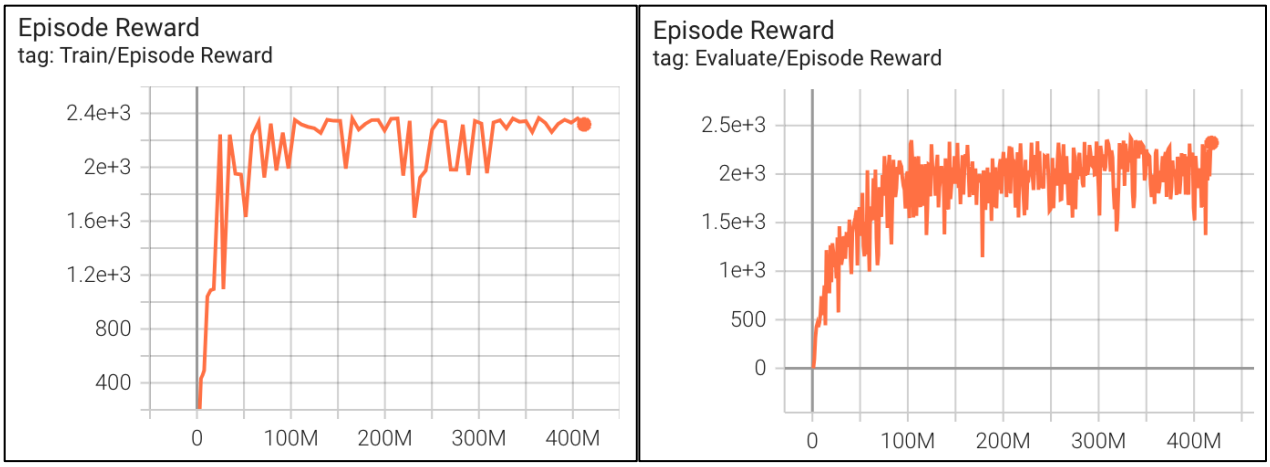
Proximal Policy Optimization
Lab Report # 3

By
312581020
許瀚丰

Selected Topics in Reinforcement Learning
Fall 2023
Date Submitted: November 12, 2023

- Screenshot of Tensorboard training curve and testing results on DQN. (30%)
  - Training curve



  - Testing results (5 games)

```
episode 1 reward: 2363.0
episode 2 reward: 2278.0
episode 3 reward: 1983.0
episode 4 reward: 2264.0
episode 5 reward: 2339.0
average score: 2245.4
```

- PPO is an on-policy or an off-policy algorithm? Why? (5%)
  - PPO 是 on-policy 的演算法，原因是因為 PPO 與 DQN、DDPG 不同，不需要使用 Replay Buffer 來儲存過去的 trajectory 來更新網路，而是透過當前與環境互動的 trajectory 來更新網路，並透讓每次更新的幅度不會過大防止網路在更新時不穩定。
- Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization. (5%)
  - 在 PPO 更新 Policy network 時使用了 clip ratio 的方式來防止每次更新的浮動不要過大，來防止網路在更新時不穩定，而在實際上，會將 ratio 限制

在 $1 - \epsilon \sim 1 + \epsilon$ ，例如，若今天 ratio $> 1 + \epsilon$ 且 $A_t > 0$，就上述的 clip ratio 機制，此時網路就不會更新了，而對於 ratio $< 1 - \epsilon$ 且 $A_t < 0$ 亦相同。

- 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%)
  - 由於使用 one-step advantages 雖然可以減少 variance，但也讓整體 Bias 較高。因此在 PPO 中透過使用 GAE 就可以同時平衡兩者，在訓練上就能在確保訓練穩定的同時也顧及的整體的準確程度。
- 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%)
  - GAE 的 lambda 與 TD(lambda)類似，皆是用來調整在當前 step t 往前看 G_t 到 G_{t+n} 所佔的權重比例，因此當 lambda 靠近 0 時，表示越靠近當前 step t 的 G 的的權重應該會越大，因此會更加強調短期的效果，相反的，若 lambda 靠近 1，則更考慮未來多步的平均，因此更強調長期的效果。