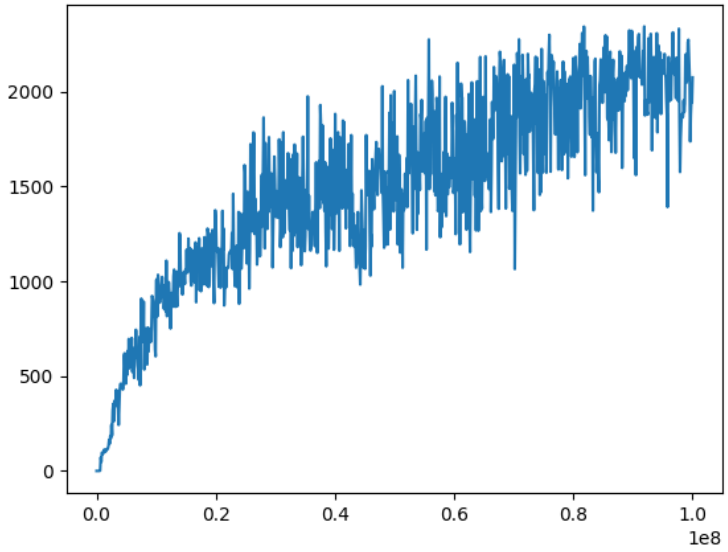
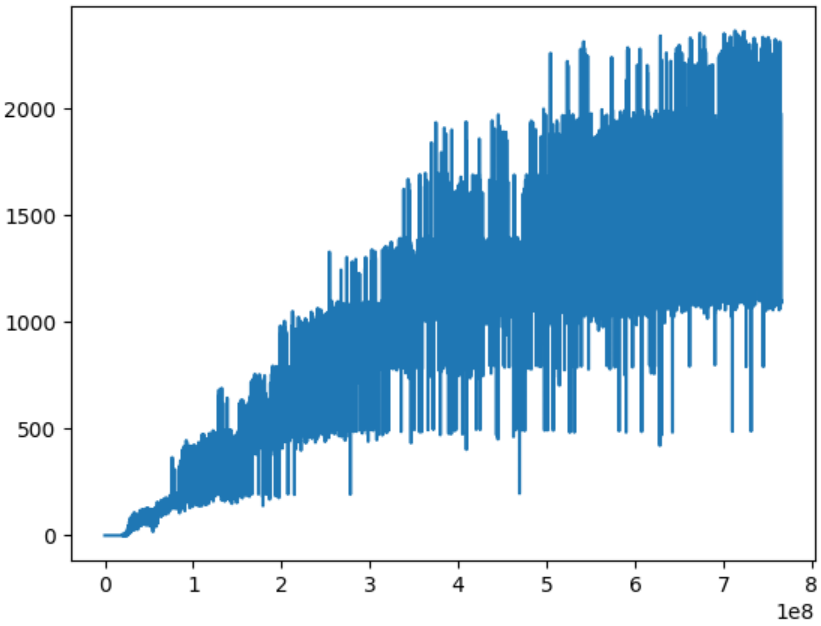
RL_Lab3 Report 312551163_陳允觀

Screenshots

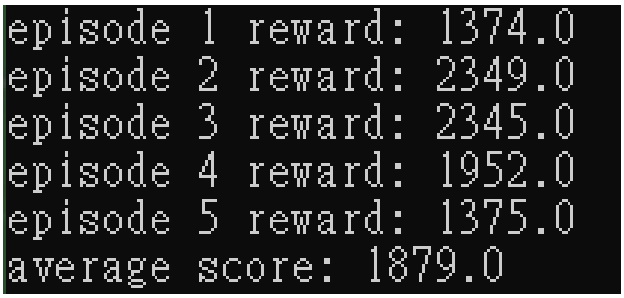
evaluate



train



Test



Questions

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

PPO is an off-policy algorithm because it does not need a replay buffer to store the past trajectory to update the net. Instead, it updates the net via trajectory from the environment and prevent unstable circumstances by huge update of net.

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

It uses clop ratio to avoid large updates at each step. It makes sure the ratio update is between (1+epsilon) and (1-epsilon). If the ratio is higher or lower than the bound, the network will not update.

- 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%)

While using one-step advantages can lower the variance, but this makes the bias higher. Therefore, PPO can use GAE to balance them and get stable training performance.

- 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%)

The lambda is like TD lambda, they are all used for calculating the weight between current step and previous steps. This enhances weights of long-term step values to make sure the agent can consider the policy from the whole game.