The neural network that I have used as part of the deep q-learning task consists of 4 inputs (provided by the OpenAI gym environment and 2 hidden layers (the first consisting of 32 neurons and the next consisting of 64 neurons). Finally, the network outputs 2 floats representing the values of the actions left and right. I first used networks of 16,16 and 32,32 and they did converge much more quickly however they were more biased so they did not score as well in the average case for their best model. I did also attempt a model with hidden layers 64,64 but it took much too long to converge.

To train the network I have implemented experience replay as it aids with faster network convergence and a target network to prevent “catastrophic forgetting” where an ai will get significantly worse at times whilst training (causing slower convergence) (Larsen, 2019). This network now converges after 5000 rounds to an average score of around 500 steps.

Also, whilst training the network I have made use of a variable exploration rate so that as the agent trains more it will become less random. I originally attempted to make the exploration rate dependent on the performance of the agent but this lead to cases where randomness caused bad scores encouraging more randomness.

References:

Larsen, N. (2019). *Why is a target network required?*. Stack Overflow. Retrieved 14 May 2021, from https://stackoverflow.com/questions/54237327/why-is-a-target-network-required.