National Taiwan University
Department of Electrical Engineering
Name: TU-CHIN CHIANG
ID: R11921038

112-1 (Fall 2023)
Reinforcement Learning

Assignment 3 Report

Q1. Discuss how you speed up your grid world environment. (10 pts.)

1. *Remove the rendering process.* The rendering process is executed when the reset function is called and cost a number of amount of time. Therefore, removing the render process can improve the efficiency of training. In my computer that runs the codes, the performance of the running time decreases the percentage of 13%.

Q2. What's your best result in 2048? (5 pts.)

The charts depicted in Figure 1 is the best result of tile freaqency with tile value 1024 appears 4 times in 100 rollouts for evaluation.

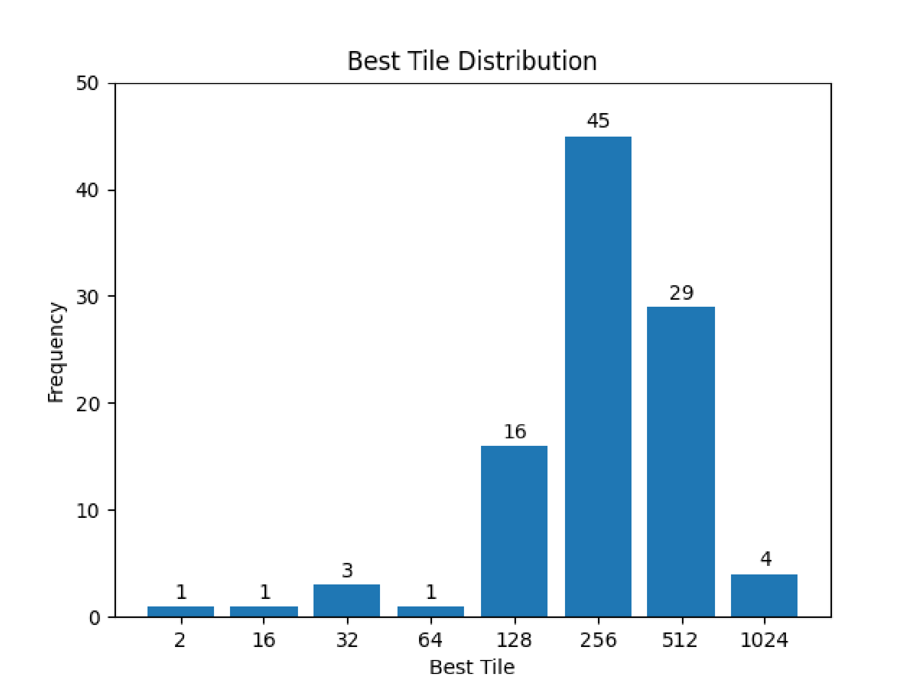

Figure 1: The tile distribution of the best model.

Q3. Describe what you have done to train your best model. (15 pts.)

The reinforcement learning algorithm used in this problem is PPO, and the evaluation during the training process is set the same as the final evaluation code. Besides, to make the model lean 2048 well, the number of timestamps, the mechanism of rewards, and the configuration of the reinforcement learning algorithm are modified.

1. *Increase timestamps.* Based on the empirical results, every time the configuration is adjusted, the algorithm might become somewhat unstable even though the final result is better than the past version. However, the model can gradually reach the higher score with a small increasing rate as long as the timestamps is large enough. Therefore, before other parameters of the model is optimized, the best result can achieve the medium baseline after 10^6 timestamps.

2. *Modify network architecture and increase learning rate of PPO.* SB3 policy consists of a feature extractor and a fully-connected network. To enhance the policy and value learning performance, the size of the element is enlarged. In addition, becasue the perturbation occurs in the learning curve of exp_4, which is depicted in Figure 2, I magnified the learning rate about three times and the corresponding result is exp_5.

3. *Adjust the reward mechanism.* For illegal moves, some additional trials are allowed by setting a foul counter and give a negative reward. For the preferred actions that encourages the tiles aggregating to one of the corners, rewards are increased by multiplying a scale factor.

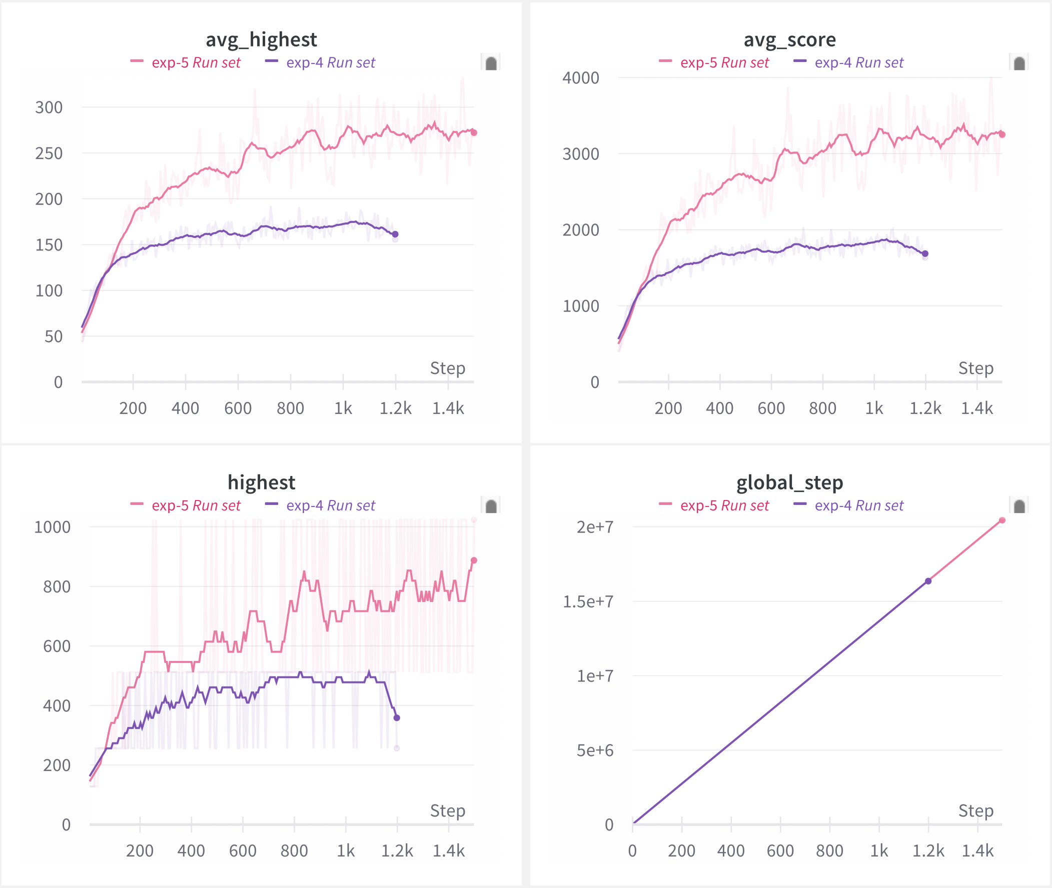

Figure 2: Training records of the two experiments.

Q4. Choose an environment from the Gymnasium library and train an agent. Show your results or share anything you like. (10 pts.)

I have chosen a classical Box2D game maintained by Gymnasium — Cart Pole. A screenshot that four environments was training in parallel and the learning curve of an experiment are illustrated in Figure 3. Due to the fact that the model is just trained naively, the result of this game is not optimized at all.

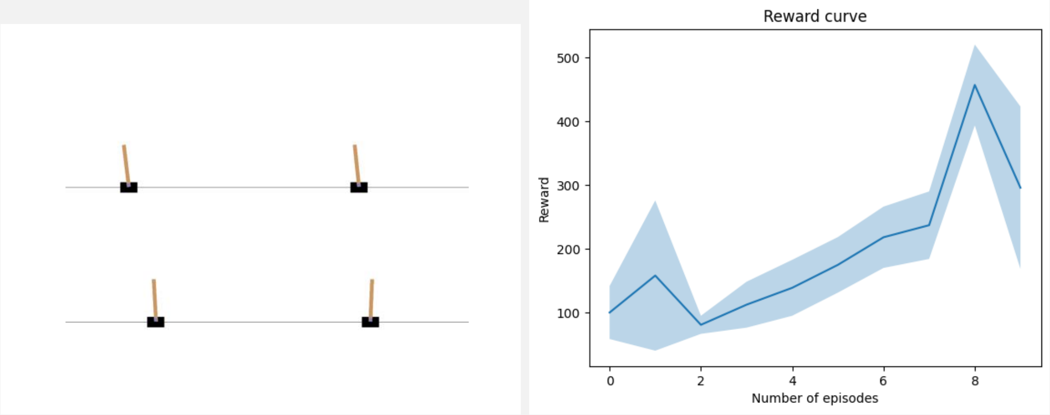

Figure 3: Four Cart Pole games and the learning curve.