Reinforcement Learning 2021/2022 Coursework

Xiao Heng (s2032451)

March 28, 2022

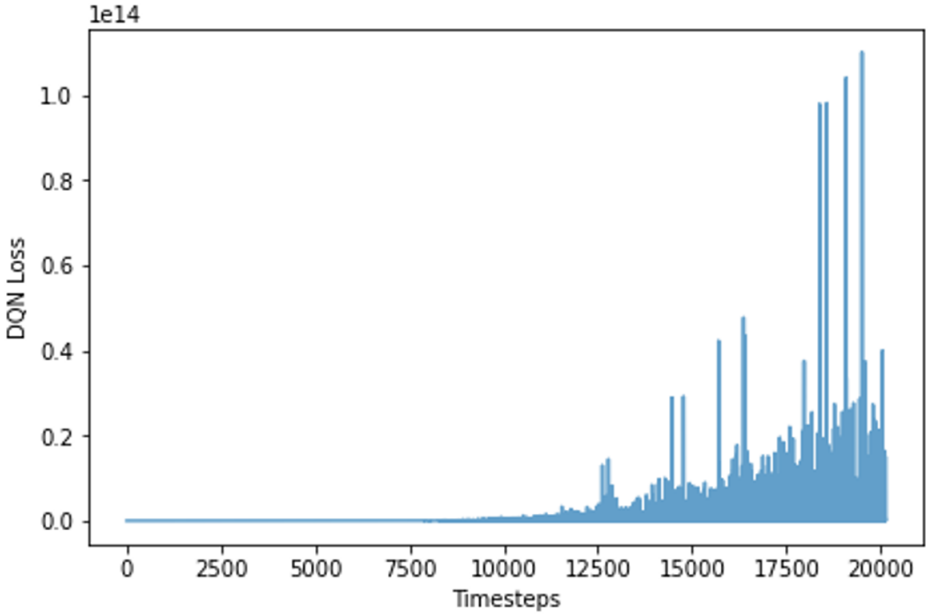

Figure 1: DQN loss during training within a single run of CartPole with the x-axis and y-axis corresponding to "timesteps trained" and the DQN loss. Associated hyper-parameter settings see table 1

Increasing loss

Along the training, the performance grows better, and the corresponding reward (as well as value of Q) increases from 0 or small random values, leading more difference. Also, with the better "balance strategy" in the CartPole game, the single episode grows longer, causing larger variance. Another reason is the drop in the buffer, that after training well, the experience becomes biased and skewed.

Explaining spikes

Here the spikes are more close, compared to the given figure in assignment, due to a smaller hyper-parameter of update frequency in my code. So the reason is clear now, the spikes appear when updating the target network every C steps. After each updating, new target network replace the old, which has been learnt well by value network, so loss increasing dramatically, and then gradually decreases.

Appendix

hyper-parameters	value
target_update_freq	10
batch_size	16
gamma	0.999
epsilon	$1 \rightarrow 0$

Table 1: hyper-parameter settings