Bowei Zhu z5059172, h5059172

COMP9444 Neural Nets Assignment 3 Cart Pole

Basic Implementation:

- Took around 1000 episodes before it reached 200 reward

Batching and Experience Replay

- Performance improved substantially, by episode count 50, it was able to reach an average reward of 200.
- However there were abrupt drops due to instability in the way training was done.
- This is likely because the network we're using to calculate the Q values has not been fully trained on extensively for most
- An improvement on the reward heuristic was also applied, this had some noticeable effect.
    - Rather that using a reward of 1 or 0:
        - A combination of the current theta location and position of the cartpole was used to calculate the current reward value.

Double Deep Q Learning

implemented to reduce the likelihood of over estimations by using a slightly order version of the ne which could be changed by a parameter based on episode count.

A copy of the network was made to evaluate Q values, it was not trainable but its weight was periodically updated (each iteration of episode) to remove some temporal associations.

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)\,.$$

Dueling DQN Network

To improve stability, the network was altered to include an advantage and value function when added together resulted in greater stability. The trade off it appeared was that the number of episodes needed to reach an average of 200 reward had increased. This appeared not to have a significant effect

$$Q(s,a) = A(s,a) + V(s)$$

Originally a batch size of 256 was used, but it was noticed that the network wouldn't always get to 200 average reward by 100 episodes. So it was increased slightly to ensure that episode 100 would reach ~200 average reward.

Results
episode: 10 epsilon: 0.40290221441642243 Evaluation Average Reward: 9.2
episode: 20 epsilon: 0.36437754069956635 Evaluation Average Reward: 9.8
episode: 30 epsilon: 0.32953651634447895 Evaluation Average Reward: 9.5
episode: 40 epsilon: 0.2980269184427927 Evaluation Average Reward: 45.1
episode: 50 epsilon: 0.2695302029097726 Evaluation Average Reward: 183.6
episode: 60 epsilon: 0.2437582841850842 Evaluation Average Reward: 192.7
episode: 70 epsilon: 0.22045062285189232 Evaluation Average Reward: 183.7
episode: 80 epsilon: 0.19937159173177776 Evaluation Average Reward: 157.7
episode: 90 epsilon: 0.18030809382819335 Evaluation Average Reward: 200.0
episode: 100 epsilon: 0.16306740803722372 Evaluation Average Reward: 200.0
episode: 110 epsilon: 0.1474752408470118 Evaluation Average Reward: 153.5
episode: 120 epsilon: 0.13337396432964374 Evaluation Average Reward: 200.0
episode: 130 epsilon: 0.12062102261259347 Evaluation Average Reward: 200.0
episode: 140 epsilon: 0.10908749072006121 Evaluation Average Reward: 200.0
episode: 150 epsilon: 0.09865677121491266 Evaluation Average Reward: 200.0
episode: 160 epsilon: 0.08922341546501161 Evaluation Average Reward: 109.2
episode: 170 epsilon: 0.08069205761761986 Evaluation Average Reward: 102.3
episode: 180 epsilon: 0.07297645050495305 Evaluation Average Reward: 105.2
episode: 190 epsilon: 0.06599859373444678 Evaluation Average Reward: 200.0
episode: 200 epsilon: 0.05968794514922206 Evaluation Average Reward: 200.0
episode: 210 epsilon: 0.05398070768706516 Evaluation Average Reward: 199.3
episode: 220 epsilon: 0.04881918442847171 Evaluation Average Reward: 200.0
episode: 230 epsilon: 0.04415119531365876 Evaluation Average Reward: 200.0