**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра МОЭВМ**

**ОТЧЕТ**

**по лабораторной работе №3**

**по дисциплине «Обучение с подкреплением»**

**Тема: Реализация SAC для среды Flappy Bird**

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

**Цель работы**.

Реализация SAC для среды нестандартной среды Flappy Bird. Исследование влияния параметра alpha для контроля энтропии. Реализация автоматической настройки alpha.

**Задание.**

1. Реализация SAC.
2. Исследование влияния alpha.
3. Реализация автоматической настройки alpha.

**Выполнение работы.**

1. **Реализация SAC**

Основными созданными сущностями для алгоритма SAC являются Actor и Critic, где в этой паре Actor рассчитывает потенциальное действие, исходя из текущей политики, а Critic оценивает рассчитанное действие. На основании оценки Critic Actor корректирует свою политику, увеличивая вероятность хороших действий (с учётом энтропии). SAC обучает два Critic аналогично тому, как обучается QNetwork в алгоритме DQN, для уменьшения переоценки. Входными параметрами для Actora, так как он выбирает действие, являются Observation Space среды FlappyBird-v0 [1], описание которых представлено на рисунке 1, а на выходе численный показатель движения или бездействия, которые выражены 1 и 0 соответственно.

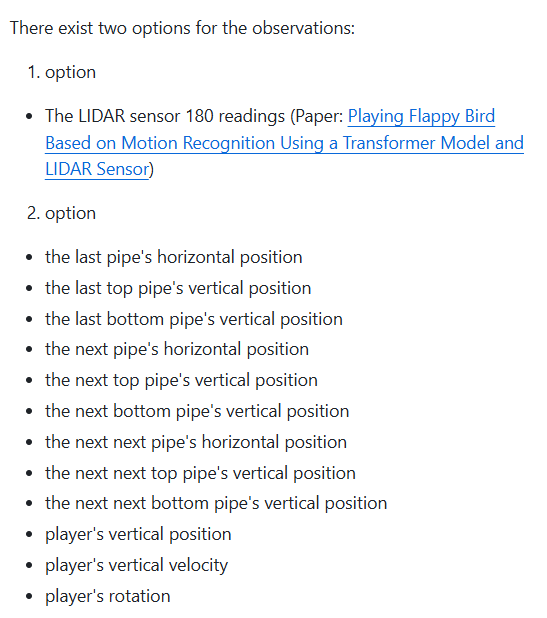


Рисунок 1 – Пространство состояние в среде MountainCarContinuous-v0

Поведение SAC определено следующими гиперпараметрами, представленными на рисунке 2.

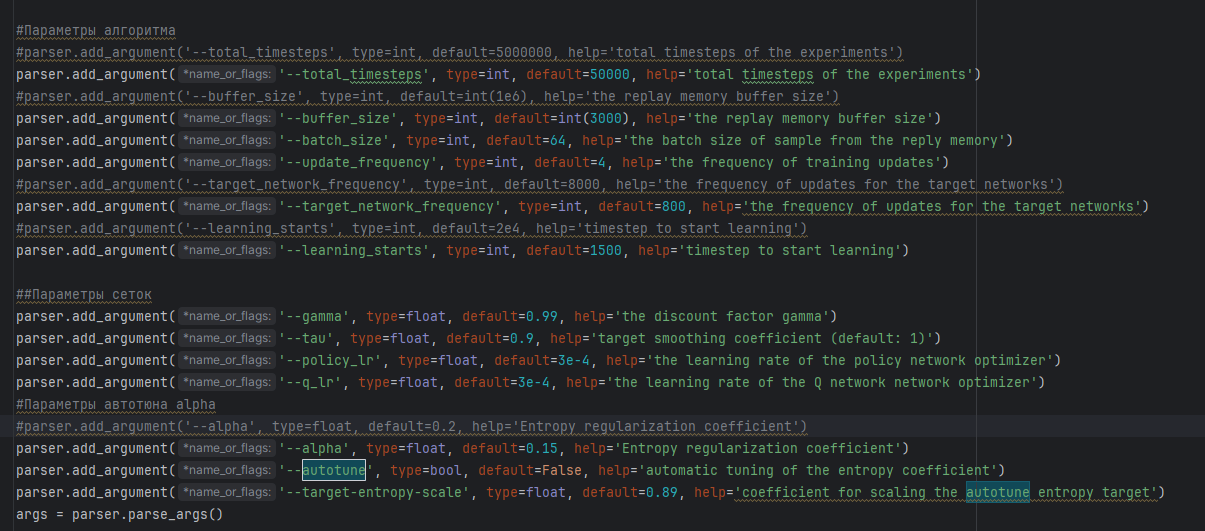


Рисунок 2 – Изначальное состояние гиперпараметров сети

Все параметры, изменяемые на дальнейших этапах работы представлены на рисунке выше. Остальной код, а именно тренировочный цикл, класс буфера ReplayBuffer, метод установки изначальных значениев параметров слоев layer\_init, автоматическое обновление alpha, представлены в исходном скрипте SACRoutine.

1. **Исследование alpha**

Важное замечание, что эксперименты приводились на пространстве состояний LIDAR. Наилучшее достигнутое число перезапусков равно 16.

При 0 значении alpha, что приводит агента к максимизации награды с минимизацией исследования, агент минимизирует число перезапусков только до 26 на интервале 800 шагов. Вывод SPS и число перезапусков при итерациях представлены на рисунках 3.



Рисунок 3 – Результат без автоматической настройки с alpha 0

При максимизации энтропии при параметре alpha равного 1 агент получает наилучшую достигнутую оптимальную стратегию за одну итерацию. Аналогичный результат достигается при изначальном значении 0.2 взятого с оригинальной реализации, выступающей референсом – sac\_atari [2].

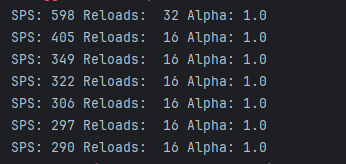


Рисунок 4 – Результат без автоматической настройки с alpha 1 и 0.2

**3. Автоматическая настройка alpha**

Реализация автоматической настройки alpha была реализована в соответствии sac\_atari [2]. Запуск при использовании автоматической настройки alpha представлена на рисунке 5.

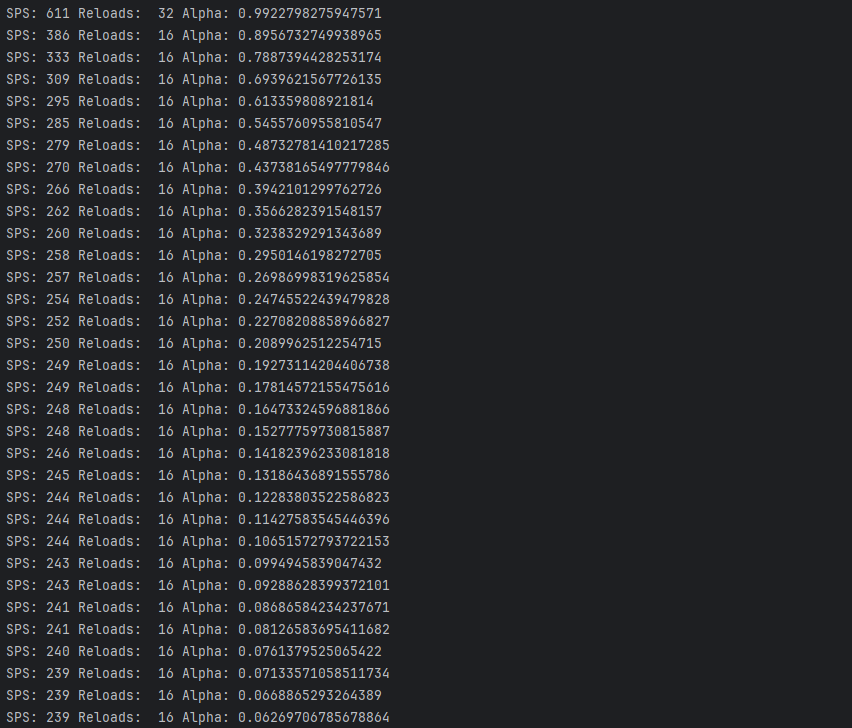


Рисунок 5 – Результат с автоматической настройкой

**Выводы.**

Была выполнена реализация SAC для нестандартной среды FlappyBird-v0 вместе с исследованием влияния параметра alpha и реализация возможности его автоматической настройки в прцоессе обучения.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Среда FlappyBird [Электронный ресурс]. URL: https://github.com/markub3327/flappy-bird-gymnasium (дата обращения: 01.05.2025).
2. Референс sac\_atari [Электронный ресурс]. URL: https://github.com/vwxyzjn/cleanrl/blob/master/cleanrl/sac\_atari.py (дата обращения: 01.05.2025).