Lab3 Report

Proximal Policy Optimization

Student ID : 313554044

Student Name : 黃梓誠

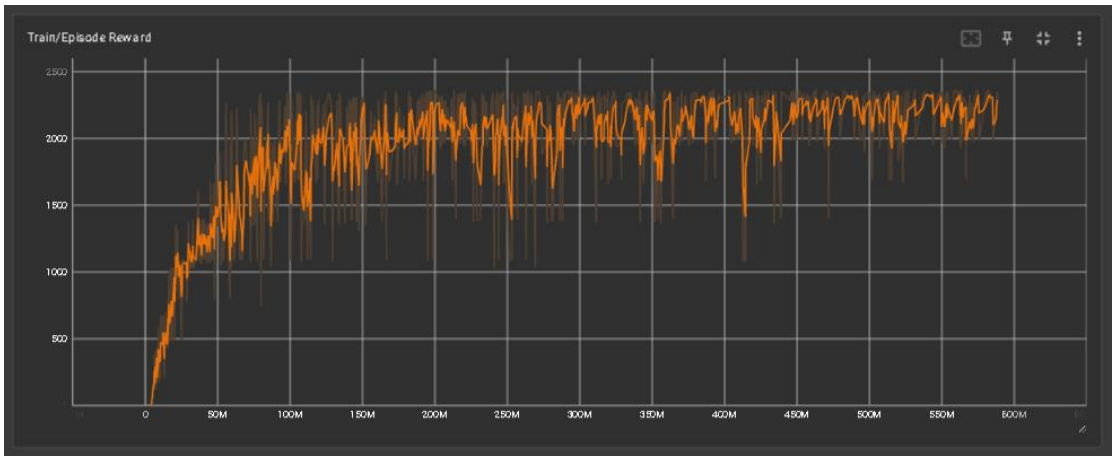
NYCU Reinforcement Learning Fall 2024

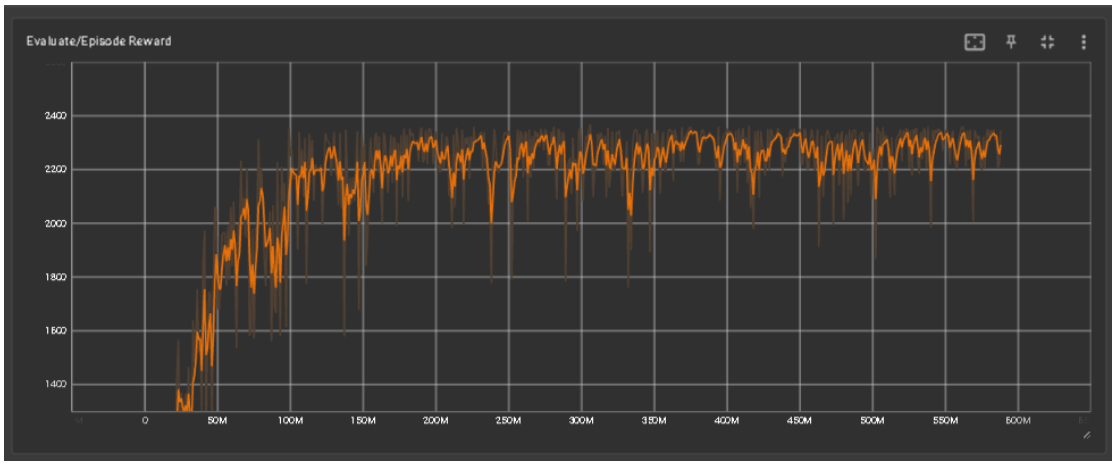
Date Submitted: November 10, 2024

■ Screenshot of Tensorboard training curve and testing results on PPO.

E.g.

Training curve:





Testing results (5 games

```
================================================
Evaluating...
episode 1 reward: 2356.0
episode 2 reward: 2319.0
episode 3 reward: 2361.0
average score: 2345.3333333333335
================================================
```

■ Answer the questions (bonus) (20%)

1. PPO is an on-policy or an off-policy algorithm? Why? (5%)

PPO 是一種 On-Policy 演算法。這是因為 PPO 在更新策略時，直接使用當前策略所收集的經驗數據。它依賴於最新的策略來生成行為，並根據這些行為與環境互動的 trajectory 來更新 network。因此，PPO 的學習過程依賴於與當前策略一致的數據，故為 on-policy 演算法。

2. Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization. (5%)

PPO 通過引入 clipping 機制來限制策略更新的幅度，從而避免策略變動過大而導致不穩定。
具體來說，PPO 的目標函數包含一個 ratio 來衡量新舊策略的變化。通過對這個 ratio 應用 clipping，可以防止 policy ratio: $r_t(\theta)$ 偏離 1 過多（即過大或過小），將其限制在 $[1-\epsilon,1+\epsilon]$ 範圍內。如果 $r_t(\theta)$ 超出這個範圍，目標函數會使用 clipping 後的值。這種方法有效地控制了每次更新的步伐，防止策略劇烈變動，從而維持學習的穩定性。

2. Why is GAE-lambda used to estimate advantages in PPO instead of just one-step advantages? How does it contribute to improving the policy learning process? (5%)
3.

PPO 使用 GAE-λ來估計 advantages 比僅使用 one-step advantages 具有更好的效果。因為使用 one-step advantages 雖然可以減少 variance，但也讓整體 Bias 較高。GAE-λ 結合了 multi-step advantages 估計，透過引入一個折扣因子 λ，平衡了 Bias 與 variance，捕捉更長期的依賴關係。使得使用 GAE-λ能提供更準確和穩定的優勢估計，從而提升 PPO 的策略學習效率和性能。

4. Please explain what the lambda parameter represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO? (5%)

在 GAE-λ中，λ 是一個介於 0 和 1 之間的參數，用來調整在當前 step t 往前看 G_t 到 G_{t+n} 所佔的權重比例。具體來說，λ 決定了在計算 advantages 時，未來回報的影響範圍：

- λ 接近 1：有助於捕捉更長期的依賴關係，從而減少偏差，但可能增加方差。可能提升策略的穩健性和最終性能，但也可能導致訓練過程中方差增大，學習不穩定。

- λ 接近 0：主要依賴於 one-step reward，使 advantages 估計更加穩定，訓練過程中 variance 較小，但可能無法充分利用長期回報的信息(增加 Bias)，限制了策略的表現。