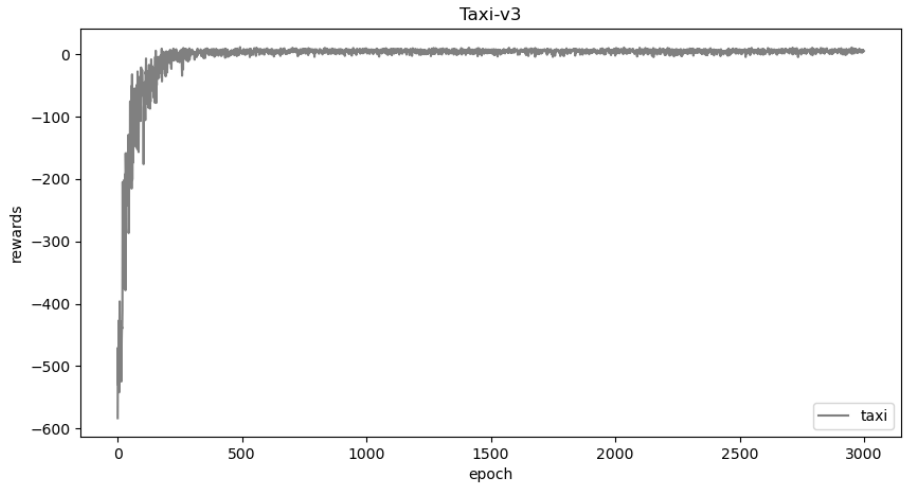
Homework 4 : Reinforcement Learning

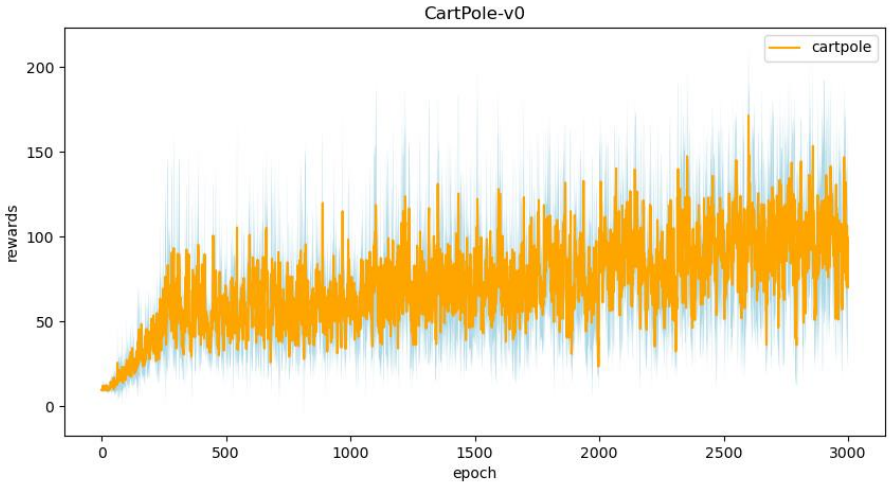
109550135 范恩宇

Part I. Experiment Results :

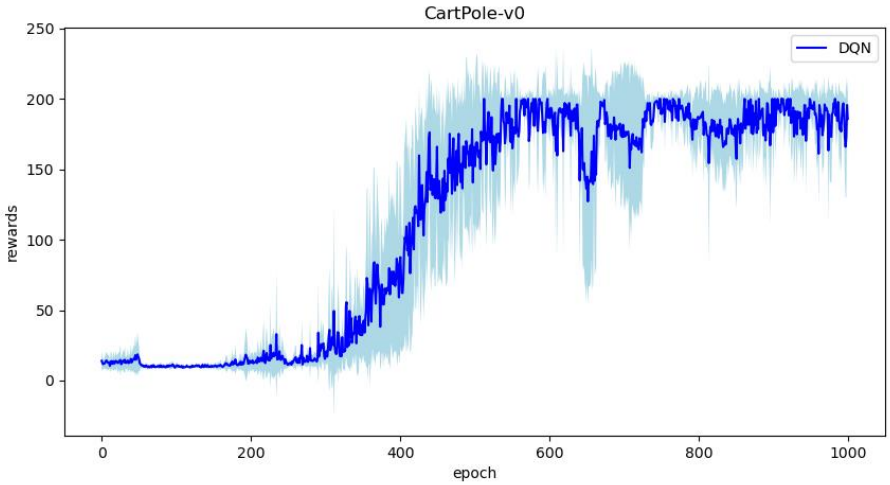
1. **taxi.png :**



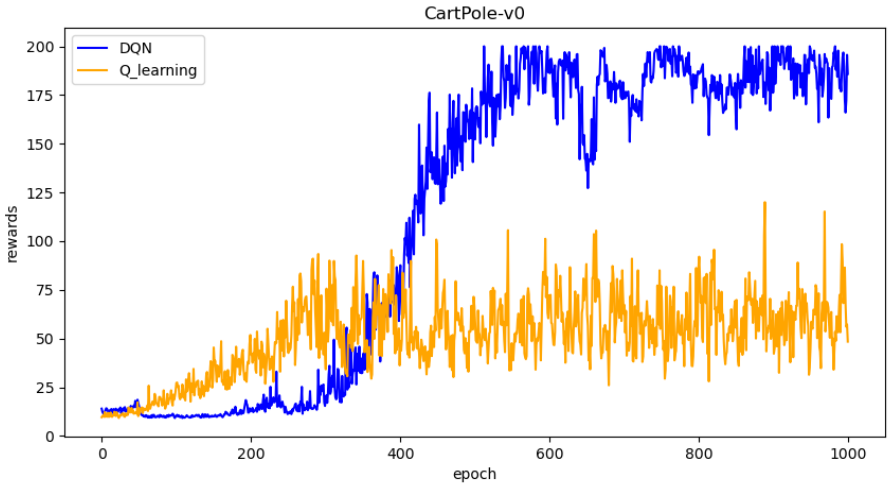
2. **cartpole.png :**



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%)**

   Optimal Q-value : -2.3745 [(0.9^13 - 1)/(1-0.9) + 20* 0.9^13]

```
(base) D:\學業\人工智慧概論\作業\AI_HW4_updated>python taxi.py
C:\Users\En-yu Fan\AppData\Roaming\Python\Python39\site-packages\gym\core.py:172: DeprecationWarning: WARN: Function `en
v.seed(seed)` is marked as deprecated and will be removed in the future. Please use `env.reset(seed=seed) instead.
  deprecation(
100%|                                                                                | 3000/3000 [00:56<00:00, 53.57it/s]
100%|                                                                                | 3000/3000 [00:55<00:00, 53.96it/s]
100%|                                                                                | 3000/3000 [00:56<00:00, 52.98it/s]
100%|                                                                                | 3000/3000 [01:02<00:00, 47.84it/s]
100%|                                                                                | 3000/3000 [01:00<00:00, 49.61it/s]
average reward: 7.55
Initail state:
taxi at (2, 2), passenger at Y, destination at G
max Q:-2.374402515013
```

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%)**

   Max Q-value : 33.2580 [1*(1-0.97^200)/(1-0.97)]

```
(base) D:\學業\人工智慧概論\作業\AI_HW4_updated>python cartpole.py
C:\Users\En-yu Fan\AppData\Roaming\Python\Python39\site-packages\gym\envs\registration.py:505: UserWarning: WARN: The en
vironment CartPole-v0 is out of date. You should consider upgrading to version `v1` with the environment ID `CartPole-v1
`.
  logger.warn(
C:\Users\En-yu Fan\AppData\Roaming\Python\Python39\site-packages\gym\core.py:172: DeprecationWarning: WARN: Function `en
v.seed(seed)` is marked as deprecated and will be removed in the future. Please use `env.reset(seed=seed) instead.
  deprecation(
#1 training progress
100%|                                                                                | 3000/3000 [02:11<00:00, 22.87it/s]
#2 training progress
100%|                                                                                | 3000/3000 [02:18<00:00, 21.61it/s]
#3 training progress
100%|                                                                                | 3000/3000 [02:21<00:00, 21.25it/s]
#4 training progress
100%|                                                                                | 3000/3000 [02:20<00:00, 21.35it/s]
#5 training progress
100%|                                                                                | 3000/3000 [02:24<00:00, 20.69it/s]
average reward: 198.54
max Q:30.254679281424785
```

3.

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

   Since the values we observe in cartpole environment are continuous , which are not suitable for being the input of a table .

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

   The accuracy of the received result will be improved , since it is closer to the original continuous data .

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

It may make the discrete parts too big for the agent to find the optimal value , and it may need larger network to model the action-value function.

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

Q-learning agent performs better than DQN at fist , since DQN needs some data before being able to train a reasonable model of Q-values. However , Q-learning agent may perform poorly due to greedy random action . But DQN can generalize to states that it hasn't encountered, making its performance become better and more stable after a while .

5.

a. What is the purpose of using the epsilon greedy algorithm while choosing an action? (2%)

It can balance exploration and exploitation by choosing between exploration and exploitation randomly . Also , it's easy to be implemented .

b. What will happen, if we don't use the epsilon greedy algorithm in the CartPole-v0 environment? (3%)

It will be difficult to balance exploration and exploitation . In addition , efficiency of choosing best action will be poorer . Besides , other methods for solving former problems are often more difficult .

c. Is it possible to achieve the same performance without the epsilon greedy algorithm in the CartPole-v0 environment? Why or Why not? (3%)

It is possible . Since there are still other ways for balancing exploration and exploitation and enhancing efficiency of choosing best action , it's still possible , but more difficult to achieve .

d. Why don't we need the epsilon greedy algorithm during the testing section? (2%)

When using epsilon greedy algorithm during testing section , if epsilon is very small , there may be a strong bias towards exploitation over exploration,making the agent choose the action with the highest q-value rather than a random action

6. **Why is there “with torch.no_grad():“ in the “choose_action” function in DQN? (3%)**

Because “with torch.no_grad():“ is generally for stopping the work of the autograd module to accelerate and save memory , putting this in “choose_action” helps stop the gradient .

7.

a. Is it necessary to have two networks when implementing DQN? (1%)

No , it’s not necessary .

b. What are the advantages of having two networks? (3%)

It improves the stability . Using a separate target network , helps keep runaway bias from bootstrapping through dominating the system numerically, causing the estimated Q values to diverge.

c. What are the disadvantages? (2%)

It lowers the speed of learning , increases the complexity of a sample and results in extra memory used .

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%)

It is for storing trajectories of experience when executing a policy in an environment . During training, replay buffers are also queried for either a sequential subset or a sample of trajectories to replay the agent's experience.

It is necessary. Because if it’s choosen sequentially, similarity between the data will be large, and the network will also converge to the local optimum value easily .

First , the agent won’t have to fit the model to the same small mini-batch for multiple iterations . Second , it helps the training process be able to use diverse mini-batch for performing updates.

b. Why do we need batch size? (3%)

Because it is an important hyperparameter that influences the dynamics of the learning algorithm , and we need it to control the accuracy of estimate of the error gradient when training neural networks .

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%)

Yes . For example , with bigger batch size, the noise in the gradients is lesser and gradient estimate is better , which allows the model to take a better step towards a minima . However, bigger batch size needs more memory , each step is time consuming and lead to poor generalization .

9.

a. What is the condition that you save your neural network? (1%)

There's no additional conditions added .

b. What are the reasons? (2%)

The result I got without adding additional conditions meets the requirement(= 200 , which is >195) .

10. What have you learned in the homework? (2%)

Actually , I didn't totally understand how Q learning exactly works in the lecture , this homework do help me understand how it operates and how each self defined function constructs the whole reinforcement learning method . In addition , the results I get and the process of implementation also helps me know what are and what cause the difference between Q learning and DQN . Implementing something practically is much more helpful than just reading theories on papers .