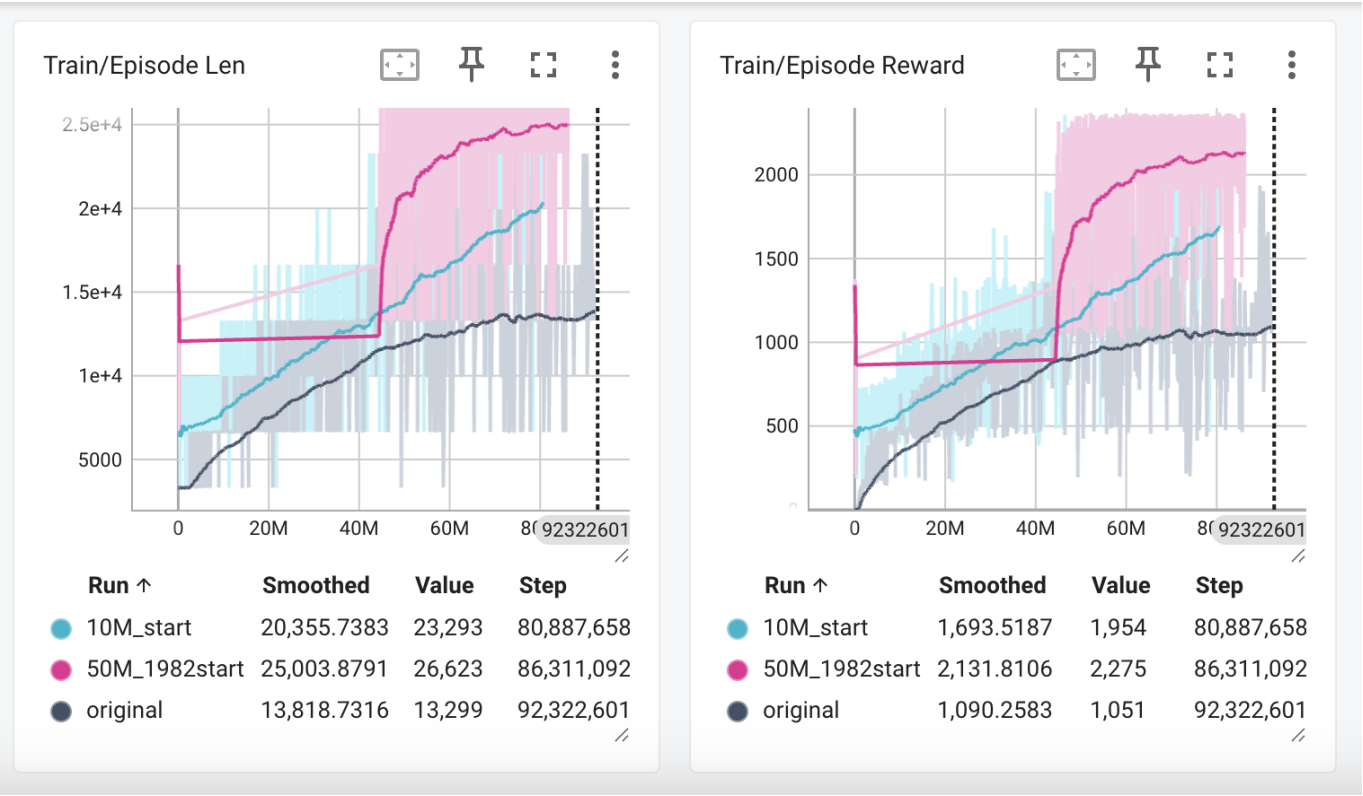
Lab 3: Proximal Policy Optimization (PPO)

314553017 蕭希敏

Results



Run ↑	Smoothed	Value	Step
10M_start	20,355.7383	23,293	80,887,658
50M_1982start	25,003.8791	26,623	86,311,092
original	13,818.7316	13,299	92,322,601

Run ↑	Smoothed	Value	Step
10M_start	1,693.5187	1,954	80,887,658
50M_1982start	2,131.8106	2,275	86,311,092
original	1,090.2583	1,051	92,322,601

> original為直接原版的code，開始train，learning rate = 2.5e-4

> 10M_start為抓original約10M時的.pth(700分)開始train，learning rate = 2.5e-4(步數前面沒更新)

> 50M_1982start為抓10M_start約40M時的.pth(1982分)開始train，learning rate = 2.5e-5(步數前面有更新)

```
正在載入模型: log/Enduro_release2/model_62526550_2332.pth
==============================================
Evaluating...
episode 1 reward: 2323.0
episode 2 reward: 1697.0
episode 3 reward: 1076.0
episode 4 reward: 2368.0
episode 5 reward: 2366.0
average score: 1966.0
==============================================
```

Question

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

PPO是一種on-policy演算法。 PPO主要是利用current policy去收集一批trajectories，並用同一批trajectories去更新Policy Network，通常會在這同一批trajectories上進行多次epochs更新。和DQN, DDQN不同，不會用Replay Buffer儲存大量過去採集到的經驗。PPO一旦策略被更新，舊的trajectories就會被丟棄，接著Agent會去收集新的trajectories去更新Policy Network。

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

  PPO會使用clipping，設定一個小參數ε，定義一個可接受的ratio範圍[1-ε, 1+ε]來限制policy updates，比如如這次作業ε=0.2，範圍就是[0.8, 1.2]，這時policy updates最多降至80%和升到120%。這樣子就能保證新policy保持在舊policy的一個Trust Region內，防止了過大浮動的更新，確保了stabilization 。

- **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 Advantages雖然因只依賴於單步的reward和下一個狀態的值估計，所以低Variance。但準確性高度依賴 V(s) 的準確性。一旦 V(s) 不準，advantage就會有偏差，所以高Bias。 GAE-lambda融合了多步的真實 reward，降低了 Bias，同時又不像Monte Carlo那樣累加所有噪聲，降低了 Variance。因此更準確、更穩定的 Advantage 估計能提供更可靠的 gradient，從而提高學習效率。

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

  lambda介於 0 到 1 之間，控制 Bias-Variance Tradeoff，影響：

  - lambda = 0：等同於「One-step Advantage」。估計值會高 Bias、低 Variance。
  - lambda = 1：等同於「Monte Carlo Advantage」。估計值會低 Bias、但高 Variance。
  - lambda = 0.95：在兩者間取得平衡點，可以看得夠遠（低 Bias），又能保持穩定（低 Variance），通常可以帶來較好的訓練效能，較常用。