Deep Reinforcement Learning Project 2 : Continuous Control

Rachel Schlossman

September 29, 2021

1 Learning Algorithm

The learning algorithm is a modified implementation of the Deep Deterministic Policy Gradient (DDPG) algorithm [1] in order to be applicable to multi-agent reinforcement learning. It is important to note how this implementation is different from the multi-agent Deep Deterministic Policy Gradient (MADDPG) algorithm as described in [2]. As the referenced paper describes, the algorithm is based on (1) each agent having its own actor network to approximate the actor function (decentralized exection) and (2) a centralized critic network with access to all agents' observations and actions (centralized learning). In contrast to this, the algorithm in this project leverages a single actor network and a single critic network (centralized execution and learning). The two networks are shared between the two agents; The actor network only has access to a single agent's observations and, similarly, the critic only has access to a single agent's observations and actions. So, at a single timestep, the same network parameters are used to devise both agents' policies and action-value functions. This strategy is appropriate for the Tennis environment because the environment is symmetric; the agents have the same observations, actions, and reward function; and neither agent is rewarded directly for beating the other agent.

1.1 Hyperparameters

The following hyperparameters were used in the implementation:

- replay buffer size : 1e6
- minibatch size: 128
- discount factor: 0.99
- for soft update of target parameters, τ: 1e-3
- learning rate of the actor: 1e-4
- learning rate of the critic: 1e-4
- L2 weight decay: 0

2 Model Architecture

The modified DDPG algorithm employs two neural networks, the actor network and the critic network. For both networks, the input passes through two linear layers with relu activation. For the actor network the input is the observation (length 24) and the hidden layers are followed by a tanh output layer. For the critic network, the input is also the observation, and in the first hidden layer the (length 2) action is appended to the input layer's outputs. The hidden layers are followed by a linear output layer. The two hidden layers for each network each are comprised of 256 nodes (excluding the concatenation of the actions).

3 Results

The two agents are able to receive an average reward (over 100 episodes, and over all both agents) of at least 0.5 in 2406 episodes, as shown in Figure 1.

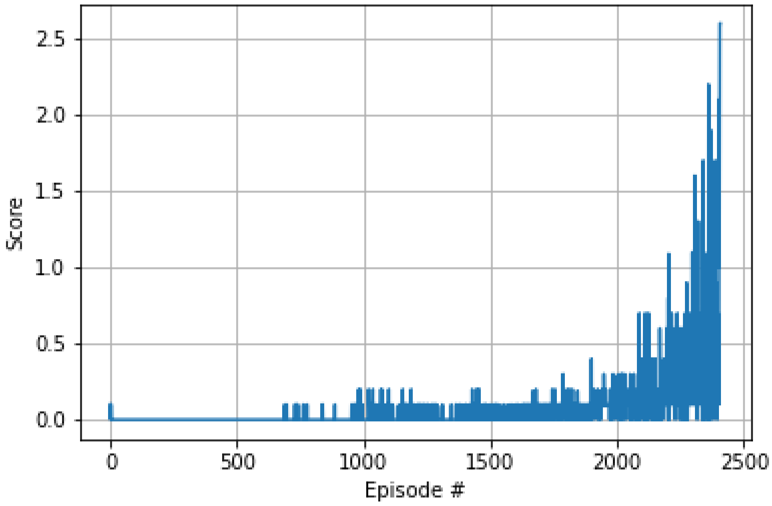

Figure 1: Average Score Plot

4 Future Work

In the future it would be of interest (1) to explore learning behavior via a prioritized replay buffer and (2) to implement decentralized policy execution via two separate action networks, one per agent.

References

[1] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[2] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.