**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра МОЭВМ**

**ОТЧЕТ**

**по лабораторной работе №1**

**по дисциплине «Обучение с подкреплением»**

**Тема: Реализация DQN для среды CartPole-v1**

|  |  |  |
| --- | --- | --- |
| Студент гр. 0306 |  | Кирсанов Д.Э. |
| Преподаватель |  | Глазунов С.А. |

**Цель работы**.

Реализация PPO для среды MountainCarContinuous-v0. Исследование зависимости результатов от макропараметров: steps (далее numSteps), clip\_ratio (далее clipRatio), ppoEpochs, а также от добавления нормализации преимуществ.

**Задание.**

1. Реализация PPO
2. Добавление нормализации
3. Влияние на обучение изменение макропараметров, указанных в цели работы.

**Выполнение работы.**

1. **Реализация PPO**

Основными созданными сущностями для алгоритма PPO являются Actor и Critic, где в этой паре Actor рассчитывает потенциальное действие, а Critic оценивает его выгоду (advantage) на основании полученной награды (на самом деле на отрезке наград), тем самым корректируя дальнейшие действия Actor. Входными параметрами для Actora, так как он выбирает действие, являются Observation Space среды MountainCarContinuous-v0, описание которых представлено на рисунке 1, а на выходе численный показатель движения влево или вправо в зависимости от знака числа в диапазоне от [-1 до 1]: движение влево (индекс 0) или движения в право (индекс 1).

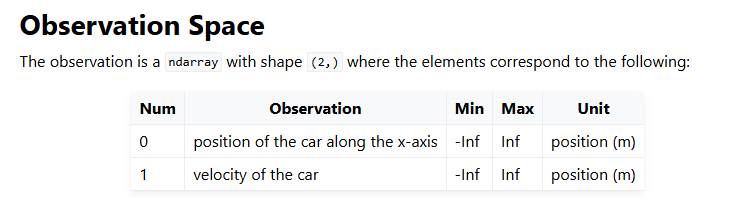


Рисунок 1 – Пространство состояние в среде MountainCarContinuous-v0

Поведение PPO определено следующими гиперпараметрами, представленными на рисунке 2.

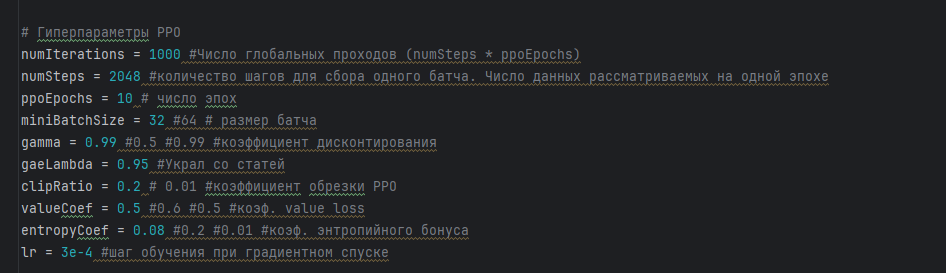


Рисунок 2 – Изначальное состояние гиперпараметров сети

Все параметры, изменяемые на дальнейших этапах работы представлены на рисунке выше. Остальной код, а именно тренировочный цикл, метод накопитель траекторий данных – collect\_trajectories, метод получения returns и advantages – cra\_func, представлен в исходном скрипте PPO\_Routine\_MoutainCar.py.

1. **Добавление нормализации**

Без нормализации награды PPO совершает слишком жадные действия, из-за разброса получившихся значений разброса, как итог награда всегда наименьшая, а машинка всегда движется на левый холм, что представлено в видео «/videos/Результат без нормализации.mp4». Значение ошибки и награды при итерациях представлены на рисунках 3.

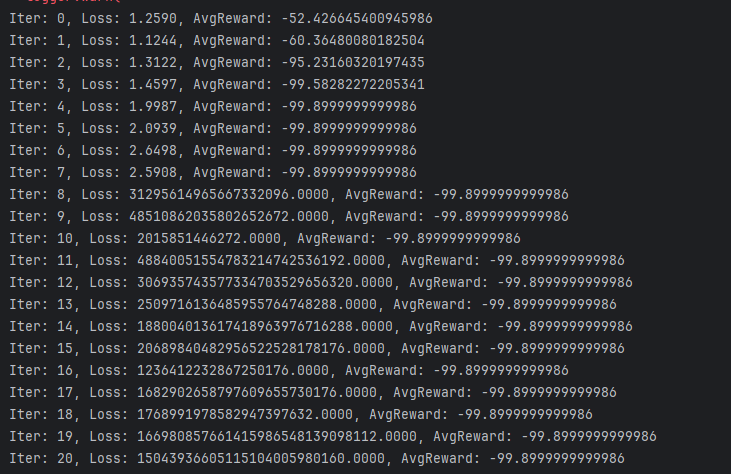


Рисунок 3 – Результат без нормализации

С добавлением нормализации средней абсолютной ошибкой (MAE), что на реализовано на строчке 122 внутри метода cra\_func, PPO достиг целевого состояния, что представлено на рисунке 4, где результат на итерациях до достижения целевого состояния, а само движение записано и представлено в ролике «Результат с нормализацией.mp4».

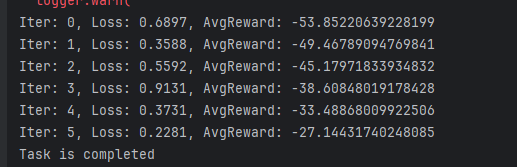


Рисунок 4 – Результат с нормализацией

1. **Влияние гиперпараметров**

Из целевых параметров для изменения весомый эффект возымеет только изменение clipRatio, так steps и обучение без одновременного изменения размера батча не имеет смысла, потому что в результате сеть в среднем будет достигать результата при одинаковом количестве шагов, обучаясь на одном размере данных (размер батча), не считая выбросов с неудачным значением весов сети при их инициализации.

Оптимальным значением clipRatio оказалось базовое значение 0.2.

При большем в два раза значении 0.4 хоть и награда растёт быстрее, но в реальности машинка раньше начинает совершать более осторожные действия, которые и поощряются средой, не позволяя ей пересечь ключевую точку и быстрее сводит её в состояние циклирования около дна котлована, без единого пересечения ключевой точки до этого состояния.

При уменьшении clipRatio сеть более верна своим изначальным параметрам, что делает её наиболее зависимой от изначального значения параметров и менее предсказуемой на практике. Например, в случае базового значения в 0.2 получилось найти значение награды, при пересечении которого модель наиболее вероятно пересекает ключевую точку, когда в случае 0.08 машинка на одном из экспериментов пересекла точку уже на 3 видеозаписи, без явного изменения в награде.

**Выводы.**

Была выполнена реализация PPO для среды MountainCarContinuous-v0. Было проведено исследование влияния изменения некоторых параметров сети на её результат, а также эффективность применения нормализации к значениям advantages.