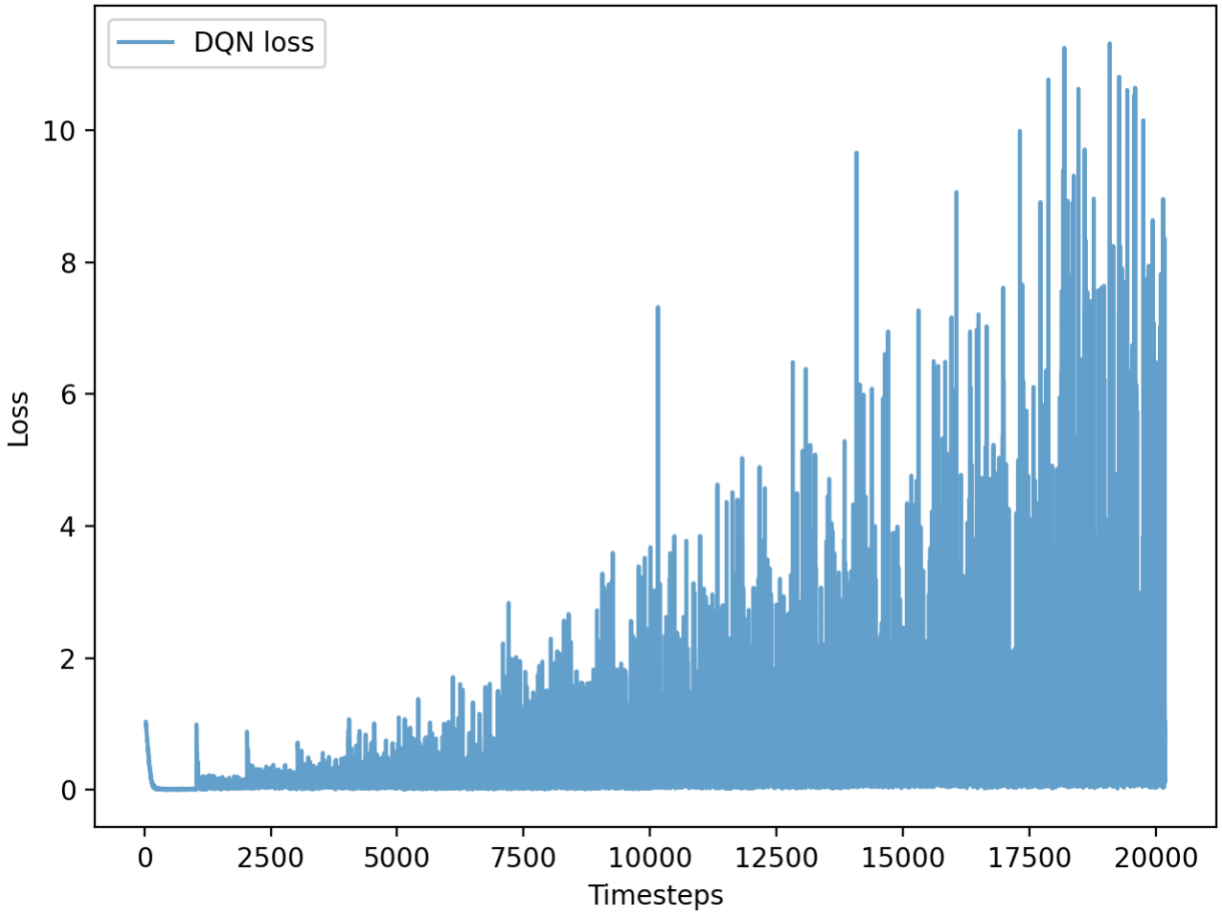

The plot shows that the loss periodically decreases and increases There are few reasons for it:

1. In the supervise learning, the data remains same during the training, so the model could be fitted to the data better after each epoch of training. In the DQN training, a batch of experience is sampled from the replay buffer to update the parameters in the network to minimise the loss. However, the data is keep changing (replay buffer) as the policy updates, which makes it different from the typical supervised learning.
2. As target network periodically updates, the parameters in the critic network that minimised the loss function is not minimising the loss function of new target network. Thus, the loss increases, resulting in the spikes. The increasing of magnitude of the loss could be due to the larger variance between loss functions of old and new target network.