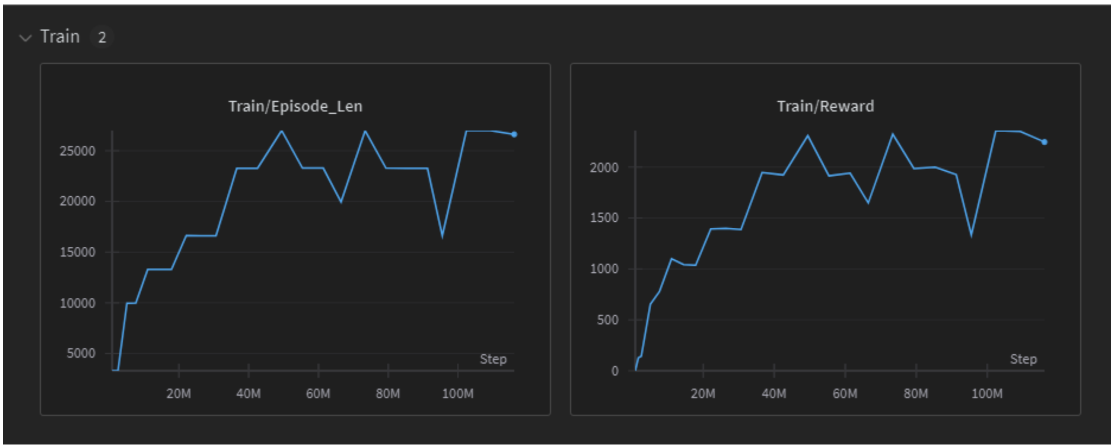
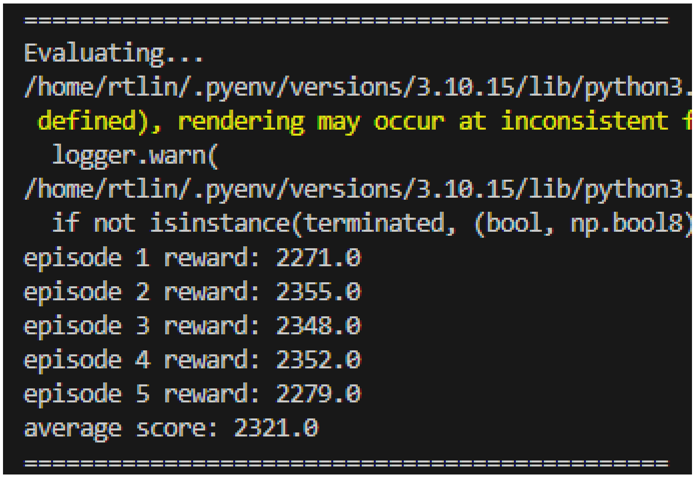
RL Topic HW3

1. Training curve and testing result of PPO Enduro:





```
=============================================
Evaluating...
/home/rtlin/.pyenv/versions/3.10.15/lib/python3.
 defined), rendering may occur at inconsistent f
  logger.warn(
/home/rtlin/.pyenv/versions/3.10.15/lib/python3.
  if not isinstance(terminated, (bool, np.bool8)
episode 1 reward: 2271.0
episode 2 reward: 2355.0
episode 3 reward: 2348.0
episode 4 reward: 2352.0
episode 5 reward: 2279.0
average score: 2321.0
=============================================
```

2. Questions:

a. PPO is an on-policy or an off-policy algorithm? Why?

PPO is an on-policy algorithm, since it only learns by the trajectories by the newest policy, instead of using trajectories from another agent.

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

PPO uses clipping to prevent the update step from being too large, it limits the difference of same action from current agent and updated agent within a range to maintain its stabilization.

c. 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?

GAE-lambda has low variance of estimation, while one-step advantage is a high variance.

Its low variance estimation can help the learning process to be more stable and efficient to reach an optimal policy.

d. 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 represents the decay of the combination of multi-step advantage estimation.
When lambda is close to 1, the algorithm considers more longer-term future rewards.
On the other hands, when it is close to 0, the agent would focus on one-step reward more