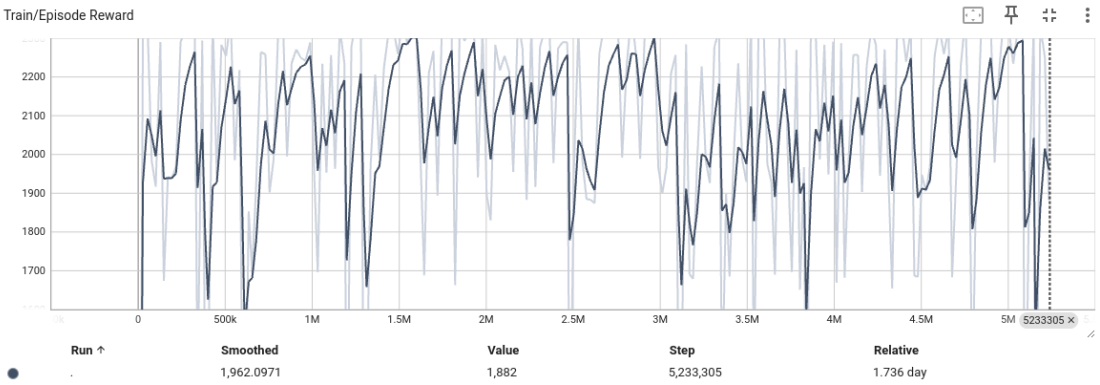
RL LAB3 report 313552042 陳品翰

Screenshot of Tensorboard training curve and testing results on PPO




```
episode 1 reward: 2327.0
episode 2 reward: 2316.0
episode 3 reward: 2351.0
average score: 2331.3333333333335
================================================
```

Bonus:

1.1 PPO 是一個 on policy 的方法

1.2 因為只會用當前策略得到的樣本來進行訓練，不像 off policy 會拿其他的策略得到的樣本數來做訓練。

2.PPO 會限制每次的更新幅度，更新幅度的限制是基於訓練者自己的設定，例如這次的作業中的限制就由 clip_epsilon 這個變數設定，更新的幅度會介於 1 + epsilon 跟 1 – epsilon 之間，以此來保證學習過程中不會有一次更新太大而導

致學習不穩定的情形。

3.1 因為藉由 GAE 可以加權往後的多個 STEP，這樣做的好處是可以讓估計值更加的平滑，也可以讓 PPO 的訓練過程更加穩定，如果用 one-step advantages，可能會無法充分利用長期的資訊，造成學習偏離最佳的方向。

3.2 因為 GAE 會看往後多步的訊息，所以在估計該動作的優勢時，可以有一個比較平滑的結果，又因為 PPO 本身有 Clipped Surrogate Objective，提供一個比較平滑的估計值可以讓他們之間相輔相成，讓整個學習的過程更穩定。

4 λ 越接近 0 代表越傾向單步估計，越接近 1 則會結合更多步的資訊，當接近 0 時，因為只看少步的資訊，所以學習的過程會比較快速，但是會比較不穩定，因為考量的資訊較少，如果越接近 1 則會考慮越多步的資訊，優點是可以有更好的學習穩定度，缺點則是比較耗時。