Project Report: Collaboration and Competition

A DDPG model was used to solve the problem. Two agents were created, each with an actor and critic pair, acting independently, but with a shared experience buffer. The actors are a feedforward neural network consisting of 2 hidden layers each with 128 neurons each. The critic networks also followed the same architecture.

While the output of the actor model was 2 (the action size), the output of the critic network is 1.

One network, the actor, learns the optimal deterministic policy. While the other network, the critic, evaluates the optimal action value function by using the actor's prediction.

Both networks had a learning rate of 1e-3.

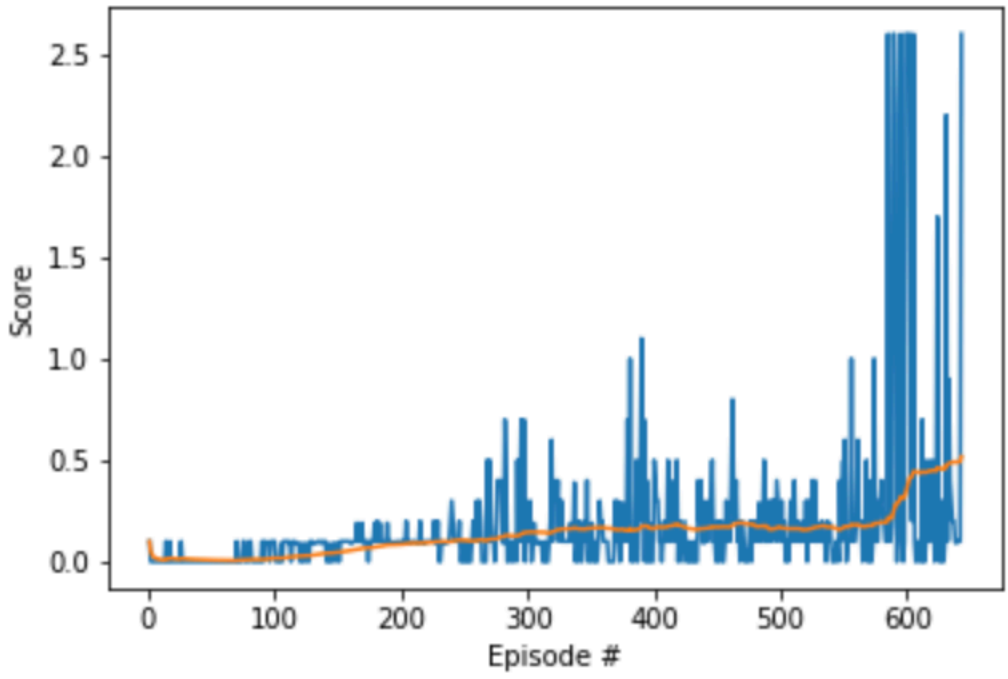
Other parameters are below:

BUFFER_SIZE = 100000; This is how much experience should be stored.

BATCH_SIZE = 128; How much data points to train the network with.

TAU = 6e-2; soft update parameter
GAMMA = 0.99; This determines the importance of future rewards. With numbers approaching 1 favouring long-term rewards, while those close to 0 favours current reward.

As depicted below, the environment was solved in 543 episodes!



Conclusion and Future Work

PPO achieves more stability on control tasks, so it'll be interesting to see how it performs in this environment. D4PG, a technique by DeepMind could also be tested on this environment. Also, bigger and deeper networks could be tried to see if the agent learns faster.

While the environment was solved, learning diverges after a while, showing instability. This behaviour needs to be investigated.