RL Lab 3 Report

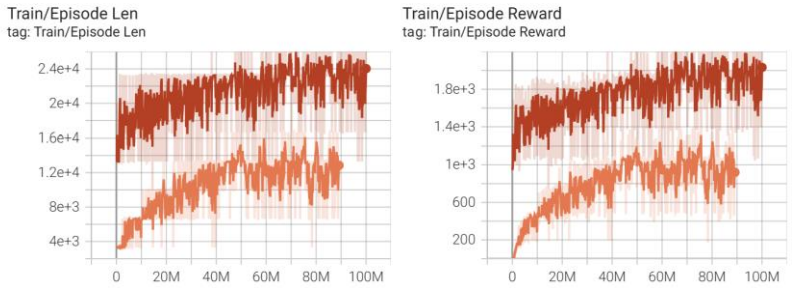
資科工碩 陳冠廷 313551058

Basic

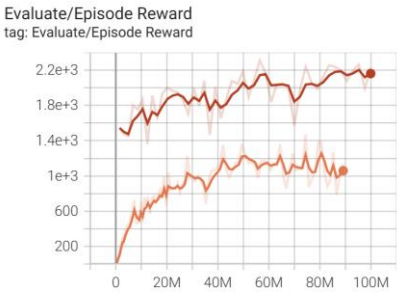
Screenshot of Tensorboard training curve and testing results on PPO.

由於原始的參數 PPO 訓練後期會卡住，所以在約 90M 的地方提早終止，並用更小的 learning rate (1e-6) 繼續訓練模型，因此下面的 curve 會有兩個 curve，橘色為第一次訓練的結果，紅色為第二次訓練的結果。

Training Curve



Evaluate Curve



Testing Results

```
================================================
Evaluating...
episode 1 reward: 1982.0
episode 2 reward: 2354.0
episode 3 reward: 2333.0
average score: 2223.0
================================================
```

Bonus

PPO is an on-policy or an off-policy algorithm? Why?

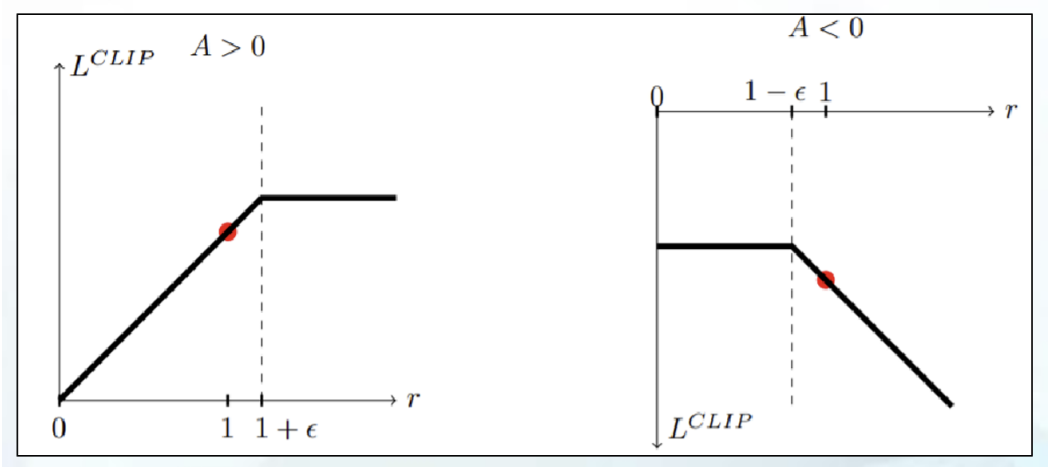
PPO 是一種 on-policy 演算法，因為從下方的 pseudo code 可以看到模型在更新參數所使用的資料是用自己的 policy 跑出來的結果。每個 iteration 中，同一個 actor 參數會執行多次產生多個 timestamp 長度為 T 的結果，最後將資料以 batch 的方式更新 actor 參數，因此為 on-policy(當前的 actor 並沒有用到其他參數所產生的 trajectories)。

Algorithm –Proximal Policy Optimization (PPO):

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

- **Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization.**

PPO 在訓練時加入了 Clip 的機制，將 policy loss 變化限縮在 1-ϵ 和 1+ϵ 之間，這樣可以確保 policy 更新不會偏離 old policy 太遠，避免了因為學習率過大而導致極端不穩定的行為，使得策略無法收斂，同時也保持了更新的有效性。(本實驗使用 epislon=0.2)



- **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?**

one-step advantage ($A_t = r_t + \gamma V(s_{t+1}) - V(s_t)$)，僅使用現在的 reward 和下一步 state 的值來估計 advantage，這種方法的 bias 小但是 variance 大，容易讓訓練不穩定。

GAE-lambda 採用超參數λ，在 one-step advantage 和 multi-step reward 之間進行平衡。GAE-lambda 使用了一系列加權的 TD-errors 估計 advantage ($\widehat{A_t^{GAE}} = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}$)，λ 值越大，估計的 advantage 就越傾向於使用長期回報； λ 值越小，估計的 advantage 就更依賴於短期回報。GAE-lambda 降低 bias 也減少 variance，使 policy 更新更加平滑。

- **Please explain what the lambda parameter represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO?**

GAE-lambda 使用了一系列加權的 TD-errors 估計 advantage ($\widehat{A_t^{GAE}} = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}$)，λ 值越大，$\gamma\lambda$的值就會愈大，$(\gamma\lambda)^l$遞減較慢，因此估計的 advantage 可以看到後期的δ值，因此會傾向於使用長期回報；反之，λ 值越小，$(\gamma\lambda)^l$遞減非常快，估計的 advantage 或著重在

前期的δ值，後期的δ值的貢獻較小，因此會傾向於使用短期回報。如果λ接近 0 則會變成類似 one-step advantage，bias 大導致訓練不穩定，容易 overfit 到短期波動；如果λ接近 1 則會變成類似 Monte Carlo，variance 大且學習效率低。(本次實驗使用 lambda=0.95)