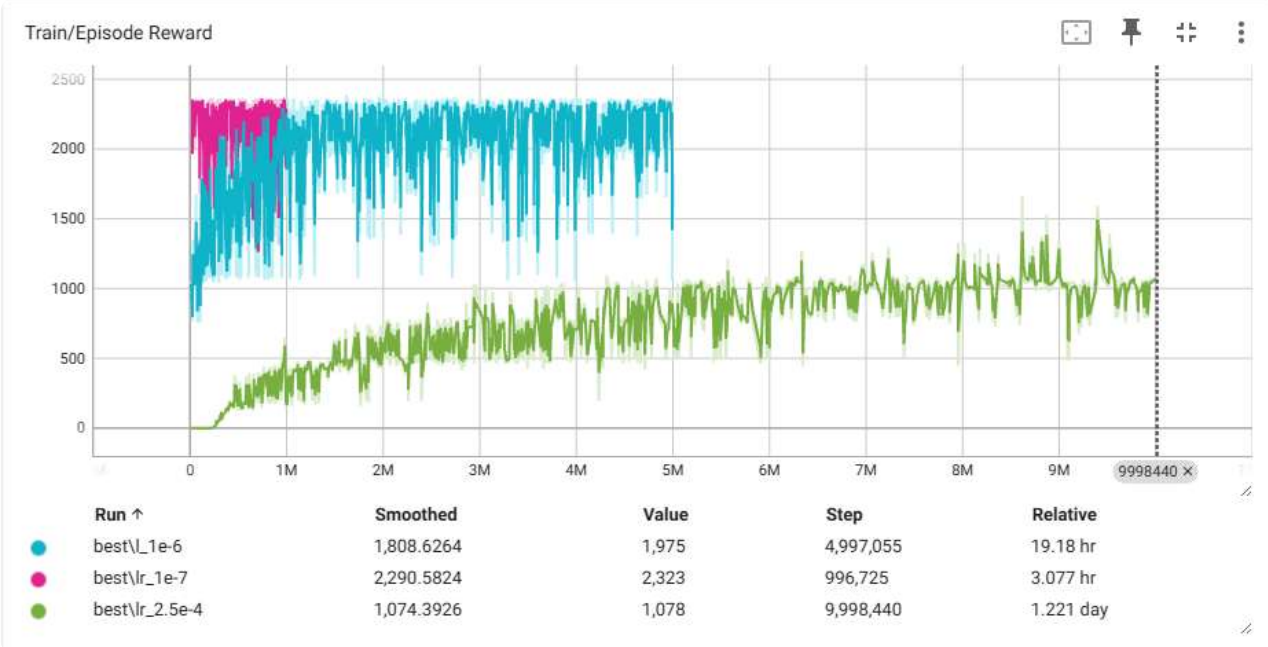
LAB3

Screenshot of Tensorboard training curve and testing results on PPO

Training curve



Run ↑	Smoothed	Value	Step	Relative
best\l_1e-6	1,808.6264	1,975	4,997,055	19.18 hr
best\lr_1e-7	2,290.5824	2,323	996,725	3.077 hr
best\lr_2.5e-4	1,074.3926	1,078	9,998,440	1.221 day

我第一遍照著 PPO 論文的Atari 參數[1]，用學習率 2.5e-4 平行 8 個 Agent 跑 10M 個 timesteps。

接著降低學習率後用 1e-5 跑 5M 個 timesteps 達到不錯的成績，最後再嘗試用更小的 1e-7 去 finetune 1M 個 timesteps，但沒有顯著成長。

Evaluate curve

Run ↑	Smoothed	Value	Step	Relative
best\l_1e-6	2,348.3876	2,346	4,698,222	17.4 hr
best\lr_1e-7	2,353.9496	2,350	954,908	1.978 hr
best\lr_2.5e-4	1,211.9028	1,055	9,895,264	1.204 day

Testing result

```
Evaluating...
episode 1 reward: 2346.0
episode 2 reward: 2346.0
episode 3 reward: 2346.0
episode 4 reward: 2346.0
episode 5 reward: 2346.0
average score: 2346.0
```

可能是我在 make test env 的時候有把 repeat action probability 設為 0，所以五次跑出來的結果相同，影片也相同。

```
1  if repeat_action_probability is not None:
2      env = gym.make(env_id, repeat_action_probability=repeat_action_probability, render_mode='rgb_array')
```

Bonus

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

PPO 原則上是 on-policy，每次更新 policy 時只使用當前 policy 產生的資料。
但是因為每次更新時會用 GAE buffer 中的資料更新 update_count 次，嚴格定義上，只有迴圈的第一圈是 on-policy，迴圈第一圈更新後，迴圈第i圈都不是用當前 policy(迴圈i-1圈更新後的policy)生成的資料去更新，所以嚴格定義的話應該算是 near on policy[2]。

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

PPO 利用裁剪機制 (Clipped Surrogate Objective) 來限制更新，將新舊 policy 比率限縮在 $[1-\epsilon, 1+\epsilon]$ 的範圍內，以此限制每次更新的幅度，避免更新過大導致不穩定。

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?

GAE-lambda 綜合了多步的資訊，降低了訓練的 bias 和 variance，幫助我們更準確的估計 advantage，並提升模型的訓練效率。

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 代表對於多步回報的 discount factor。
當 lambda 接近 1 時，GAE 會更傾向長期回報，以此降低 bias，增加訓練的穩定性，但會增加 variance 使得收斂速度變慢，適合長期學習。
當 lambda 接近 0 時，GAE 會更傾向即時回報，這樣會減少 variance，加快模型的收斂速度，但會使 bias 較大，使得 policy 偏離最佳解，適合短期的學習。

Appendix

1. PPO 論文的 Atari 參數

Hyperparameter	Value
Horizon (T)	128
Adam stepsize	$2.5 \times 10^{-4} \times \alpha$
Num. epochs	3
Minibatch size	32×8
Discount (γ)	0.99
GAE parameter (λ)	0.95
Number of actors	8
Clipping parameter ϵ	$0.1 \times \alpha$
VF coeff. c_1 (9)	1
Entropy coeff. c_2 (9)	0.01

Table 5: PPO hyperparameters used in Atari experiments. α is linearly annealed from 1 to 0 over the course of learning.

↩

2. 深度解读：Policy Gradient，PPO及PPG - 知乎 (https://zhuanlan.zhihu.com/p/342150033) ↩