Learning Algorithm

The algorithm used was DDPG due to the fact that the action space in this case is continuous. Note that additionally noise in the form of a normal distribution was added to each action value in order to aid with exploration

Note that although the project description suggested that we adapt the following code bases to this problem
https://github.com/ShangtongZhang/DeepRL/blob/master/deep_rl/agent/DDPG_agent.py and
https://github.com/udacity/deep-reinforcement-learning/tree/master/ddpg-pendulum

I was instead focused on CleanRL's implementation, with attributions in the code where necessary https://github.com/vwxyzjn/cleanrl/blob/master/cleanrl/ddpg_continuous_action.py

Hyperparameters

The environment was developed using the 20 agent version.

The hyperparameters used are listed below

```
buffer_size = int(1e6)
gamma = 0.99
tao = 0.005
max_grad_norm = 0.5
batch_size = 256
exploration_noise = 0.1
learning_starts = 5e3
policy_frequency = 5
noise_clip = 0.5
learning_rate = 3e-4
```

Note in particular that the learning only starts after a "warm up" period of 5000 steps.

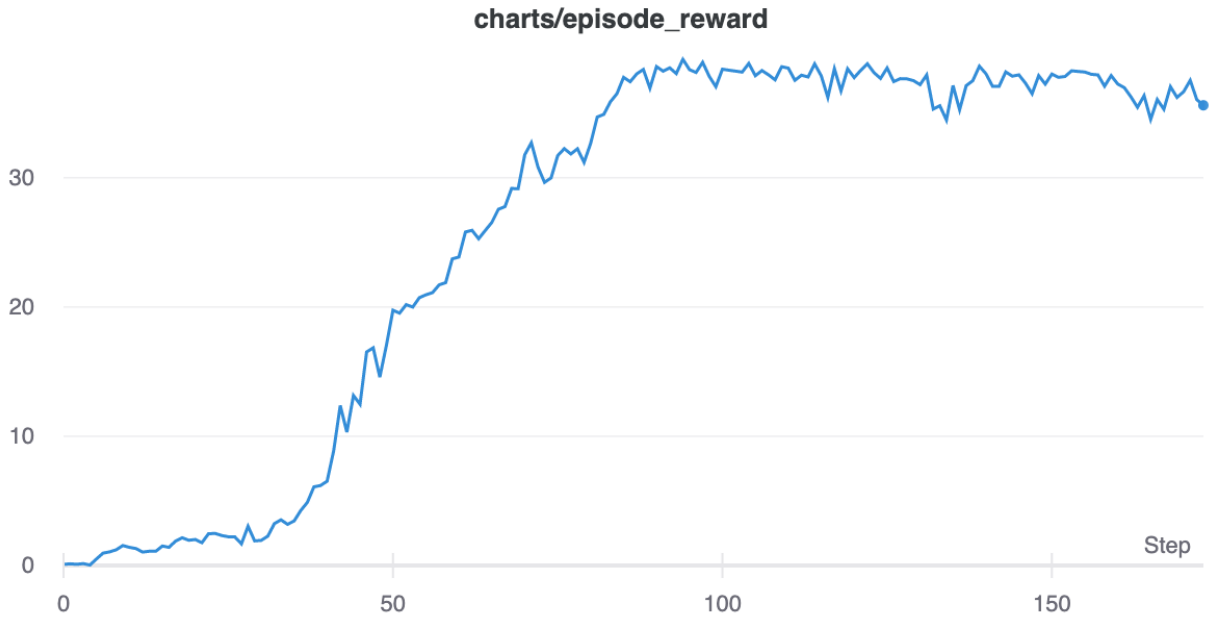
Model Architectures

There were two neural networks involved here:

1. The Q network. This was used to estimate the value of each taken action. This took as input into the first layer both the current action and the current state. This was followed by two consecutive hidden layers of 256 ReLU neurons each, followed by a linear output neuron.

2. The Actor network. This was used to produce the actions to be taken, given the state at each point. This had two consecutive layers of 256 ReLU neurons each, with an output layer producing one tanh-capped action value for each action in the space.

Results



The above chart denotes the average reward per episode. As you can see, the environment is solved by episode 150, producing an average reward of above 30.

Future Work

The efficiency of training could be increased by implementing Self Adaptive Double Bootstrapped DDPG, for example https://www.ijcai.org/Proceedings/2018/0444.pdf
This may be increase the speed with which the environment is solved.