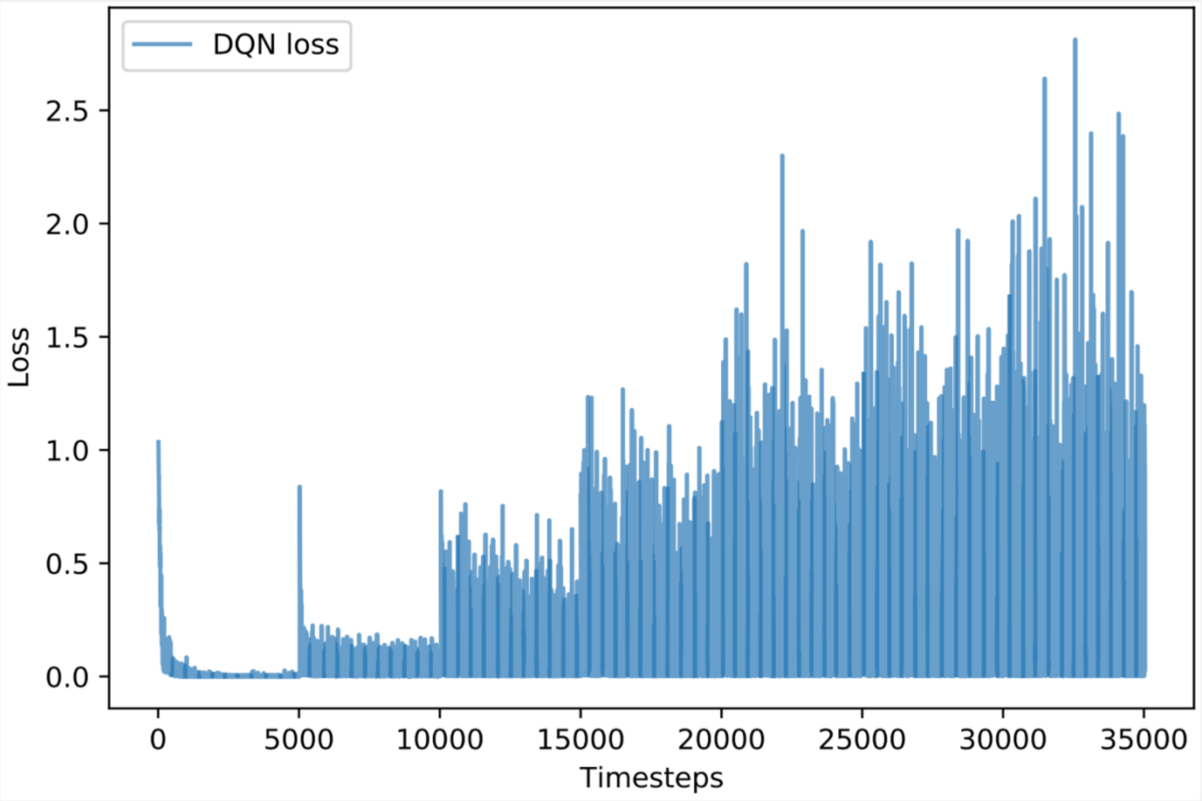
DQN Loss Explanation – Stephen Gallagher



With typical supervised learning approaches, the model is modified to decrease the loss against the training data. In the case of DQN, the algorithm is optimising the loss against the target network. At each 5000th time step, the parameters of the target network are changed and then we optimise the loss against a new target network, causing a spike. After each spike, the loss descends up until the next spike, as DQN uses an epsilon-greedy method of exploration, so the critic and target choose different actions. The loss continues to increase as because with CartPole, episodes get longer as the model improves (because the model will take a longer time to fail), therefore the variance in the losses will be larger as the model is constantly improving (which is why there is higher variance in the loss at later timesteps).