Udacity Deep Reinforcement Learning Nano Degree
Project 2 Continuous Control

December 18, 2018

1 Introduction

Continuous Control is the second project for the Udacity Deep Reinforcement Learning specialisation. In this project students are required to develop an agent that is capable of manoeuvring a double-jointed arm so as to maintain its position at a target location for as many time steps as possible. In this regard, the agent is given a reward of +0.1 for each time step that it maintains its hand in the target location. The observation space consists of 33 variables corresponding to position, rotation, velocity, and angular velocities of the arm. Each action is a vector with four numbers, corresponding to torque applicable to two joints. Every entry in the action vector should be a number between -1 and 1.

Note that for the purpose of this submission, the second version of the Reacher environment is to be solved. For this environment it is specified that the agents must get an average score of +30 (over 100 consecutive episodes, and over all agents). Specifically,

After each episode, we add up the rewards that each agent received (without discounting), to get a score for each agent. This yields 20 (potentially different) scores. We then take the average of these 20 scores. This yields an average score for each episode (where the average is over all 20 agents). The environment is considered solved, when the average (over 100 episodes) of those average scores is at least +30.

2 Learning Algorithm

In this submission an implementation of a DDPG is included. The hyperparameters used are as follow: the maximum number of training episodes is 200, the seed to initialise the pseudo random number generator is 0, the maximum size of buffer is 1e5, the size of each training batch is 128, the discount factor is

0.99, the interpolation parameter for soft updating is 1e-3, the learning rate for the actor and the critic is 1e-4 and the weight decay parameter is set to 0.

The network architecture for both the agent and the critic consists of a single neural network with two hidden layers consisting of 256 neurons in the first hidden layer and 128 neurons in the final hidden layer. Each layers uses ReLu activation.

2.1 Training Results

The results obtained during training of the agent are

Episode: 115 Score: 36.34 Average Score: 30.31
Environment solved in 115 episodes! Average Score: 30.31

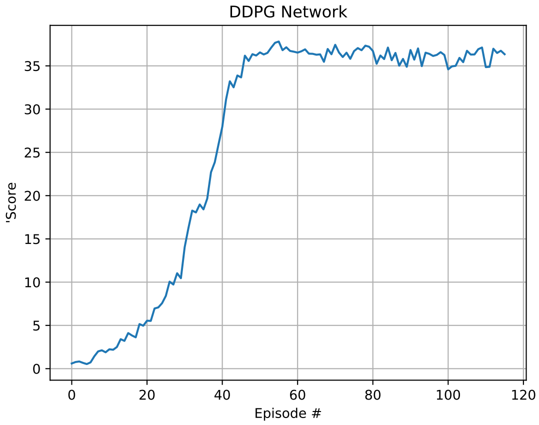

Figure 1: Training performance of Double Deep Q Network

3 Ideas for Future Work

In the future I would like to experiment with Proximal Policy Optimisation (PPO) with Generalised Advantage Estimation (GAE) as this was not investigated.