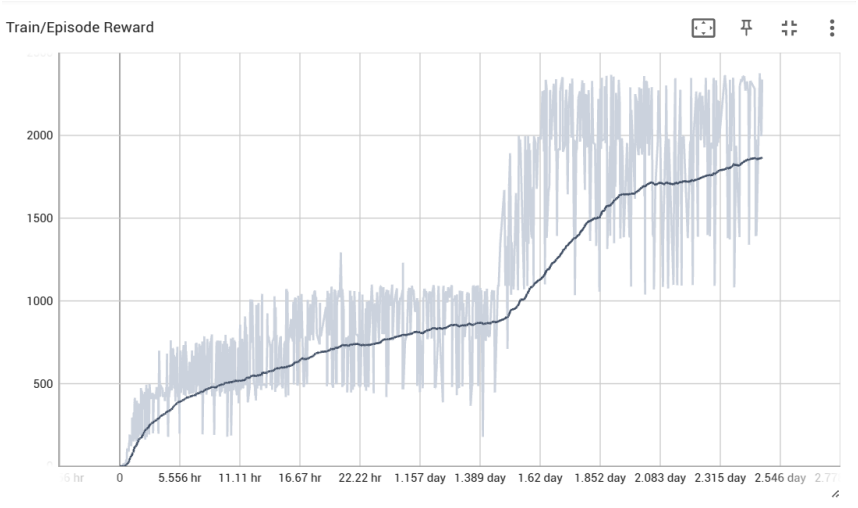
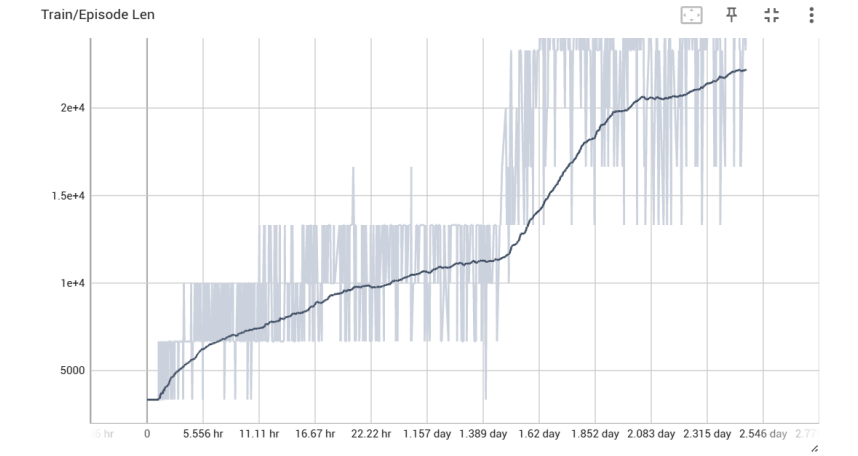
Experimental Results

Training curve

Learning rate for first half: 2.5e-4; second half: 2.5e-5





Testing Result

```
(RLenv) C:\Users\stanl\OneDrive - 國立陽明交通大學\桌面\RL3>python main.py
A.L.E: Arcade Learning Environment (version 0.8.1+53f58b7)
[Powered by Stella]
==========================================
Evaluating...
C:\Users\stanl\Envs\RLenv\lib\site-packages\gym\utils\passive_env_checker.py:233: DeprecationWarning: `np.bool8` is a
deprecated alias for `np.bool_`.  (Deprecated NumPy 1.24)
  if not isinstance(terminated, (bool, np.bool8)):
episode 1 reward: 2363.0
episode 2 reward: 1993.0
episode 3 reward: 1885.0
average score: 2080.3333333333335
==========================================
```

Bonus

1. PPO is an on-policy or an off-policy algorithm?

   PPO is a on-policy algorithm since it is updated by sampling actions according to the latest version of its policy.

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

   It avoids the large step using (1) learning rate, (2) the clipping strategy

   $$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}\ (r_t(\theta), 1-\epsilon, 1+\epsilon\ )\ \hat{A}_t \right) \right]$$

   If $A_t > 0$ and the ratio is not clipped, L_{CLIP} could be large and cause destabilization when updating.

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?

   If one-step advantages are used, the advantage equation becomes $r_t + V(S_{t+1}) - V(S_t)$, if V is not accurate enough, the advantage will be incorrect. Compared to GAE-lambda, bias of one-step method is higher and harder to converge. The GAE-lambda $\hat{A}^{\text{GAE}(\gamma,\lambda)} = \sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i}^{\pi_\theta}$ is less susceptible to error for V(). GAE-lambda take several steps into account, you can trade-off between bias and variance. If you had lambda properly set, you can have variance and bias small enough to make your model converge faster.

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?

   $$\hat{A}^{\text{GAE}(\gamma,\lambda)} = \sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i}^{\pi_\theta}$$

   Lambda is used to calculate the advantage (0<=lambda<=1). If we have a larger lambda, we will discount less in the future (take future value more into consideration), and since you take more future steps into consideration, the variance will be higher and the bias will be smaller. If you had smaller lambda, you will discount more in the furture (focus in short-term), making bias higher and variance smaller.