Домашнее задание 2: Реализация базового алгоритма RL и запуск первых экспериментов

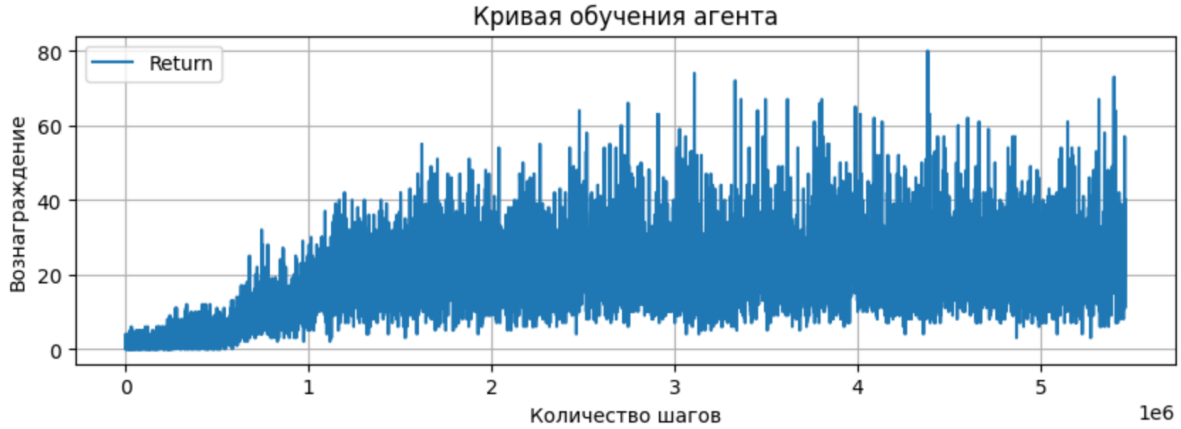
Зимин Евгений Евгеньевич

Задача: Мы выбрали задачу обучения агента игре **Breakout**. Эта игра является частью библиотеки Atari и подходит для методов обучения с подкреплением, поскольку обладает следующими особенностями:

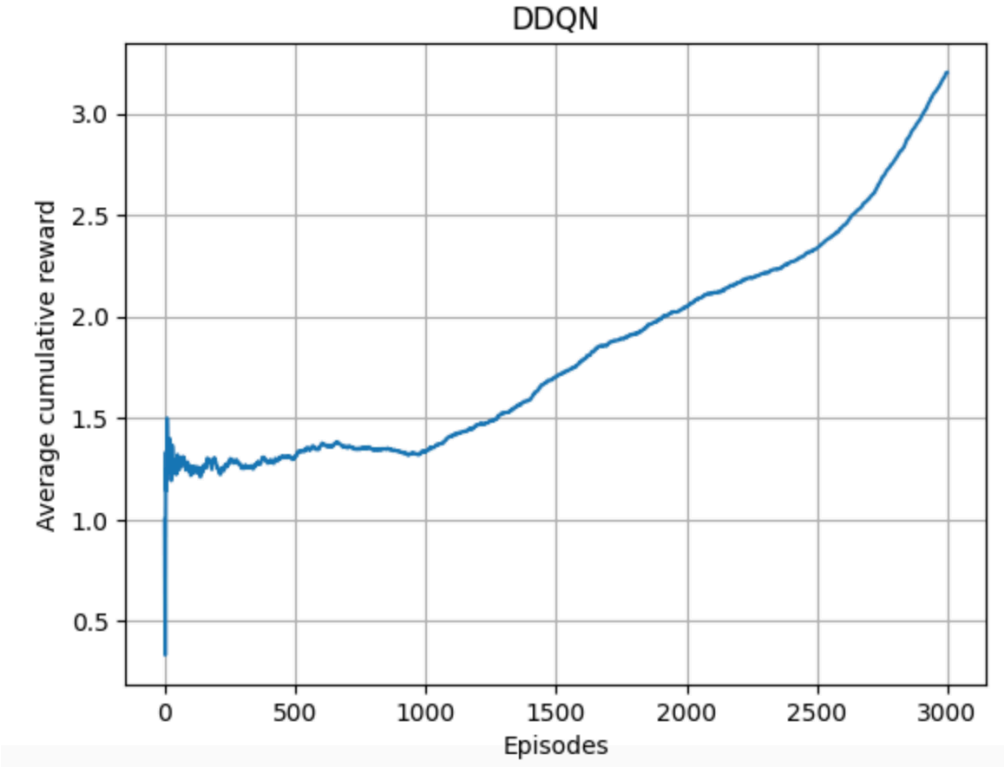
- Набор состояний и действий достаточно прост для начального уровня реализации.
- Игра имеет четкую структуру наград, что упрощает формализацию целевой функции.
- Прогресс агента легко оценивать через счет, а также визуально наблюдать процесс обучения.

Результаты экспериментов[1]

Кривая обучения



[1] Результаты экспериментов подробно представлены в .ipynb файле, в репозитории Github.



На графике, построенном в ходе экспериментов, видно, что **среднее накопленное вознаграждение** (Average Cumulative Reward) неуклонно увеличивается с увеличением числа эпизодов. Это указывает на успешное обучение агента:

- На начальных этапах обучения среднее вознаграждение было низким, так как агент находился в фазе исследования среды и выполнял случайные действия из-за высокого значения ϵ (эпсилон).
- По мере снижения ϵ (в рамках ϵ-жадной стратегии) агент стал принимать более оптимальные действия на основе опыта, что привело к росту среднего вознаграждения.

Анализ кривой

1. **Период исследования среды**: В первые 1000 эпизодов агент в основном выполнял случайные действия, стремясь собрать разнообразные данные о среде. На графике это отражается медленным ростом кривой.
2. **Фаза оптимизации**: Начиная с середины обучения, агент стал активно использовать накопленный опыт. Это выразилось в более крутом наклоне графика.
3. **Стабилизация**: В текущей версии проекта модель демонстрирует уверенный рост среднего вознаграждения, что свидетельствует о постепенном улучшении её стратегии. Однако до стабилизации и выхода на плато, которые указывают на достижение эффективной и устойчивой политики, обучение ещё не завершено. В дальнейшем, на финальной стадии проекта, мы планируем продолжить обучение модели, чтобы достичь стабилизации результатов и более высокого уровня игровой эффективности.

Сравнение с базовой линией

Для оценки эффективности была выбрана **случайная стратегия** (Random Agent) в качестве базовой линии.

Агент DDQN значительно превзошёл случайного агента.

Метрики производительности

- **Среднее вознаграждение за последние 100 эпизодов**: для DDQN составляет 23.37, в то время как для случайной стратегии 1.27.
- **Количество победных эпизодов**: продемонстрировало устойчивый рост по сравнению с началом обучения.
- **Скорость обучения**: использование Double DQN и Prioritized Replay значительно сократило колебания в результатах, улучшив общую эффективность обучения.