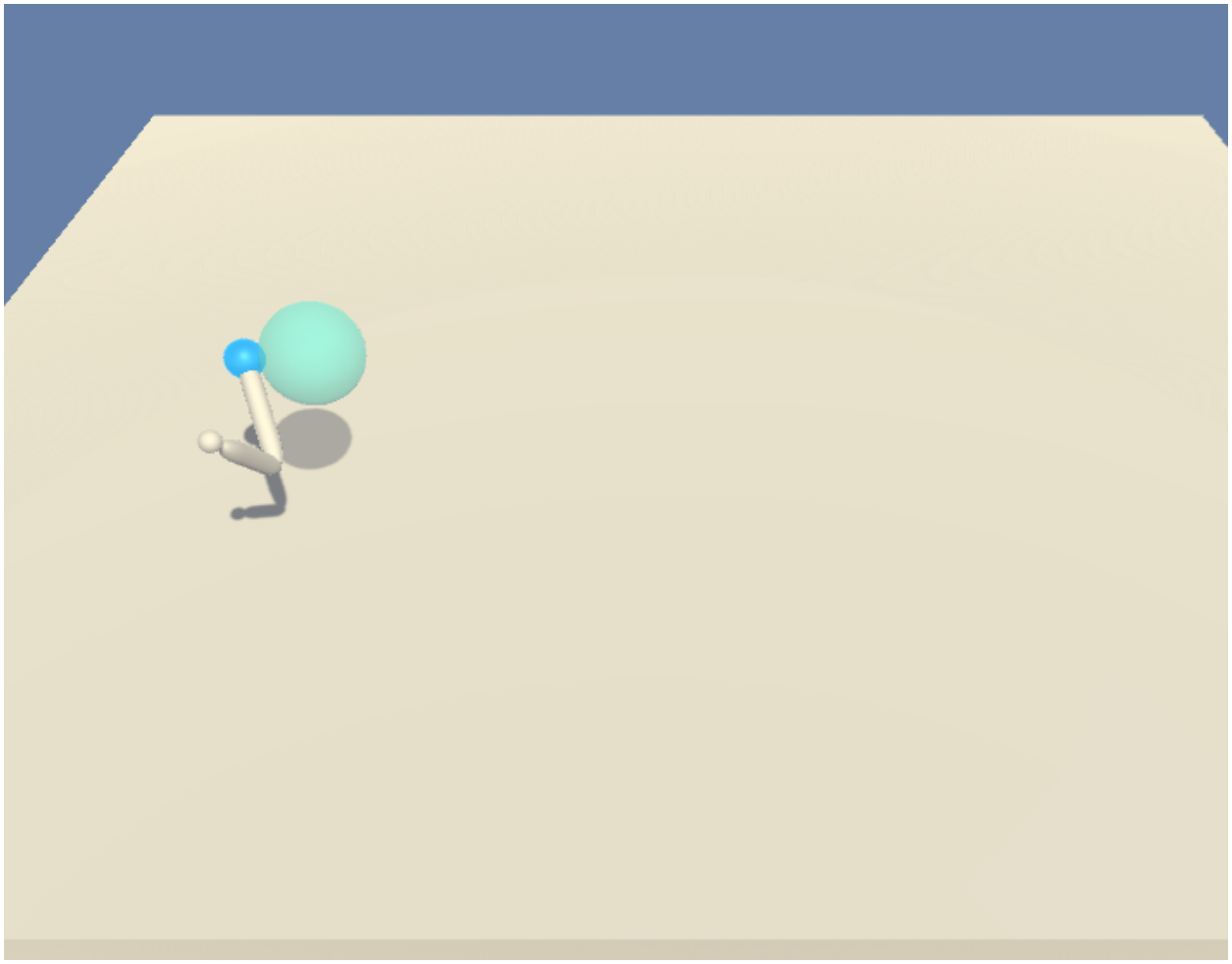
Project 2: Continuous Control Report



I: Project Summary

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.

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.

II: Learning Algorithm

We use a Deep Deterministic Policy Gradient (DDPG) Actor-Critic algorithm to teach an agent how to interact with the environment to learn how to achieve its goal of maintaining its position at the target location for as many time steps as possible.
The main idea and details behind DDPG can be found in this paper here, but in general, in DDPG, we have to train two neural networks, one for the actor, and another for the critic.

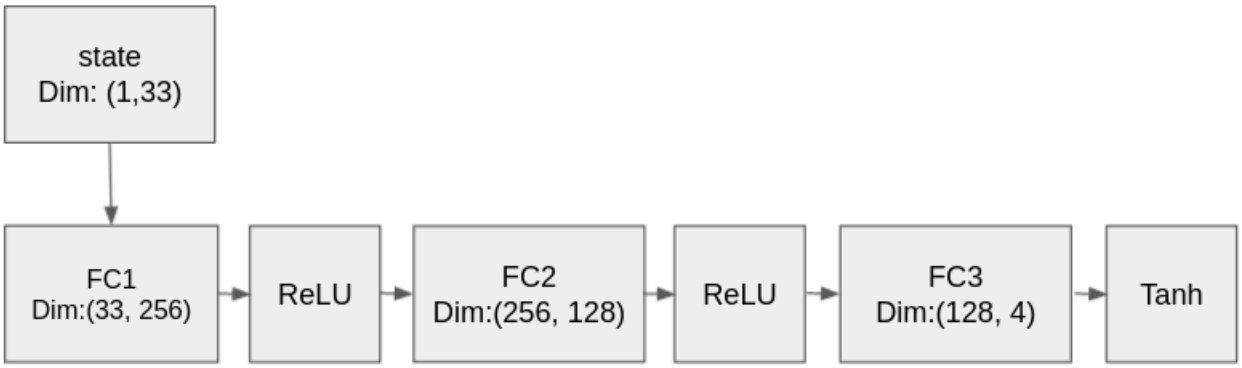
In addition to the vanilla DDPG implementation, the experience relay technique is used to learn from individual experience tuples multiple times, recall rare occurrences, and make better use of our experience. For experience relay, a replay buffer of size 100,000 is used.

The table below shows the hyperparameters used to train the model:

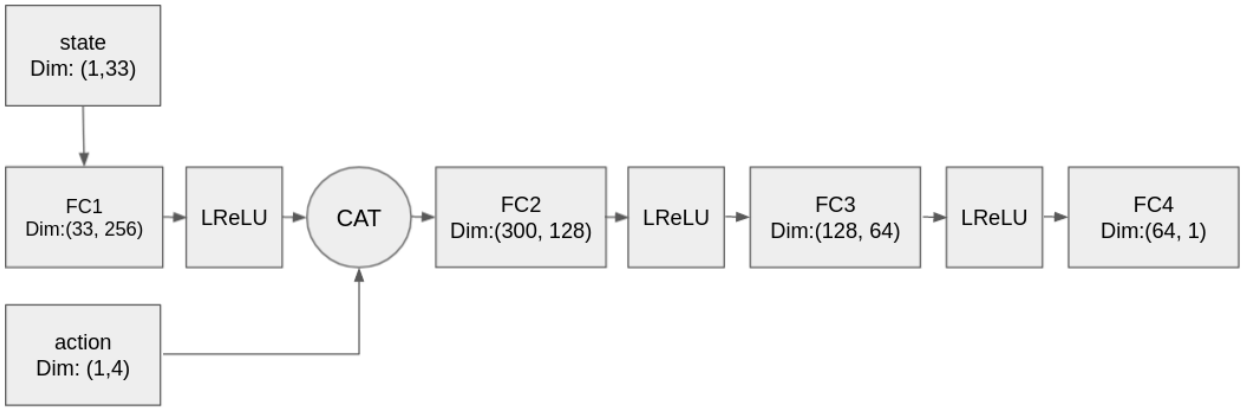
BUFFER_SIZE (replay buffer size)	100,000
BATCH_SIZE (minibatch size)	128
GAMMA (discount factor)	0.99
TAU (for soft update of target parameters)	1e-3
Actor LR (learning rate)	1e-4
Critic LR (learning rate)	1e-4
UPDATE_EVERY (how often to update the network)	10
Noise Epsilon Initial Value	0.9
Noise Epsilon Decrement Value	2e-6

The figure below shows the deep neural network architecture used for the actor and critic networks, where:
FC: Fully Connected Layer with its dimensions
ReLU: Rectified-Linear Unit
LReLU: Leaky Rectified-Linear Unit

A: Actor Deep Neural Network



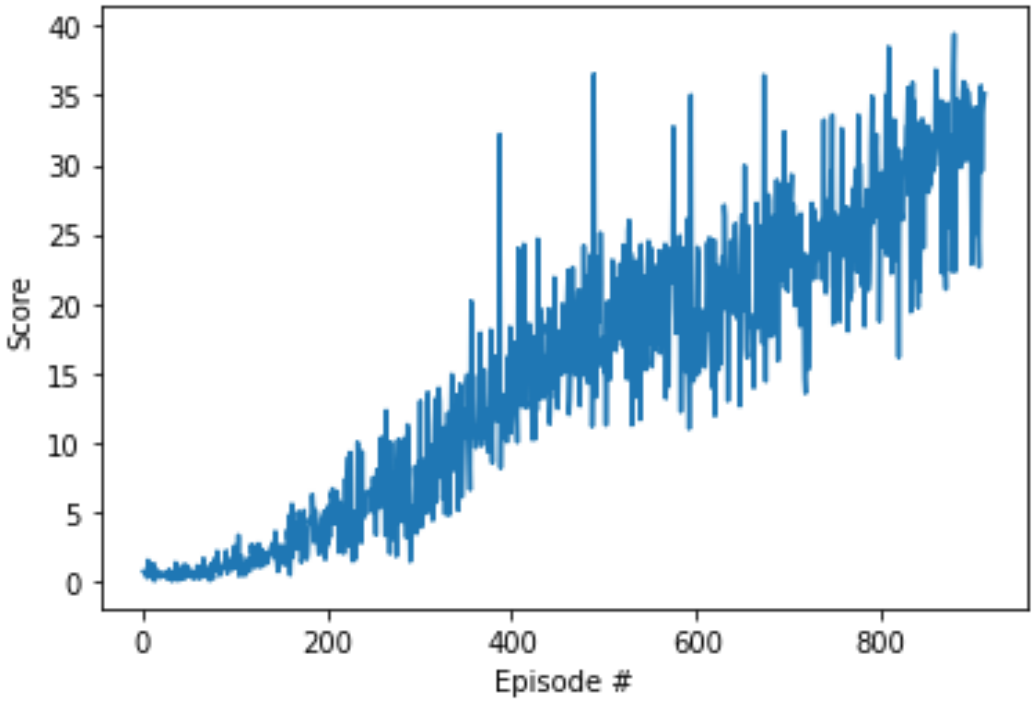
B: Critic Deep Neural Network



III: Rewards Plot

A: Training Plot

The plot below shows the rewards per episode that the agent received an average reward over training for 912 episodes to solve the environment.



IV: Future Work

Although the agent is able to maintain its position at the target location for a significant number of time steps, the agent fails to follow the target when the target is at certain locations, or when transitioning from one state to another.
It would also be interesting to try other actor-critic methods such as A2C or A3C to see how they affect the agent's performance.