МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ
по лабораторной работе №2
по дисциплине «Обучение с подкреплением»
Тема: Реализация PPO для среды MountainCarContinuous-v0

Студент гр. 0310 _____ Якушкин Д.А.

Преподаватель _____ Глазунов С.А.

Санкт-Петербург
2025 г.

Цель работы.

Реализовать алгоритм PPO для обучения агента в среде MountainCarContinuous-v0.

Задание.

1. Изменить длину траектории (steps).
2. Подобрать оптимальный коэффициент clip_ratio.
3. Добавить нормализацию преимуществ.
4. Сравните обучение при разных количествах эпох.

Выполнение работы.

1. **Реализация PPO**

Для работы этого алгоритма требуются классы Actor и Critic. Actor - класс, отвечающий за действия агента, Critic - класс, определяющий функцию ценности, которая отвечает за награду для каждого состояния.

Базовые параметры, используемые для тестирования следующие:

numIterations = 300
numSteps = 2048
ppoEpochs = 10
miniBatchSize = 64
gamma = 0.99
gaeLambda = 0.95
clipRatio = 0.2
valueCoef = 0.5
entropyCoef = 0.01
lr = 3e-4

Также в функцию рассчета возвратов и преимуществ была включена нормализация. Это необходимо для того, чтобы избежать слишком маленьких или слишком больших шагов обновления.

2. **Влияние длины траектории на обучение**

Для экспериментов были выбраны следующие значения: 1024, 2048,

4096.

Были получены сводные графики avg_reward для каждого эксперимента. График можно увидеть на рисунке
1

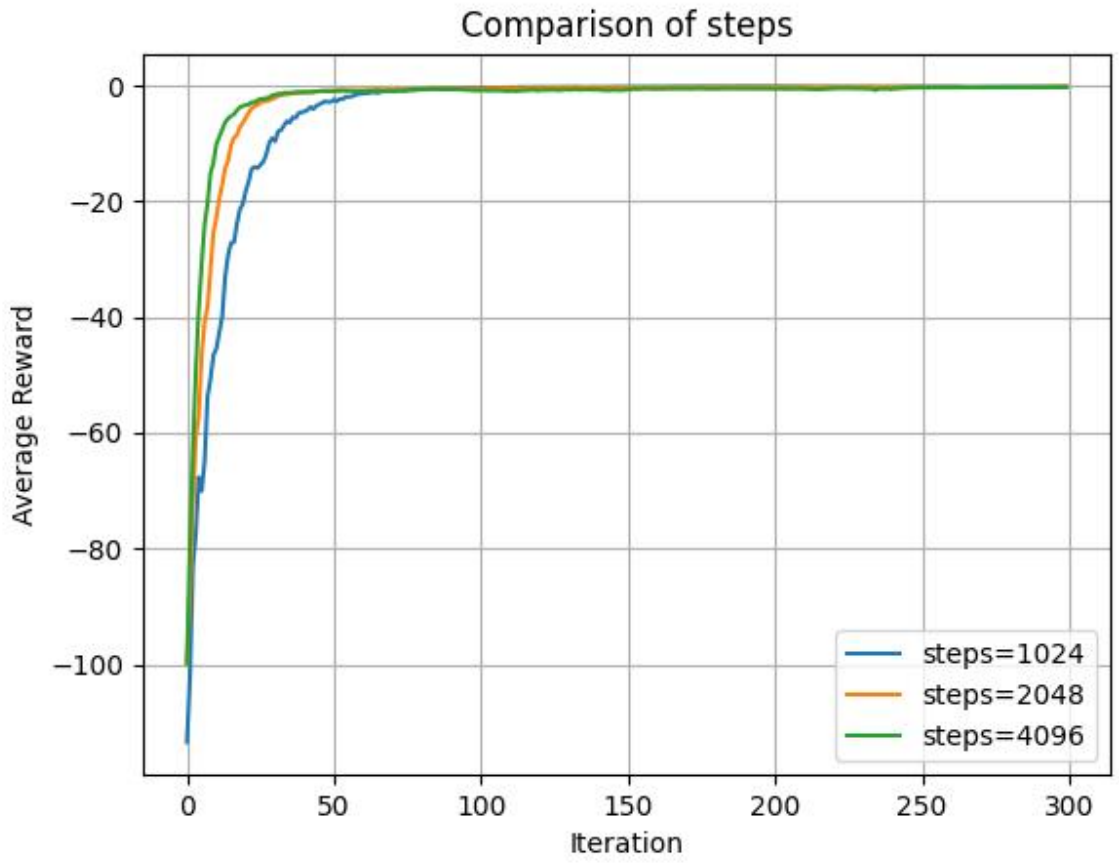

Рисунок 1 - Сравнение avg_revard для различных значений траектории

Исходя из графиков можно судить что быстрее всего модель обучается на количестве шагов 4096.

3. Влияния clip_ratio на обучение

Clip_ratio - это коэффициент обрезки. Он указывает то, насколько новая политика может отличаться от предыдущей. Он нужен для предотвращение резкого изменения политики, которое может повлиять на стабильность обучения.

Для эксперимента были выбраны значения 0.1, 0.2 и 0.3. Результаты можно увидеть на рисунке 2.

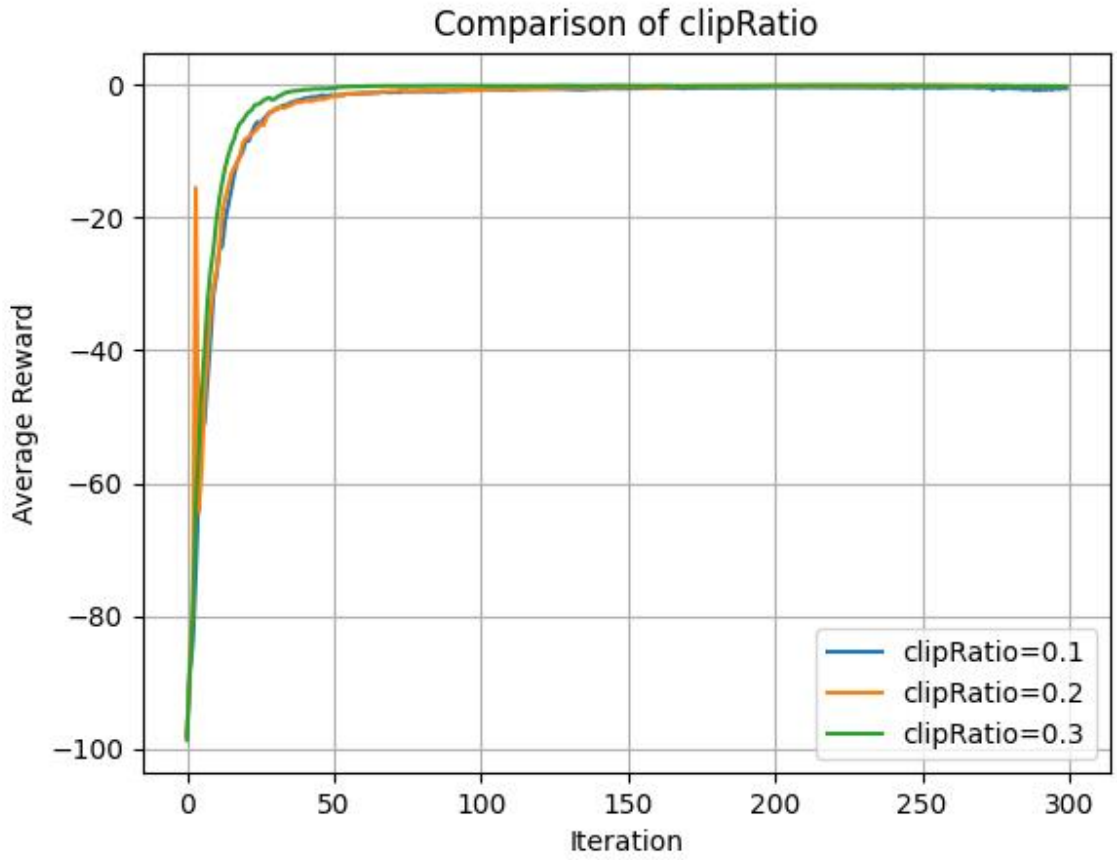

Рисунок 2 - Сравнение avg_revard для различных параметров clip_ratio

Исходя из графиков можно сделать вывод что clip_ratio=0.3 даёт более быстрое достижение результата.

4. Влияние количества эпох на обучение

Эпоха - это проход по всем данным. Чем больше эпох тем медленнее модель будет учиться, но тем более точно обновляются политики, однако на маленьких наборах данных это может привести к переобучению. Для эксперимента были выбраны количества эпох 3, 5 и 10. Результаты можно увидеть на рисунке 3.

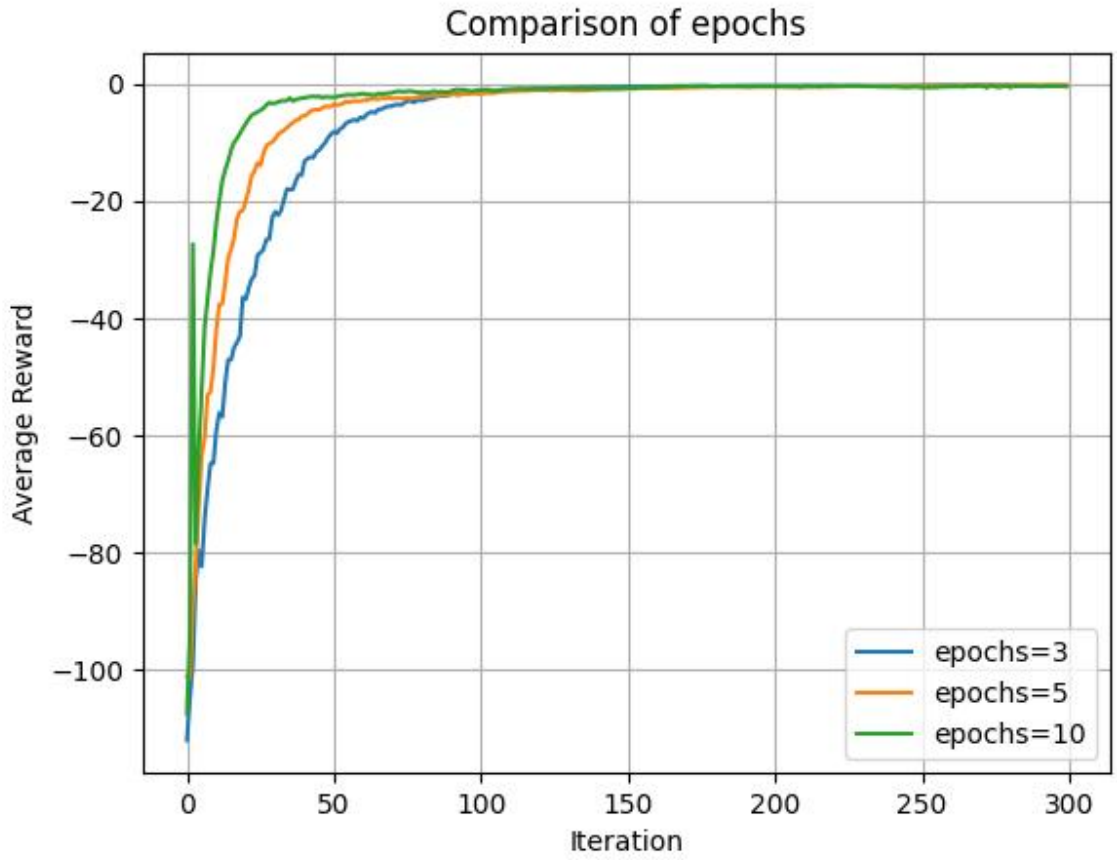

Рисунок 3 - Сравнение avg_reward для различных параметров количества эпох

Исходя из графиков можно сказать что оптимальным количеством эпох для нашего случая являются 10.

Выводы.

В ходе лабораторной работы был реализован алгоритм PPO для обучения агента в среде MountainCarContinuous-v0. В алгоритм была добавлена нормализация а также были проведены эксперименты влияние накоторых параметров на работу нейронной сети.