

Project Report: Deep Reinforcement Learning Nanodegree - Project 2: Continuous Control

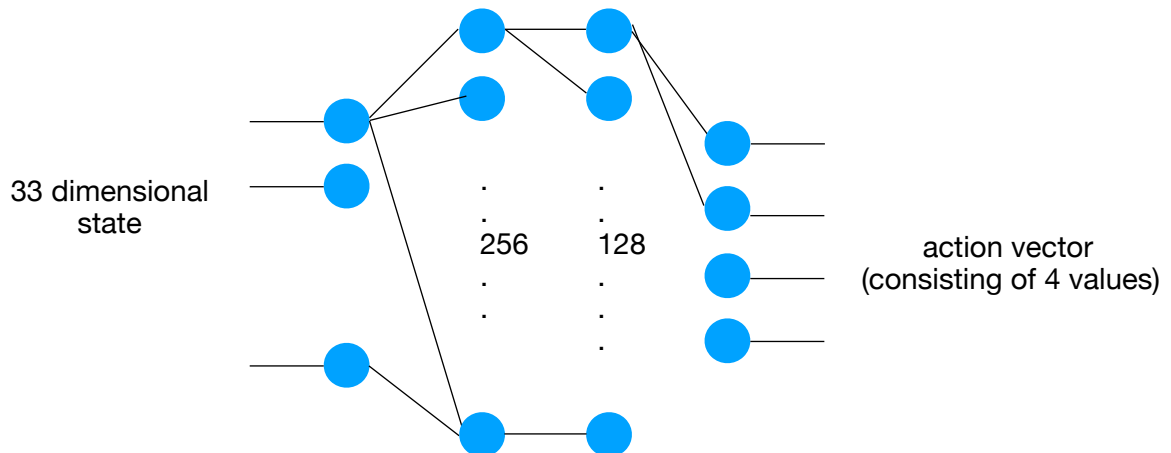
Learning Algorithm

In the project we train a Deep Reinforcement Learning agent based on the Deep Deterministic Policy Gradient (DDPG) approach. The task in hand requires a model that can generate continuous action values.

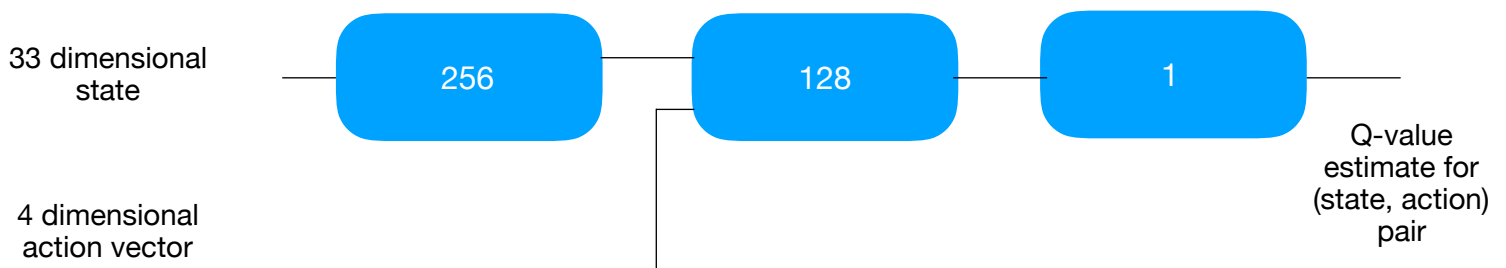
The DDPG model consists of 2 parts: an Actor and a Critic

- Actor: The actor takes the current state as input and generates a deterministic action corresponding to this state.
- Critic: The critic takes in both the current state and the action chosen by the actor. The critic's job is to estimate the Q-value corresponding to this (state, action) pair.

Actor network



Critic network

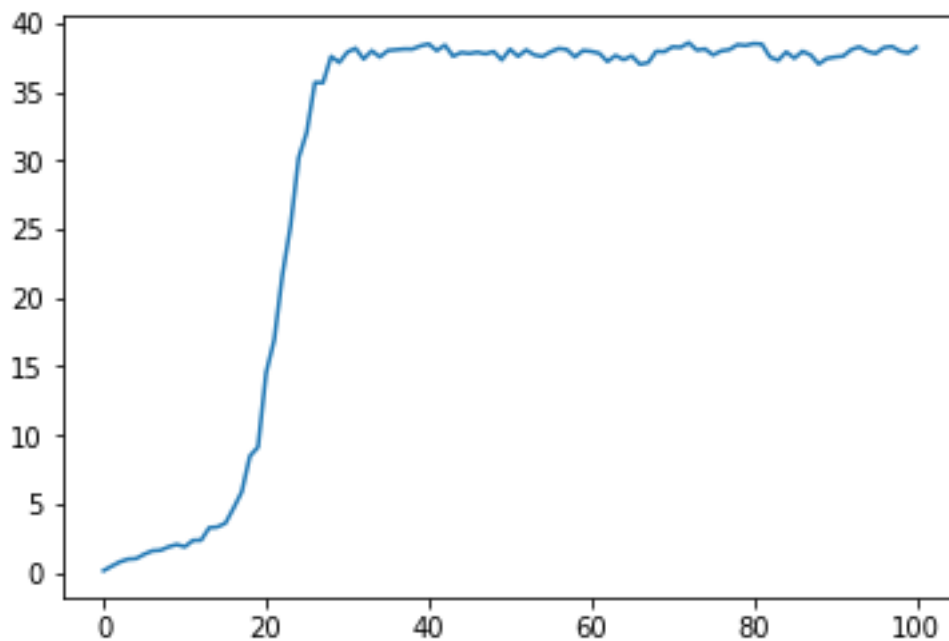


Experience Replay

An experience is defined as a tuple of (state, action, reward, next_state, done). As we progress through training, these tuples are stored in a deque of buffer size 1e6. Batches of size 128 are drawn from this deque, and used to train the DDPG model.

Rewards

The below plot shows the rewards obtained by the agent as training progressed. On the x-axis is the number of episodes. The y axis shows scores averaged over the last 30 episodes.



Ideas for Future Work

Convergence with DDPG can be somewhat unstable. Duan et al (Benchmarking Deep Reinforcement Learning for Continuous Control, <https://arxiv.org/pdf/1604.06778.pdf>), observed that approaches such as TNPG (Truncated Natural Policy Gradient), TRPO (Trust Region Policy Optimization) performed much better. This can be an avenue for further exploration.