Deep Q learning and Double Deep Q Learning

Suchet Aggarwal

2018105

In Deep Q learning, the target is estimated using the target network, with the action selection being greedy with respect to the function target function itself. This action selection often picks the suboptimal action due to the overestimation bias in the target network and leads to overall sub optimal policies.

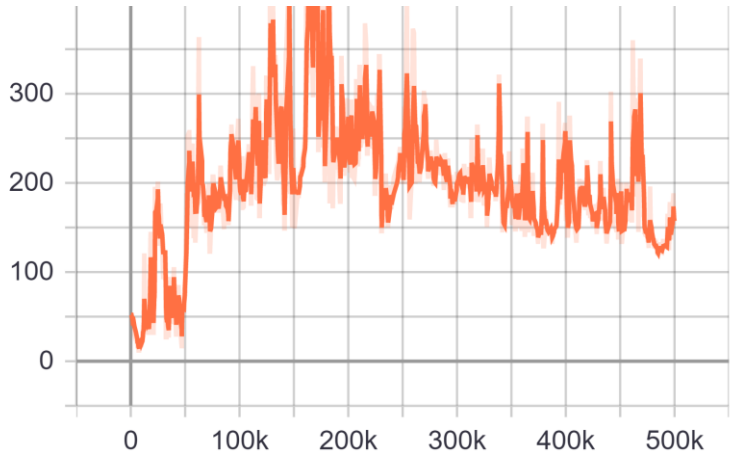


Fig 1 : Cumulative reward vs Iteration for DQN

Changes in the Deep Q learning implementation to incorporate Double Deep Q Learning:

1. The idea is decouple the action selection and action-value evaluation, which in case of DQN is in the form of a max operation over all actions in the target network.
2. In the standard Double Q learning approach we maintain 2 Q functions and randomly use one for picking the greedy action and the other for evaluation for that action for evaluating the target value. Here We already have two networks namely: qNetwork and TargetNetwork. We use the qNetwork for picking the greedy action with its respect and use this to estimate the value from targetNetwork. We thus use the online version of the network for picking the action, thus eliminating the case of picking the suboptimal action only because it had a higher offline q value (i.e. value from TargetNetwork), while the update of the weights of the target network still remains a periodic function

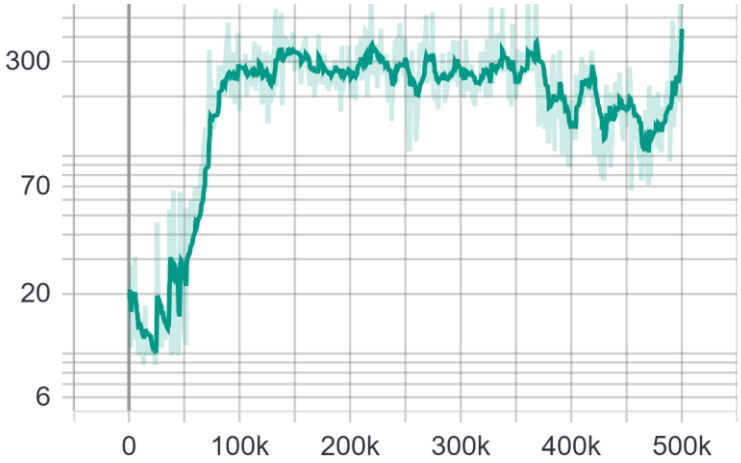
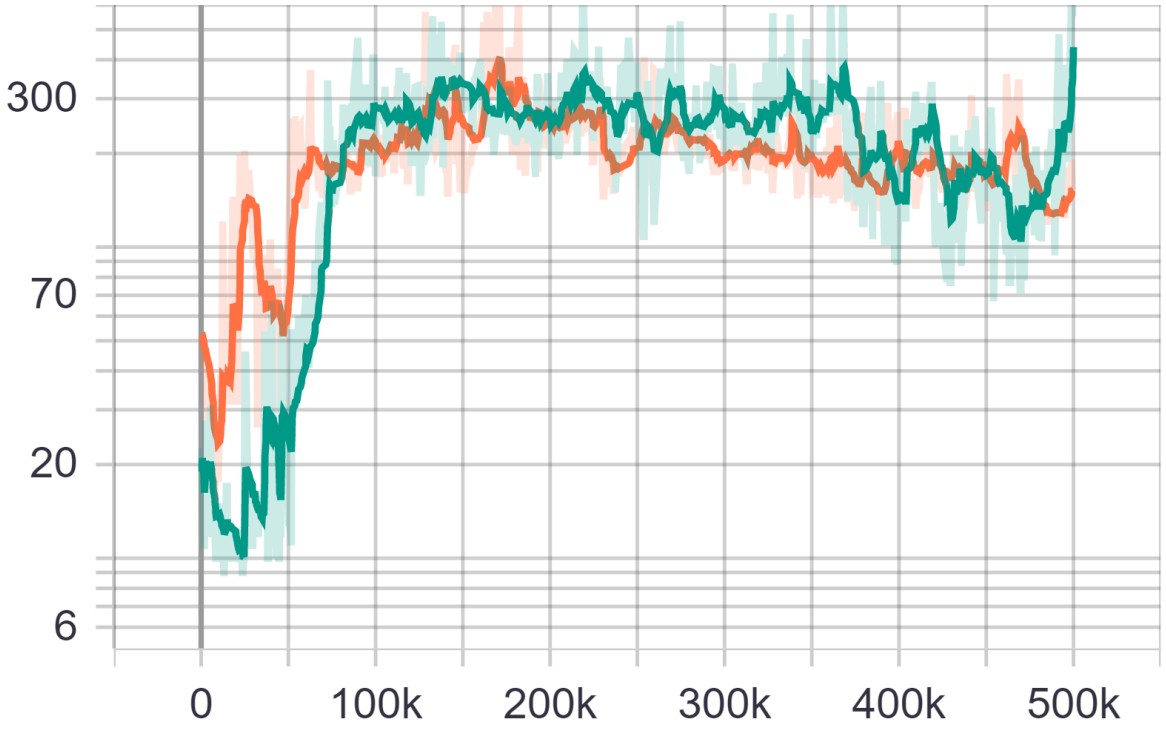


Fig 2: Cumulative reward vs Iteration for DDQN

The overall comparison between the two is shown as below



Cumulative reward vs Iteration

1. The DQN agent is a bit unstable and generally has a lower overall reward mainly because it picks sub optimal actions due to the overestimation bias, while the DDQN agent is better able to estimate the Q values leading to an overall higher cumulative reward.
2. The DDQN has higher spikes over the iterations signifying it is better able to achieve higher rewards.
3. DDQN learns faster than DQN, although with the initial reward being lower for DDQN, the reward thereafter is higher and converges quickly and to a higher value than DQN.