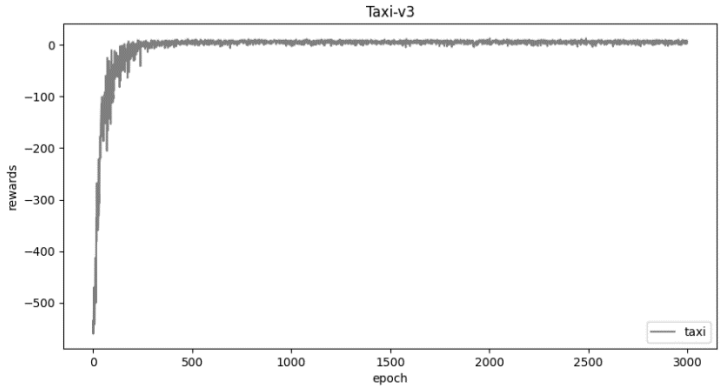
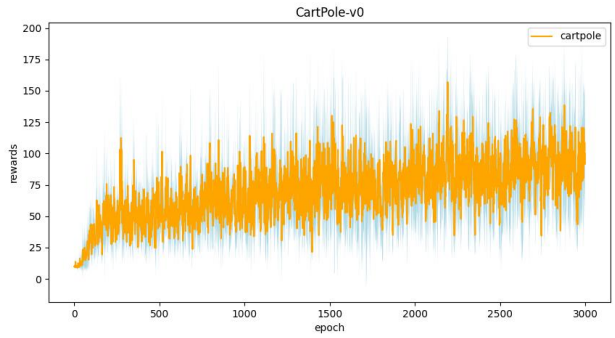
Part I. Experiment Results (the score here is included in your implementation):

Please paste taxi.png, cartpole.png, DQN.png and compare.png here.

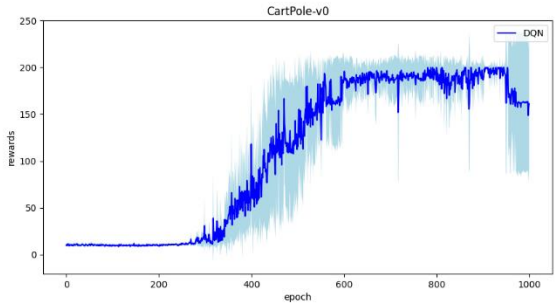
1. taxi.png:



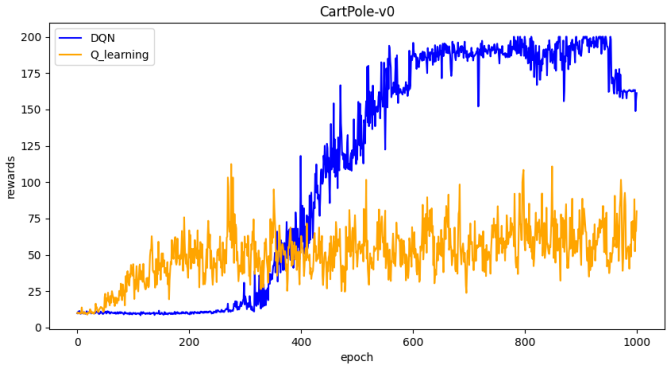
2. cartpole.png:



3. DQN.png:

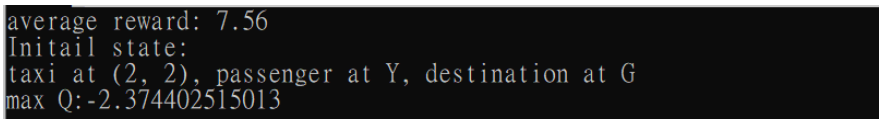


4. compare.png:



Part II. Question Answering (50%):

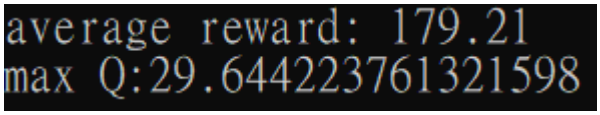
1. Calculate the optimal Q-value of a given state in Taxi-v3 (the state is assigned in google sheet), and compare with the Q-value you learned (Please screenshot the result of the “check_max_Q” function to show the Q-value you learned). (4%)



The optimal Q value = -1(1+r+r^2+r^3+r^4+r^5+...+r^12+r^13)+20*r^13 = -2.62858909785

The result is slightly different since the Q value we learned is an estimated value, it’s reasonable to be different from the optimal Q value.

2. Calculate the max Q-value of the initial state in CartPole-v0, and compare with the Q-value you learned. (Please screenshot the result of the “check_max_Q” function to show the Q-value you learned) (4%)



The optimal Q value = 1(1+r+r^2+….+r^199) = 33.25795863300011

The result is slightly different from we learned, since the learned value is an estimate value, it’s reasonable to be different from the optimal Q value.

3.

a. Why do we need to discretize the observation in Part 2? (2%)

Ans: Since we can’t record continuous infinite states with a finite size q table array.

b. How do you expect the performance will be if we increase “num_bins”? (2%)

Ans: It is expected to be better, since more states means it approximate the continuous states better.

c. Is there any concern if we increase “num_bins”? (2%)

Ans: As the num_bins increases, the size of array would grow, and would require more memory.

4. Which model (DQN, discretized Q learning) performs better in Cartpole-v0, and what are the reasons? (3%)

Ans: DQN will perform better since it can compute q value for continuous states, which can express every possible states.

5.
a. What is the purpose of using the epsilon greedy algorithm while choosing an action? (2%)
Ans: At first, since the machine hasn't learned anything, we had no choice but to randomly pick actions, as the machine learns more, we would prefer the machine's action more as the machine becomes better.

b. What will happen, if we don't use the epsilon greedy algorithm in the CartPole-v0 environment? (3%)
Ans: The machine can't use its current optimal action to improve itself, which means it could learn nothing at all.
c. Is it possible to achieve the same performance without the epsilon greedy algorithm in the CartPole-v0 environment? Why or Why not? (3%)
Ans: it's possible for achieving the same performance if we can find another function adequately choose action between the machine's output or random one.

d. Why don't we need the epsilon greedy algorithm during the testing section? (2%)
Ans: since we believe the machine is fully learned, we don't need random actions but to believe the machine's optimal choice.

6. Why is there "with torch.no_grad():" in the "choose_action" function in DQN? (3%)
Ans: If it chooses a random action, there's no need to take gradient. If it chooses the action by argmax, taking gradient on argmax is undefined. So use "with torch.no_grad():" to exclude taking gradient.

7.
a. Is it necessary to have two networks when implementing DQN? (1%)
Ans: No, we could use only one network to calculate the q value for both current states and next states.
b. What are the advantages of having two networks? (3%)
Ans: If target net is updated every 100 times, the learning process would be more stable.

c. What are the disadvantages? (2%)
Ans: We need extra memory to save the other network.

8.
a. What is a replay buffer(memory)? Is it necessary to implement a replay buffer? What are the advantages of implementing a replay buffer? (5%)
Ans: Replay buffer is to give adequate amount of states for learning. It samples a batch of states and therefore disorder the data.
It is not necessary, the model could still learns in the way as q learning does, but, the results could be undesired.
The advantage is that, since we may not have all states calculate at one time with limited memory, replay buffer could return adequate amount of states with batch size, also, the sampling helps disorder the states and prevent the network from local minimum.
b. Why do we need batch size? (3%)
Ans: We may not have enough memory to calculate all possible states at once, also, use larger batch size would better approximate
c. Is there any effect if we adjust the size of the replay buffer(memory) or batch size? Please list some advantages and disadvantages. (2%)
Ans: If we increase the batch size, the program may run faster but increase the load of memory.
If we decrease the batch size, memory is reduced., but would take longer time for running.

9.
a. What is the condition that you save your neural network? (1%)
Ans: We test our network during learning, and save the network when average reward is larger than a specific value.

b. What are the reasons? (2%)
Ans: Since the code perform test to report reward, we could save the network with good performance (rewards) during learning.

10. What have you learned in the homework? (2%)
Ans: I learned how to use lots of different libraries to help me deal with the homework and how to train neural networks.