Project: Continuous Control

1) Goal:

In this environment, a double-jointed arm can move to target locations. A reward of +0.1 is provided for each step that the agent's hand is in the goal location. Thus, the goal of your agent is to maintain its position at the target location for as many time steps as possible. In this project, we train a single agent to perform this task. A snapshot of the environment is as follows:

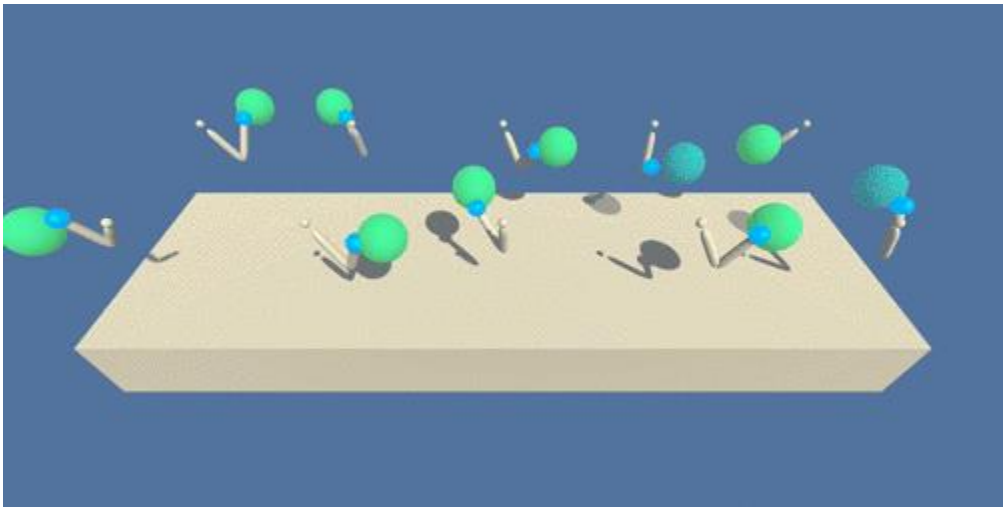


Figure 1 Reacher Environment.

2) Algorithm:

We use Deep Deterministic Policy Gradient (DDPG) algorithm to solve this environment. DDPG is an actor-critic algorithm where a critic estimates the state value function via Temporal difference estimate. The architecture for actor and critic is as shown in the following figure

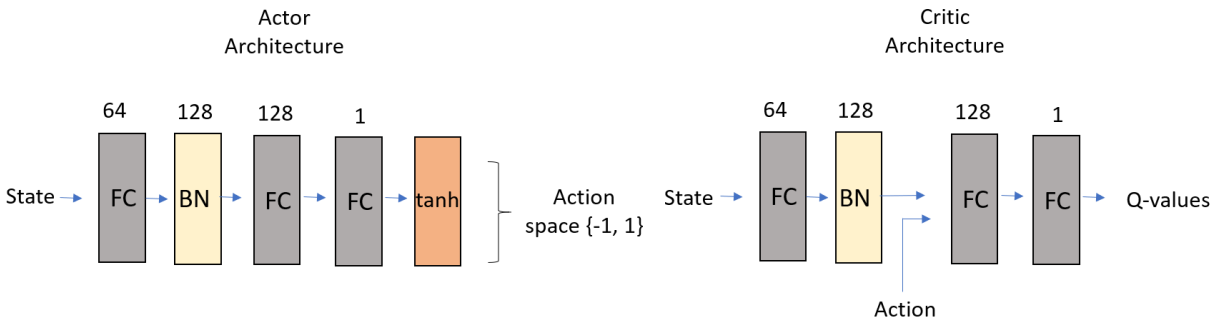


Figure 2 Model Architecture.

Hyperparameters chosen for this experiment:

BUFFER_SIZE = int(1e5) # replay buffer size

BATCH_SIZE = 128 # minibatch size

GAMMA = 0.99 # discount factor

TAU = 1e-3 # for soft update of target parameters

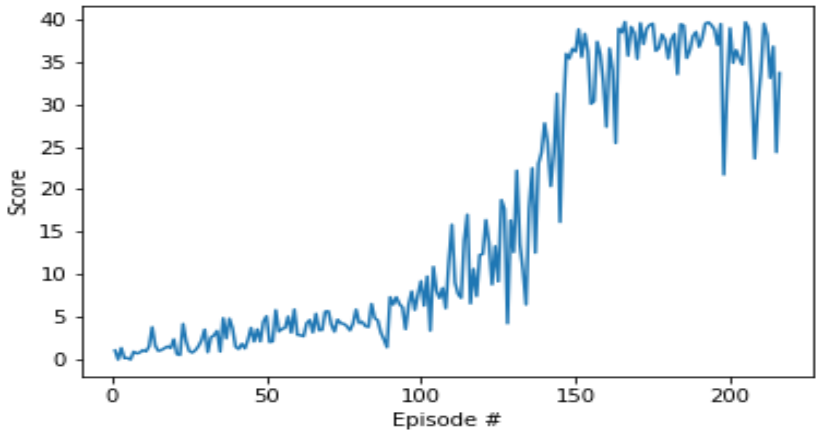
LR_ACTOR = 1e-3 # learning rate of the actor

LR_CRITIC = 2e-4 # learning rate of the critic

WEIGHT_DECAY = 0 # L2 weight decay

3) Results:

The rewards plot is shown in Figure 3.



The environment was solved in 216 episodes with the average reward of 30.08. The convergence of the model was poor without batch normalization. The average scores were not improving beyond 17.

4) Future work:

a) Analyze the effects of architecture of actor and critic on performance.

b) As mentioned in the project instructions, I would like to implement TRPO, TNPG, D4PG methods and compare all of them.