Project Report: Continuous Control

A DDPG model was used to solve the problem. The actor was a feedforward neural network consisting of 2 hidden layers each with 128 neurons each. The critic network also followed the same architecture.

While the output of the actor model was 4 (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.

Actor and critic networks were updated 10 times every 20 timesteps. Both networks had a learning rate of 1e-3.

Other parameters are below:

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

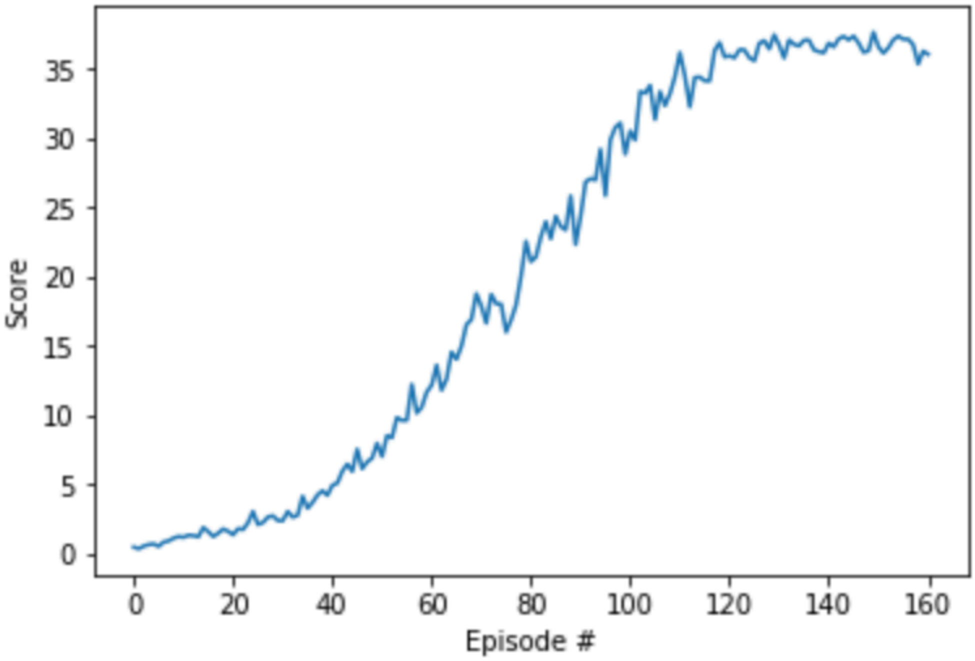
BATCH_SIZE = 64; How much data points to train the network with.

TAU = 1e-3; 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.

Rewards

Using the hyperparameters above, the agent solved the problem in 61 episodes with a mean score over 100 episodes of 30.07. Below is a plot of the rewards.



Ideas for Future Work

PPO achieves more stability on control tasks, so it'll be interesting to see how it performs in this environment.

Also, bigger and deeper networks could be tried to see if the agent learns faster.