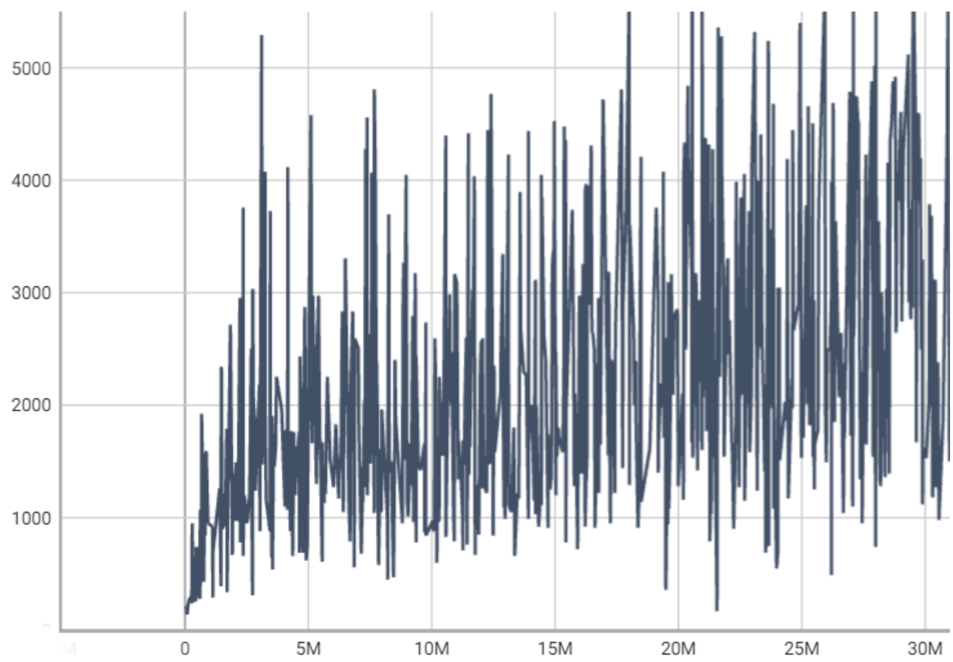
RL Lab2 Report

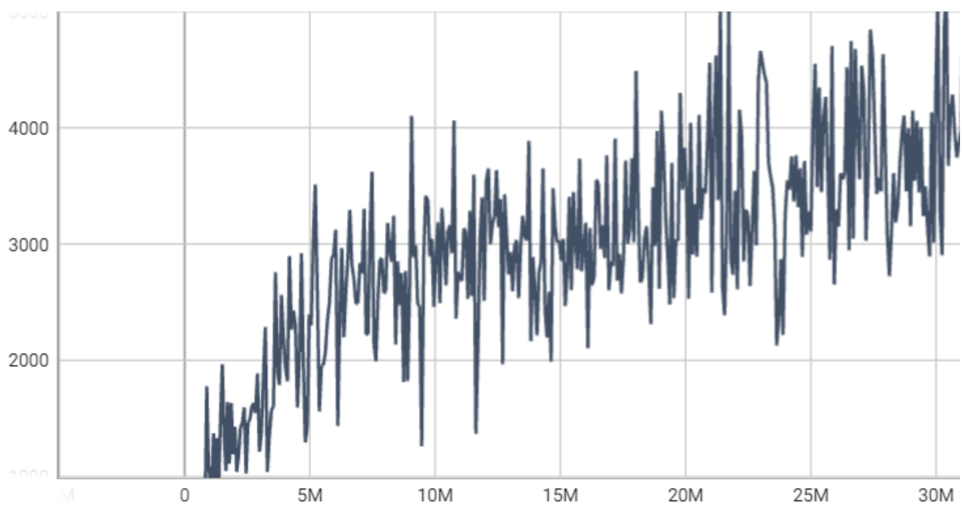
Student ID: 110550014 Name: 吳權祐

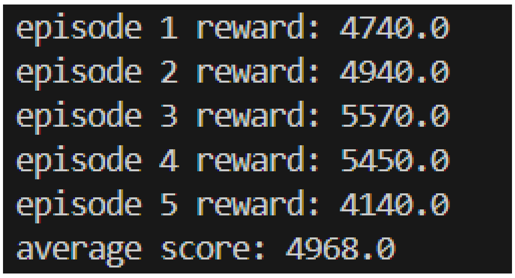
1. Training Curve & Testing Result

Train/Episode Reward



Evaluate/Episode Reward



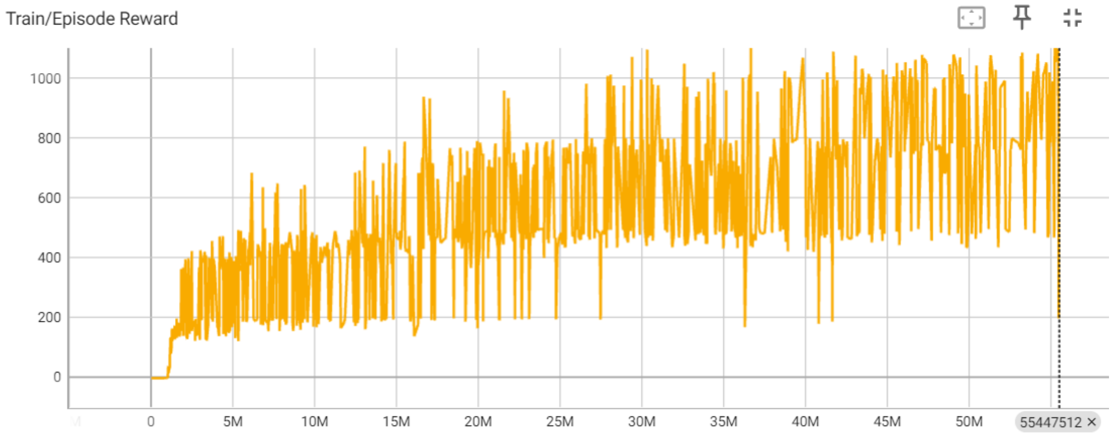


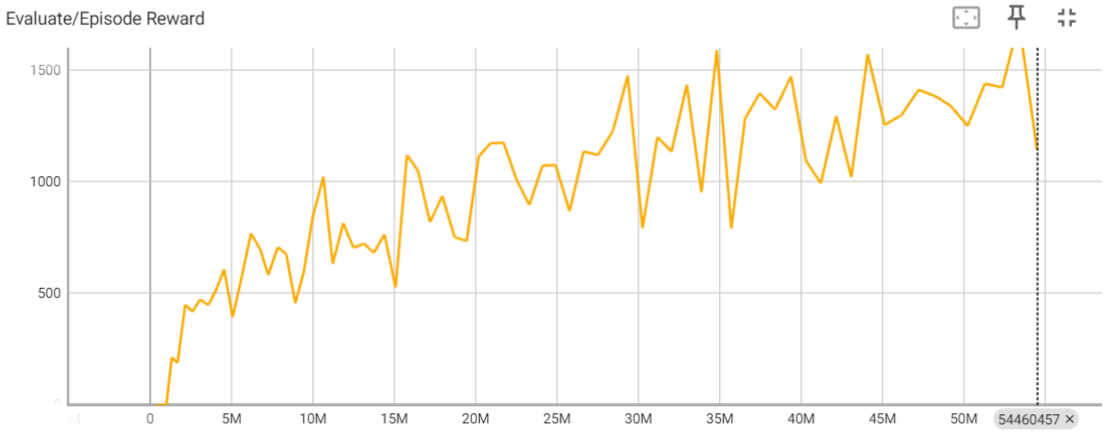
```
episode 1 reward: 4740.0
episode 2 reward: 4940.0
episode 3 reward: 5570.0
episode 4 reward: 5450.0
episode 5 reward: 4140.0
average score: 4968.0
```

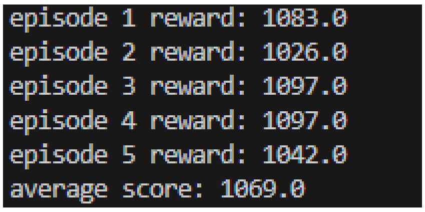
2. Bonus

I. Enduro-v5 (10%)

(1) Training Curve & Testing Result



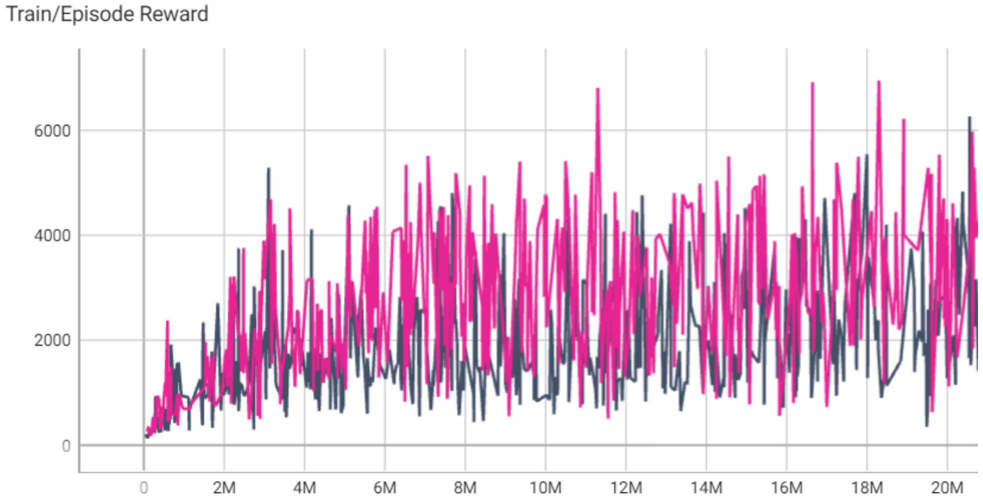


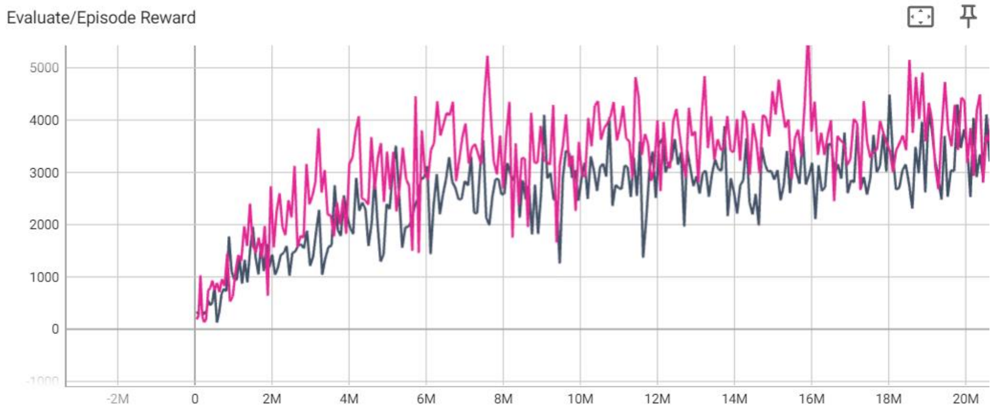


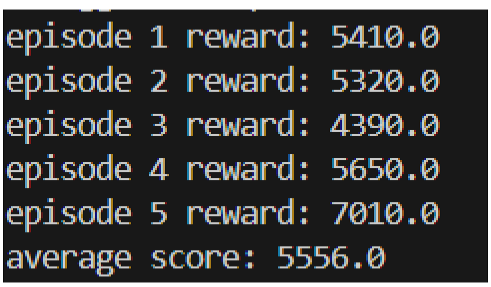
```
episode 1 reward: 1083.0
episode 2 reward: 1026.0
episode 3 reward: 1097.0
episode 4 reward: 1097.0
episode 5 reward: 1042.0
average score: 1069.0
```

II. DDQN (3%)

(1) Training Curve & Testing Result







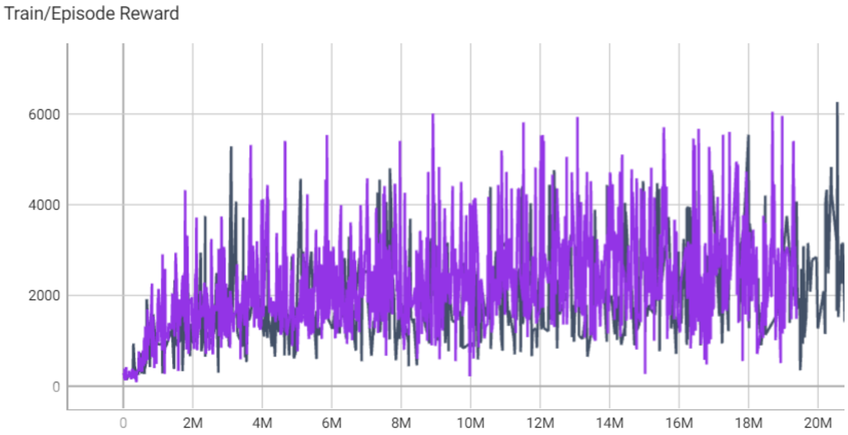
```
episode 1 reward: 5410.0
episode 2 reward: 5320.0
episode 3 reward: 4390.0
episode 4 reward: 5650.0
episode 5 reward: 7010.0
average score: 5556.0
```

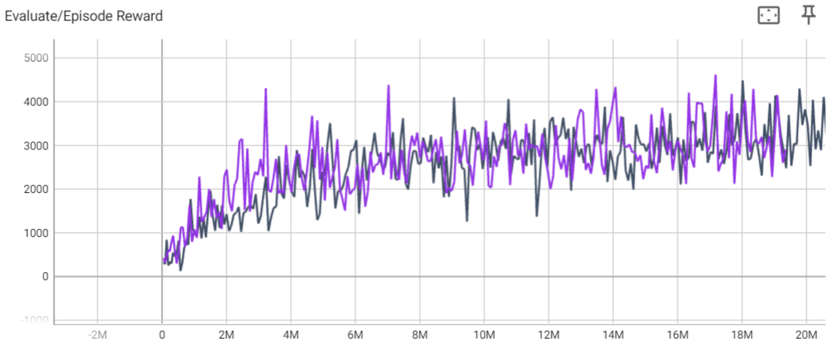
(2) Discussion

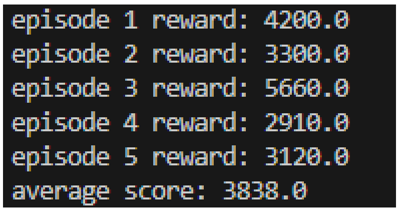
DDQN uses one network to select the best action first, then uses another network to estimate Q value, which avoids overestimation and leads to stable and better performance.

III. Duel DQN (3%)

(1) Training Curve & Testing Result







```
episode 1 reward: 4200.0
episode 2 reward: 3300.0
episode 3 reward: 5660.0
episode 4 reward: 2910.0
episode 5 reward: 3120.0
average score: 3838.0
```

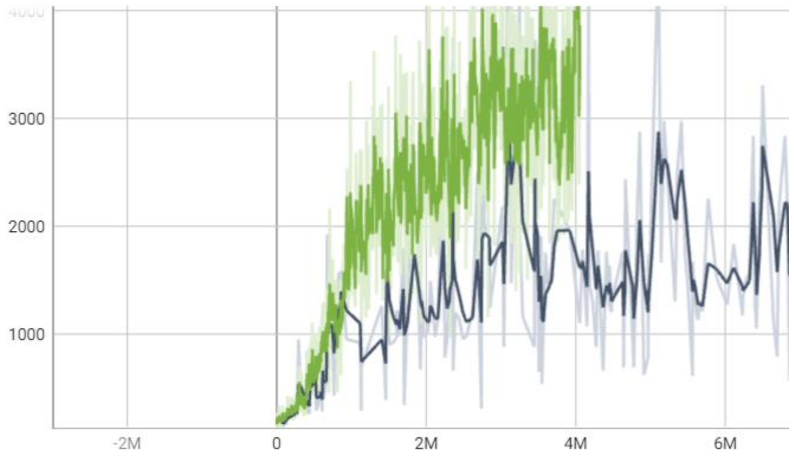
(2) Discussion

Dueling DQN modifies the network architecture to estimate a scalar value for the state and a vector for the actions, then uses the sum of these two predictions as the estimated Q value. And it gets better performance than typical DQN.

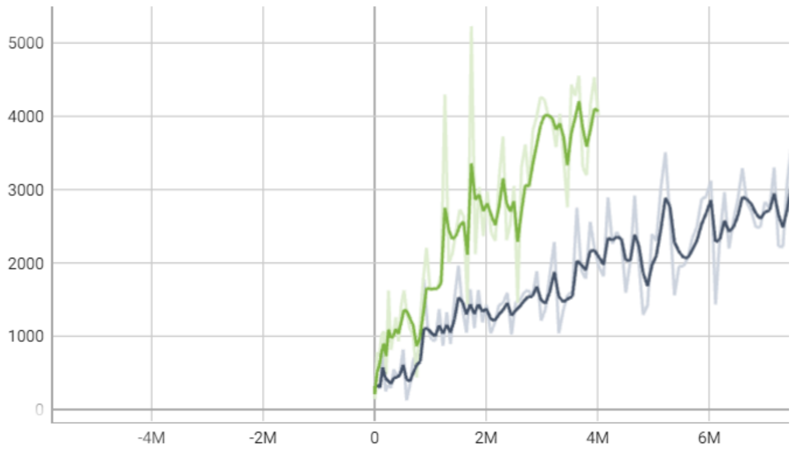
IV. DQN with parallelized rollout (4%)

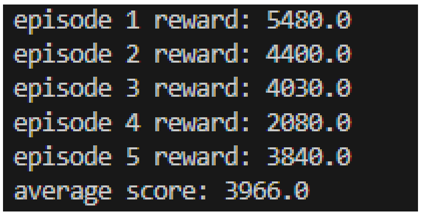
(1) Training Curve & Testing Result

Train/Episode Reward



Evaluate/Episode Reward





```
episode 1 reward: 5480.0
episode 2 reward: 4400.0
episode 3 reward: 4030.0
episode 4 reward: 2080.0
episode 5 reward: 3840.0
average score: 3966.0
```

(2) Discussion

By running multiple environments in parallel, the agent can collect more experience at the same time. This allows for faster data collection and reduces the time required to gather enough training samples, leading to quicker policy updates and overall faster training.