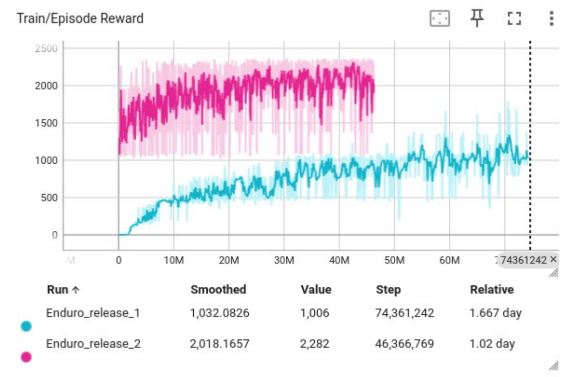
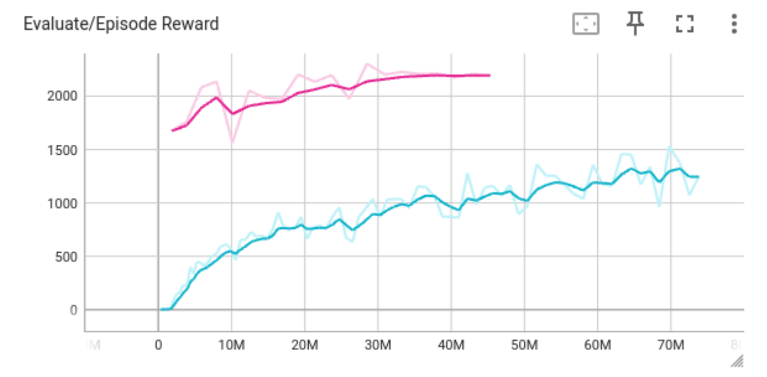
RL_LAB_3_313552041_洪日昇

Report (30% + Bonus 20%)

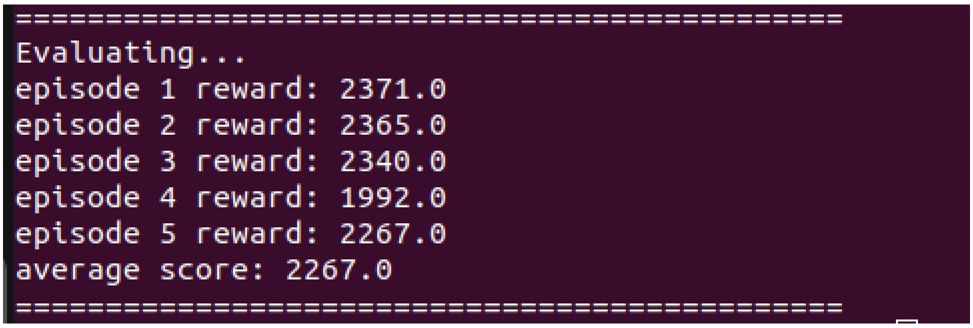
Screenshot of Tensorboard training curve and testing results on PPO.



	Run ↑	Smoothed	Value	Step	Relative
●	Enduro_release_1	1,032.0826	1,006	74,361,242	1.667 day
●	Enduro_release_2	2,018.1657	2,282	46,366,769	1.02 day



逐步改小 lerning rate 會幫助模型再次增長，同時也不能一開始就用過小的 lerning rate。
release_1 用 2.5e-4, release_2 用 2.5e-5。



```
============================================
Evaluating...
episode 1 reward: 2371.0
episode 2 reward: 2365.0
episode 3 reward: 2340.0
episode 4 reward: 1992.0
episode 5 reward: 2267.0
average score: 2267.0
============================================
```

Questions: (20%)

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

PPO is an on-policy reinforcement learning algorithm.
PPO 是仰賴 current policy 所收集的數據，PPO 利用 current policy 與 old policy 之間的 probability ratio 來確保新的 policy 不會偏離舊的 policy 太多，利用 clip ratio 來防止每次更新的幅度不要太大導致不穩定。

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

PPO 利用 clipping 讓 policy updates 在 safe range between 1−ϵ and 1+ϵ, where ϵ is a small constant，確保 policy 只會改變最多 ϵ (e.g., ϵ = 0.1, 10%)。

PPO 利用 clipping 來限制，確保 gradual, stable improvements，allows PPO to maintain both stability and effectiveness, avoiding the large, destabilizing policy changes that can occur in other policy gradient methods.

3. 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%)

PPO 利用調整 GAE-λ，調整高低(0-1)，也就是考慮長期效果的多寡，來達成 bias and variance 的平衡，可以增進 stability and efficiency of policy learning。

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%)

如果 λ 接近 1，GAE-λ 就會使用更多步驟來計算 advantage，這樣可以通過考慮 long-term effects 來減少 variance。計算出的 advantage 更平滑，對於短期回報的波動不會過於敏感，進而穩定策略更新。

如果 λ 接近 0，GAE-λ 就會類似於 one-step advantages，這種情況下，bias 會更小，但 variance 較高。這是因為 one-step advantages 計算，可能會受短期回報的波動影響，使得策略更新較不穩定。