Report of Collaboration and Competition Project

1 Problems

In this environment, two agents control rackets to bounce a ball over a net. If an agent hits the ball over the net, it receives a reward of +0.1. If an agent lets a ball hit the ground or hits the ball out of bounds, it receives a reward of -0.01. Thus, the goal of each agent is to keep the ball in play.

The observation space consists of 8 variables corresponding to the position and velocity of the ball and racket. Each agent receives its own, local observation. Two continuous actions are available, corresponding to movement toward (or away from) the net, and jumping.

The task is episodic, and in order to solve the environment, your agents must get an average score of +0.5 (over 100 consecutive episodes, after taking the maximum over both agents).

2 Learning Algorithm

2.1 Multi-agent Deep Deterministic Policy Gradient

Multi-agent Deep Deterministic Policy Gradient (Lowe *et al.*, 2017), also known as MADDPG, is employed to solve this problem. MADDPG is an extension of DDPG on a multi-agent setting. For each agent in MADDPG, it has one actor (and a target copy) for approximating the action value that maximize the Q function, and one critic (and a target copy) for estimating the Q function. During training, MADDPG employed a centralized strategy, that is each agent's critic have access to all other agents' information (for example, other's polices and observations, etc.). But the actor only expose to its own observations.

2.2 Neural Network Architecture

A three-layer fully connected neural network is employed to estimate the parameterized policy and state-value function at the same time. The first two hidden layers are with 200 and 150 neurons respectively. For actor, the input dimension is the state size, and output a vector with length action size. For critic, the input dimension is number of agents * (state size + action size), and output a single value.

2.3 Hyperparameters

- Discount factor: 0.99
- Learning rate for actor: 1e-4
- Learning rate for critic: 1e-3

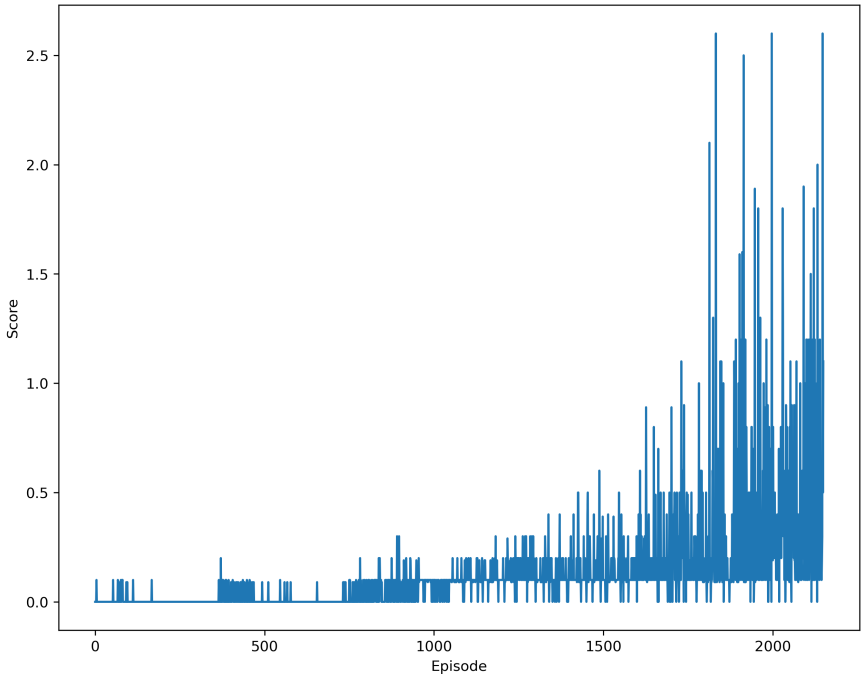

Figure 1: Score plot of MADDPG

- Buffer size: 1e5
- Batch size: 256

3 Results

The plot of maximum (over the 2 agents) accumulated scores per episode is shown in Figure 1. In total, 5000 episodes are run.

3.1 Summary

- MADDPG need 230 episodes to solve the problem.
- It takes abouts 2.5 hours for training (using CPU).
- As the training continues, the costed time for each episode is increasing in average. That is because agents are getting better and better, so the episode tends to be longer.
- Because the non-stationary property of multi-agent settings, algorithms are usually not that stable as for the conventional MDP setting.

4 Future Work

- Try other hyperparameters.

- Try PPO algorithm.
- Try other more recent proposed methods.

References

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pages 6379–6390.