gamma=0.99, initial_alpha=0.3, tau=1e-2, batch_size=64, pi_lr=1e-3, q_lr=1e-3
Так же для того, чтобы сделать алгоритм более стабильным было принято решение уменьшать альфа со временем. Так он стабильнее обучался, и увеличивал exploitation в конце.



В целом алгоритм за 600 эпизодов достигал достаточно неплохой точности, но все равно это было мало, и нестабильно относительно PPO и CEM. В интернете посоветовали написать replay buffer, но я решил не делать этого для того, чтобы было сравниваемые алгоримы были наиболее близки к алгоритмам с практик.





Наверху графики средней награды относительно сгенерированных траекторий. Каждый агент запускался 5 раз и бралось среднее относительно сгенерированных траекторий. Я старался как можно менее изменять алгоритмы относительно изначальных вариаций, чтобы сравнение было правдивое настолько насколько это возможно. Помимо этого везде старался сохранять одинаковые lr и batch_size, и нейросеть настолько насколько это было возможно. Это не всегда получалось. Некоторые алгоритмы пришлось немного переписать относительно практик. В целом CEM и PPO получились самыми стабильными на данной задаче, а DQN и SAC

достаточно быстро переобучались, и им необходим был еще какой-то механизм от переобучения, и для стабильности. Так же при обучении некоторые алгоритмы застревали в локальных минимумах, и приходилось играться с exploration, чтобы они могли оттуда выходить.

Достаточно сложно судить о производительности алгоритмов ориентируясь на количество сгенерированных траекторий, да и вообще. По ощущениям каждый алгоритм надо подкручивать по разному для разных сред. Сравнение на разных средах особенно если к нему еще приложить время обучения мне кажется было бы показательнее.

Вывод

При сравнении такого количества алгоритмов появилось разные мысли относительно того, как вообще происходит обучение с подкреплением. Было интересно отмечать то как разные алгоритмы реагируют на параметры по разному. Грустно то что не удалось достичь очень хороших результатов с SAC. По идее это SOTA для lunar lander continious, но в коде с гх есть большее количество модификаций, которые я не хотел прикручивать.