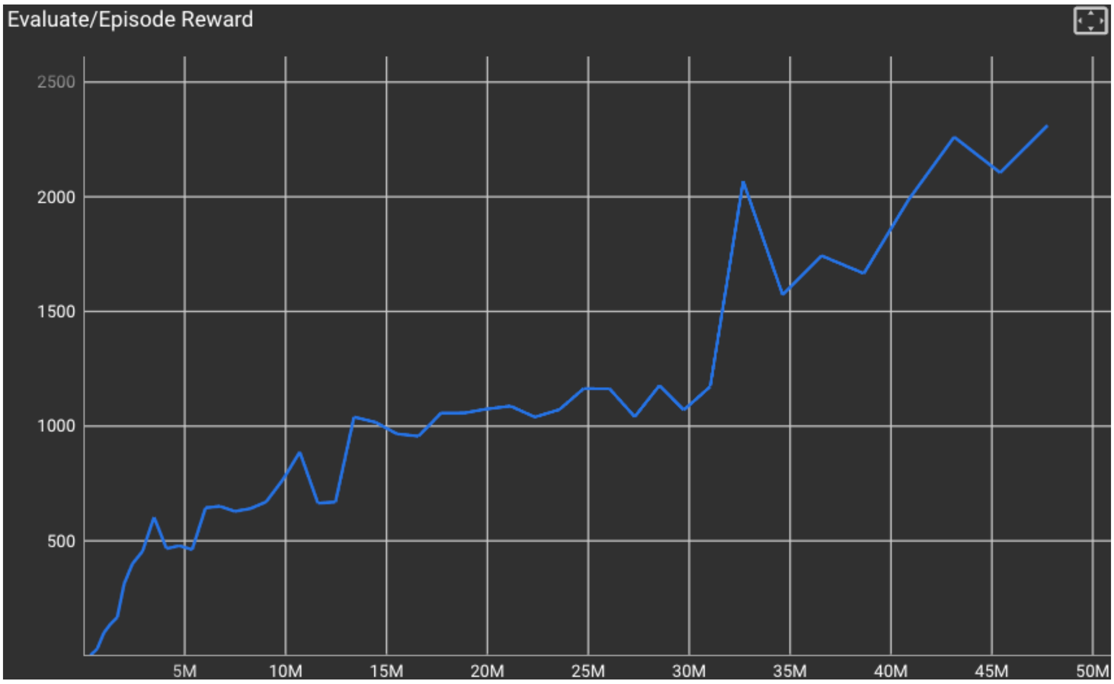
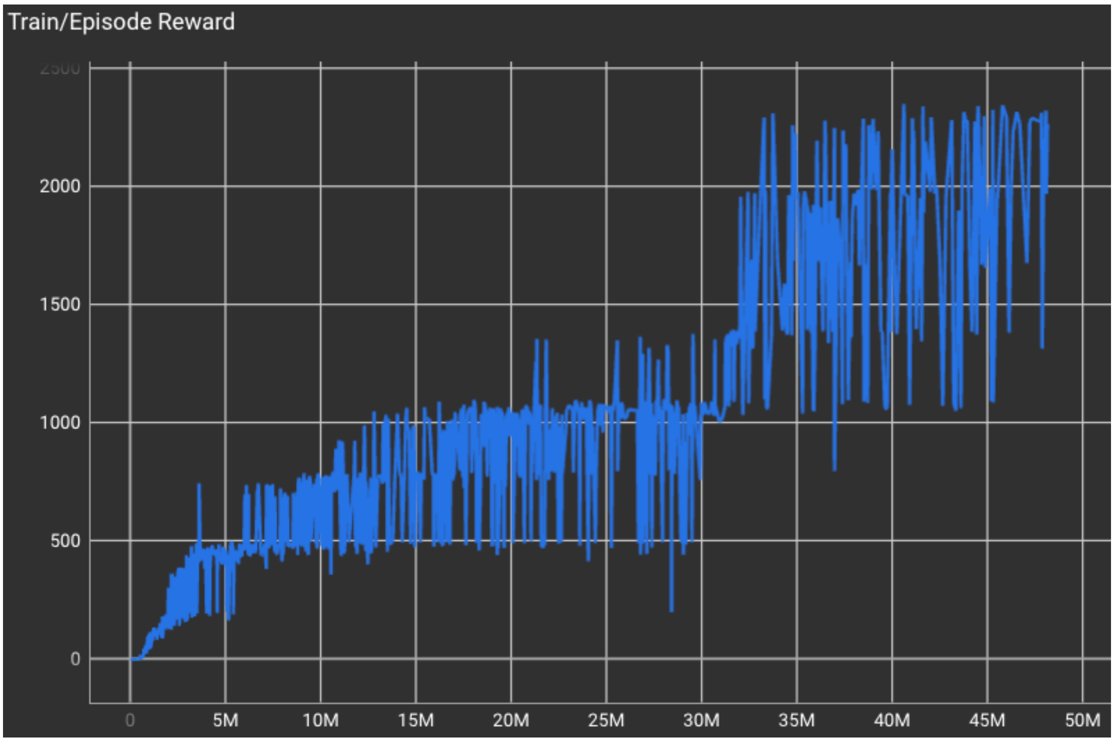
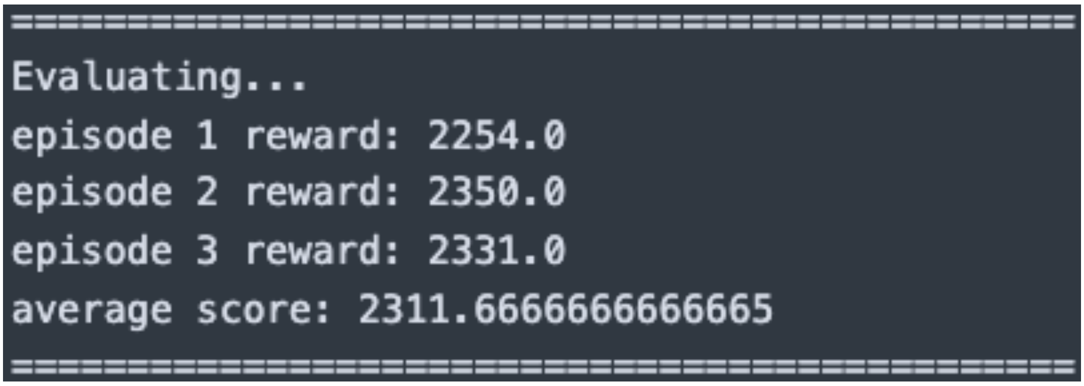
Lab3_311352004_童政瑜_report

1. Experimental Results







```
==========================================
Evaluating...
episode 1 reward: 2254.0
episode 2 reward: 2350.0
episode 3 reward: 2331.0
average score: 2311.6666666666665
==========================================
```

2. Bonus

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

   PPO is an on-policy algorithm. It operates on the current policy and collects new data by following the current policy. This means that the data used for policy updates is generated by the same policy being updated.

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

   PPO ensures that policy updates are not too large to avoid destabilization through the use of a "clipping" mechanism. When computing the surrogate loss, PPO introduces a constraint that limits how much the policy can change in each update. This constraint is controlled by the hyperparameter " clip_epsilon" in main.py. If the policy change (measured by the ratio of new policy probabilities to old policy probabilities) exceeds this clip parameter, the update is scaled down to stay within the clip range. This prevents large policy updates and helps maintain stability during training.

   (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?

   Generalized Advantage Estimation (GAE) with a lambda parameter is used in PPO to estimate advantages instead of just one-step advantages because it provides a more informative and accurate estimate of the advantages over multiple time steps. By taking into account not only the immediate reward but also the potential rewards in the future, GAE-lambda better captures the long-term effects of actions and helps stabilize the policy learning process. It reduces the high variance that can occur with one-step advantages and makes the updates less sensitive to the choice of a single time step.

   (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?

   The lambda parameter in GAE-lambda represents the weight given to the trade-off between considering short-term and long-term advantages. A lambda value of 0 corresponds to using only one-step advantages, and a lambda value of 1 corresponds to using advantages that consider all future rewards. Adjusting the lambda parameter affects the balance between bias and variance in advantage estimation. A lower lambda value (closer to 0) reduces the bias but increases the variance, while a higher lambda value (closer to 1) reduces variance but may introduce more bias. The choice of the lambda parameter can impact the training process and performance of PPO, with the ideal value often depending on the specific problem and environment. It's often tuned empirically to achieve the best results.