Continuous Control Project Report

Learning Algorithm

This model used a DDPG algorithm with the following parameters:

Replay buffer size: int(1e6) # replay buffer size
Minibatch size: 256
Discount factor: 0.9
Tau for soft update of target parameters: 1e-3
Learning rate of the actor: 1e-3
Learning rate of the critic: 1e-3
L2 weight decay: 0

I used DDPG, a type of actor-critic algorithm, to solve this environment. In DDPG, both the actor and the critic have a target and a local network (for a total of 4 networks). The actor approximates the optimal policy deterministically by using the argmax of Q(s,a). The critic evaluates the policy and is updated using the TD-error, and the actor is trained using the deterministic policy gradient algorithm.

DDPG uses a replay buffer to prevent oscillation and divergence in the action values. The state (S), actions (A), rewards (R), and next states (S') are stored as tuples and small batches of the tuples are sampled to learn from. This is called experience replay and it helps break harmful correlations, plus the model can learn better from rare occurrences.

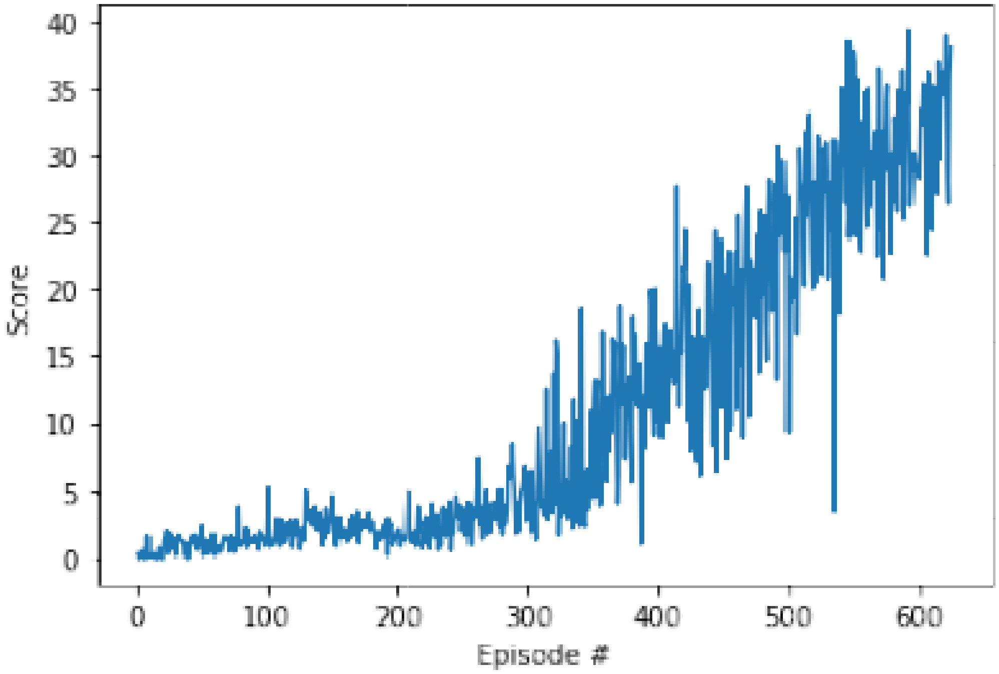
DDPG also uses soft updates to update the target networks. This means that target network is slowly blended with the regular network at every timestep. For instance, the target network will take 99.99% of the target network weights and only 0.01% of the regular network weights. This practice leads to faster convergence.

DDPG utilizes Ornstein-Uhlenbeck (O-U) noise for exploration. The algorithm probabilistically selects a random action like Boltzmann exploration or epsilon-greedy in order to ensure that it doesn't get stuck in the rut of just sticking with it's previous experiences.

The actor and critic each contained 2 hidden layers in each network, and I used batch normalization in the critic.

Plot of Rewards

```
Episode 100	Average Score: 1.00
Episode 200	Average Score: 2.17
Episode 300	Average Score: 3.03
Episode 400	Average Score: 8.95
Episode 500	Average Score: 18.12
Episode 600	Average Score: 28.28
Episode 624	Average Score: 30.07
Environment solved in 624 episodes!	Average Score: 30.07
```



Ideas for Future Work

I would be interested in comparing the performance of my current model that uses the DDPG algorithm to other models that use other algorithms, such as TRPO, TNPG, PPO or D4PG.