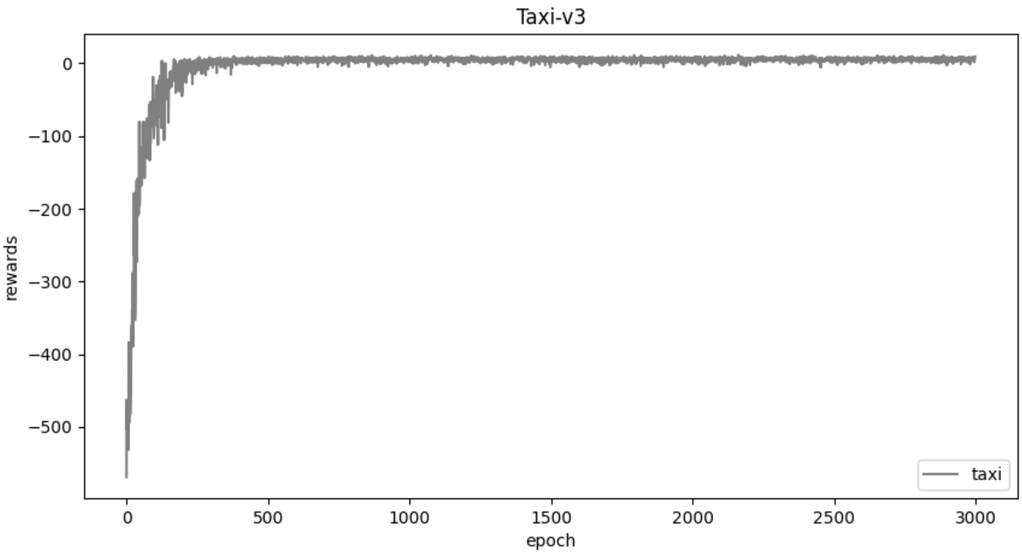
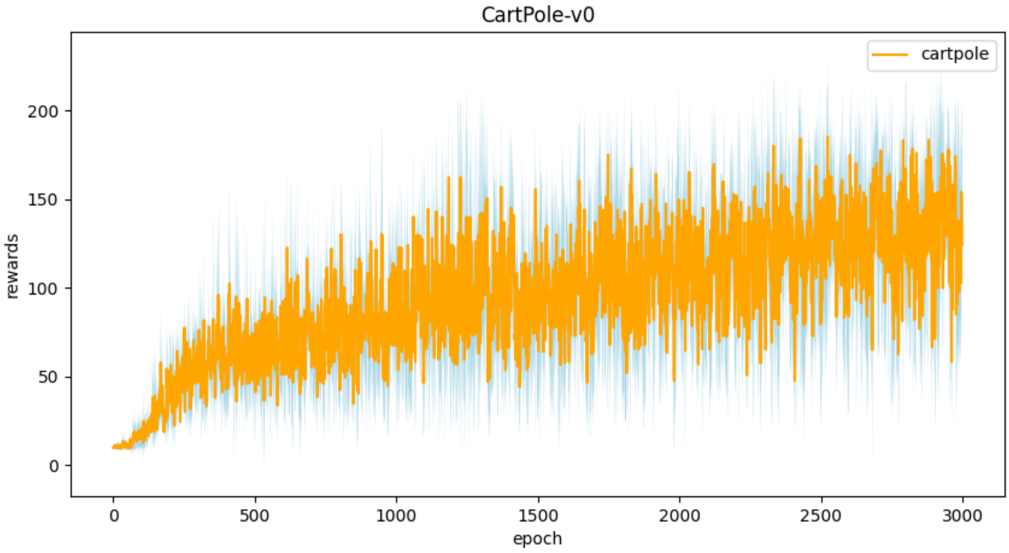
Homework 4: Reinforcement Learning Report Template

Part I. Experiment Results

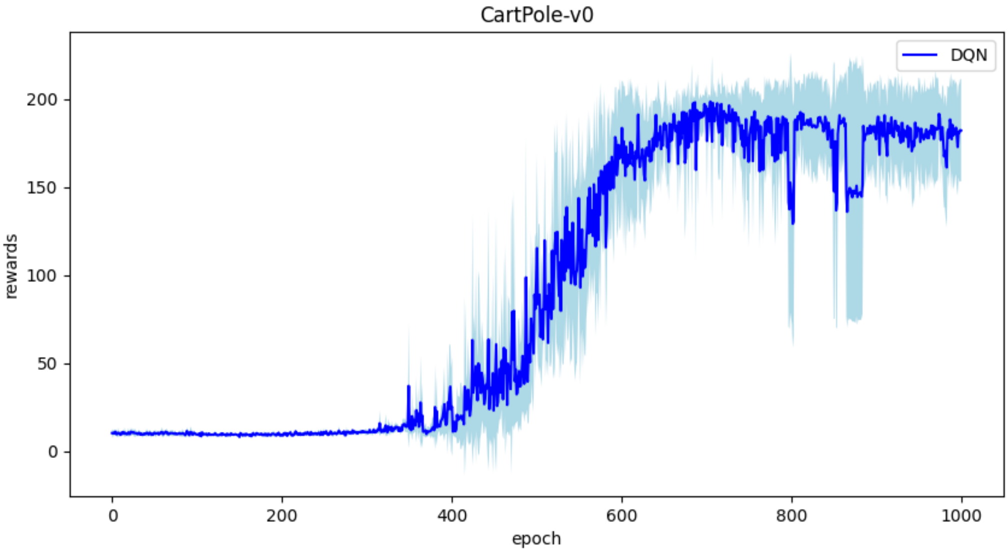
1. taxi.png:



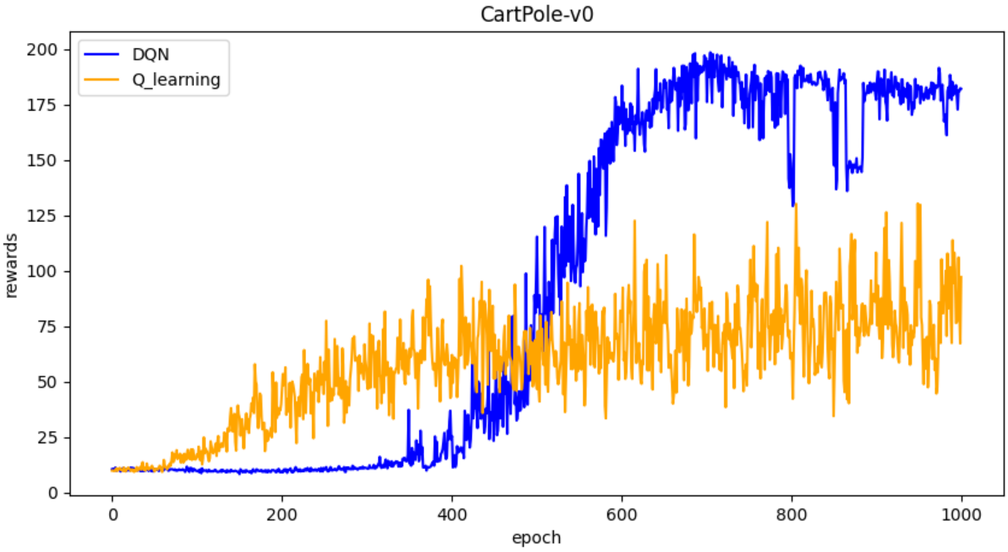
2. cartpole.png



3. DQN.png



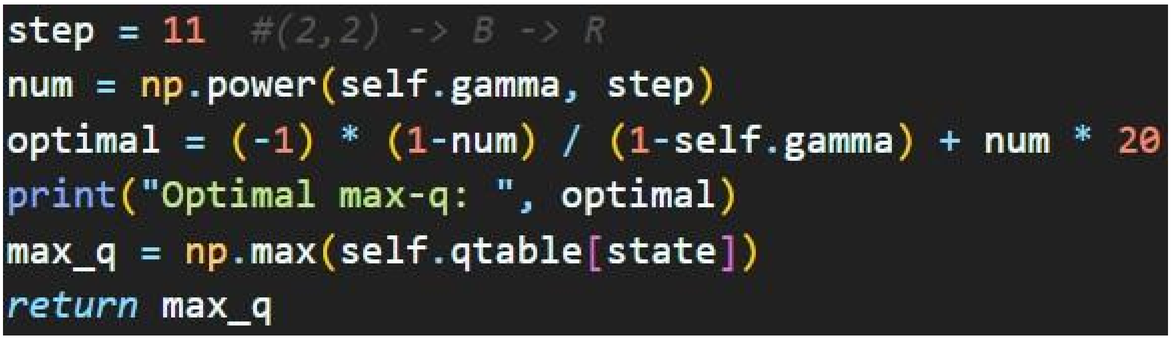
4. compare.png



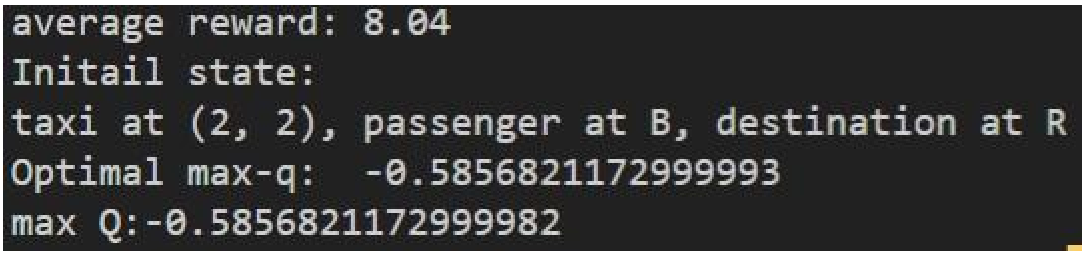
Part II. Question Answering:

1. Calculate the optimal Q-value of a given state in Taxi-v3, and compare with the Q-value you learned

Ans:



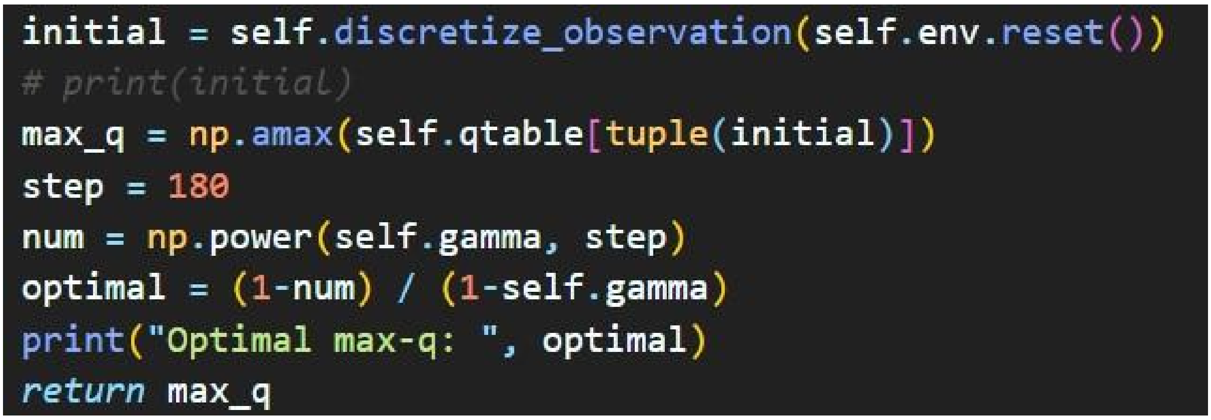
```
step = 11  #(2,2) -> B -> R
num = np.power(self.gamma, step)
optimal = (-1) * (1-num) / (1-self.gamma) + num * 20
print("Optimal max-q: ", optimal)
max_q = np.max(self.qtable[state])
return max_q
```



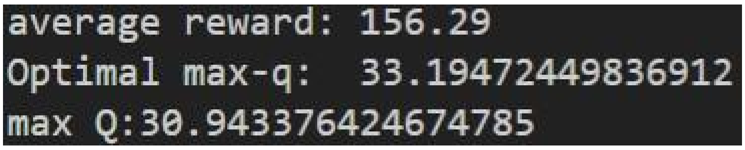
```
average reward: 8.04
Initail state:
taxi at (2, 2), passenger at B, destination at R
Optimal max-q:  -0.585682117299993
max Q:-0.585682117299982
```

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)

Ans:



```
initial = self.discretize_observation(self.env.reset())
# print(initial)
max_q = np.amax(self.qtable[tuple(initial)])
step = 180
num = np.power(self.gamma, step)
optimal = (1-num) / (1-self.gamma)
print("Optimal max-q: ", optimal)
return max_q
```



```
average reward: 156.29
Optimal max-q:  33.19472449836912
max Q:30.943376424674785
```

3.

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

Ans: Because state in cartpole is sequential, so we need to discretize them in order to build a table

b. How do you expect the performance will be if we increase "num_bins"?

Ans: I expect the performance while be better because we have more precise state, which will lead to better choice

c. Is there any concern if we increase "num_bins"?

Ans: If we discretize too much, the cost of space and time will increase extremely

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

Ans: DQN is better than discretized Q learning.
1.avoid large space usage when state and action increase dramatically
2.nural network

5.

a. What is the purpose of using the epsilon greedy algorithm while choosing an action?

Ans: a simple method to balance exploration and exploitation by choosing between exploration and exploitation randomly

b. What will happen, if we don't use the epsilon greedy algorithm in the CartPole-v0 environment?

Ans: the agent will use optimal solution in the training process which is not good for the model

c. Is it possible to achieve the same performance without the epsilon greedy algorithm in the CartPole-v0 environment? Why or Why not?

Ans: No, because as I mentioned above the agent will only use limited state to train

d. Why don't we need the epsilon greedy algorithm during the testing section?
Ans:

6. Why is there “with torch.no_grad():“ in the “choose_action” function in DQN?

Ans: because we don't need to calculate gradient and build graph in the data of with torch.no_grad():, we only to pick the right action

7.

a. Is it necessary to have two networks when implementing DQN?

Ans: No it is not necessary, but if we use two networks will stabilise the Q-training that otherwise has problems converging

b. What are the advantages of having two networks?

Ans: use two network will keep runaway bias from bootstrapping from dominating the system numerically, causing the estimated Q values to diverge

c. What are the disadvantages?

Ans:The training time will increase if we use two 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?

Ans:

1.replay memory is a memory to store past experience and randomly to refresh network

2.It is not necessary to implement a replay buffer, but it will make dqn better

3. avoid bad choose or diverging, have more state to experience rather then use most recent experience, gradually change the past experience for next choose action

b. Why do we need batch size?

Ans: we need state to keep track of earlier state train in the neural network which control the accuracy o estimate of the error gradient

c. Is there any effect if we adjust the size of the replay buffer(memory) or batch size? Please list some advantages and disadvantages.

And:If we increase the size of batch size, the accuracy of train network will increase, but the training time will also increase

9.

a. What is the condition that you save your neural network?

Ans: If loss smaller than 0.5, I will store my neural network

b. What are the reasons?

Ans: I first print loss value, and I observe the value of loss, most value is larger than loss. Since loss means difference between target network and eval network, I think if the difference is small enough, I will store the network.

10. What have you learned in the homework?

Ans:I learned how to implement RL in a gym environment and how to use a tensor in artificial intelligence as a beginner. I also learned the difference between DQ and deep learning.