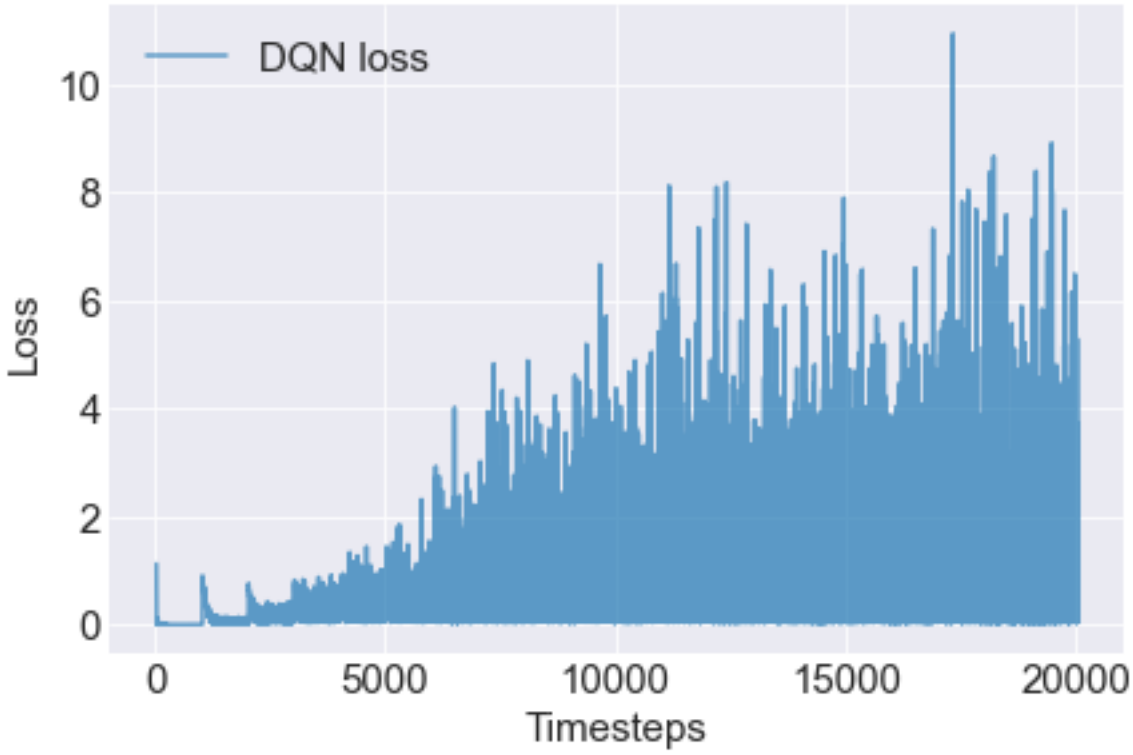

i) The loss is not decreasing steadily as in the case in supervised learning because in DQN, every time the value function is updated, the target values change. This makes estimating the weights of the networks that reduce the loss difficult as it is chasing after a moving target. This is unlike the supervised learning case where the labels and data do not change over time, so with multiple iterations of running the learning algorithms, they are able to fit suitable weights that reduces the loss with each iteration.

ii) The spikes are likely due to the mini-batch gradient descent in the Adam optimiser where the data from one mini-batch could result in higher loss than another by chance.