Tennis Report

Caijiawen

March 2019

1 Learning Algorithm

1.1 MADDPG algorithm

We use MADDPG to solve the tennis project in unity environment.

MADDPG is multi-agents version of DDPG.

Every agent use its own state to train actor network , while using a complete state(contains own state and other agents' state) critic network to estimate value function.

The critic network's loss is:

$$Q_{crtiticLocal}(S_t, a_t) - (r_t + \gamma Q_{critecTarget}(S_{t+1}, a_{t+1}))$$

Actor network's loss is:

$$Q_{criticLocal}(S_t, a_{tPrediction})$$

the $a_{tPrediction}$ is predicted using actor local network.

1.2 Implementation Detail

1.2.1 Neural Network Models

We use 4 networks for every agent : actor local network , actor target network , critic local network , critic target network.

Local network and target network has same architecture , for actor network:

- (fc1): Linear(InFeatures=24, OutFeatures=512, bias=True)
- (fc2): Linear(InFeatures=512, OutFeatures=256, bias=True)
- (fc3): Linear(InFeatures=256, OutFeatures=1, bias=True)

And critic network is:

- (fc1): Linear(InFeatures=(24+2)*2, OutFeatures=512, bias=True)
- (fc2): Linear(InFeatures=512, OutFeatures=256, bias=True)
- (fc3): Linear(InFeatures=256, OutFeatures=1, bias=True)

1.2.2 MADDPG agent

Act In every round , we use actor network to predict action , we use predicted action plus a noise as the real acted action.

Learn In every round , we add sample to ReplayBuffer and use sample from ReplayBuffer to train the networks.

1.2.3 Hyperparameters

- buffer size : 100000
- batch size : 200
- discount rate: 0.99
- critic learning rate: 0.0003
- actor learning rate: 0.0001
- tau(soft update parameter): 0.001
- update every step: 2

2 Results

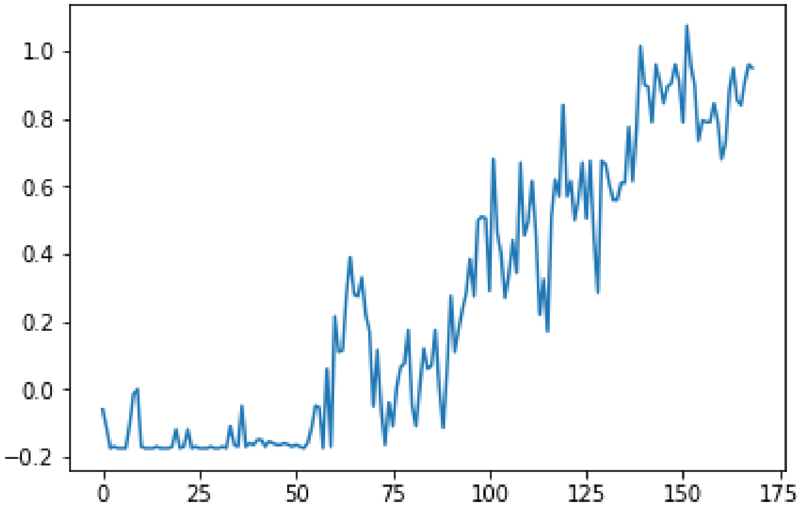

Figure 1: Score over Episodes

The environment is solved by our MADDPG agent in 164 episodes.

3 Conclusion and Future Work

In multiple agents' setting , we need to use both global information and agent self's information.

In our setting , MADDPG use a global critic network and shared replay buffer to make convergence happen.

There are other ways to attack this kind of problem , for example openai's dota2 robot , use PPO algorithm and a shared reward to master the game , in future we can try this method.