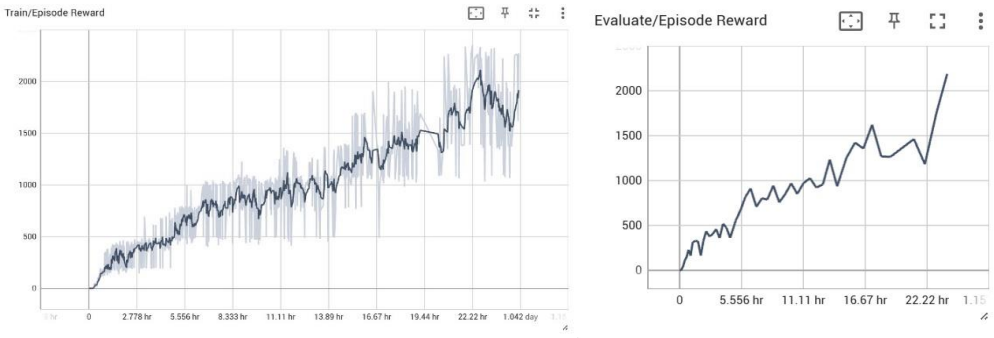
RL lab3

Name: 陳以瑄 Student ID:109705001

- **Screenshot of Tensorboard training curve.**



I conducted model training in four separate phases to manually perform learning rate decay.

Settings:

phase	1	2	3	4
learning_rate	2.5e-4	1.5e-4	5e-5	2.5e-5
training_steps	1e7	2e7	1e7	1e7

- **Screenshot of testing results.**

```
episode 1 reward: 2331.0
episode 2 reward: 1376.0
episode 3 reward: 2221.0
episode 4 reward: 1944.0
episode 5 reward: 1969.0
average score: 1968.2
```

- **Bonus Q1. PPO is an on-policy or an off-policy algorithm? Why?**

On-policy**.** In PPO algorithm, we run the policy $\pi_{\theta old}$ in the environment to collect experiences, and then use these experiences to update the same policy.

- **Bonus Q2. Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization.**

PPO uses a “clip” mechanism that sets the bound on L^{CPI}, in order to limit policy changes in each update, preventing large, destabilizing shifts.

- **Bonus Q3. 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?**

  Using one-step advantages might lead to a high bias in the estimated advantages because it considers only the immediate consequences of actions, ignoring potential long-term effects. The GAE method takes information from multiple time steps into account, providing a more accurate estimation. This results in more consistent and stable advantage estimates, making the learning process more efficient.

- **Bonus Q4. Please explain what the lambda parameter represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO?**

  The lambda parameter in GAE is used to balance the trade-off between bias and variance in estimating the advantage function. The range of lambda is [0,1]. If lambda is small, like 0, it mainly looks at recent rewards, and the policy updates respond quickly. However, using a small lambda might lead to high bias. If lambda is higher, it considers more future steps and gets a more accurate estimation. Since the estimation is less noisy, the training process is more stable, but a high lambda might lead to high variance and slow the convergence.