Deep Learning HW3_2 Report

0853411 劉書維

1. **Explain the purpose of the following hyperparameters.**

   α：指的是 updating step，也就像是機器學習中的 learning rate，根據 gradient descent 來更新參數，可以決定下一步要走大步或小步（變化量），進而決定下一個 state。

   γ：是 discount factor，可以決定未來還是當下 reward 的影響。要是時間越長，影響力就會越小，也就是執行的每一個 action 都會影響到後面的 reward，所以前面的 action 對後面的 reward 影響越小。

   τ：是 target network update period，是為了穩定 DQN 所以增加的網路參數，Q-network 分為兩個，一為實際進行訓練的 evaluation network，一為訓練目標 target network，其中 target network 久久更新一次，這個就是更新參數。

   ε：用於 greedy policy，這個參數決定我們利用目前的所有的 Q 值來找出一個最好的動作。就像是 exploration 去嘗試其他 action。在這次實作上也可以看出 ε 會隨著時間遞減，意思是在一開始還不知道哪個 action 比較好，所以會一直 exploration 來尋找最佳 action。

2. **Please show the total reward for the configuration.**

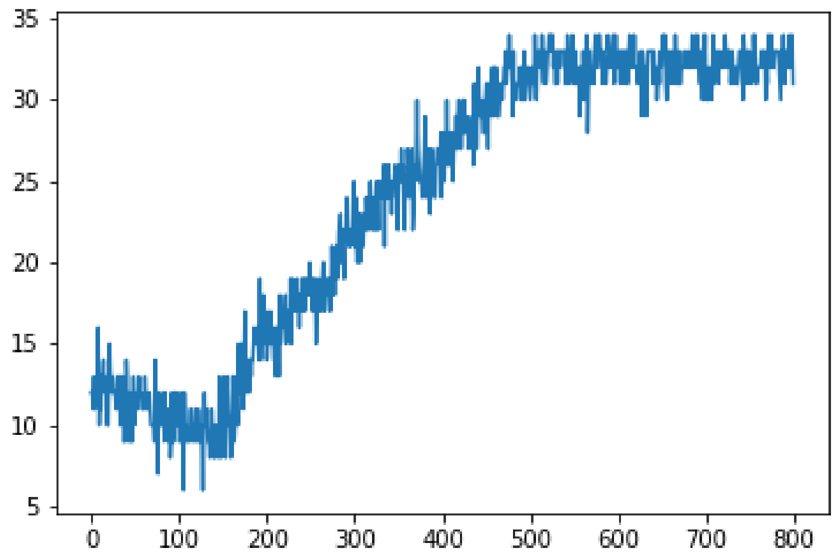
為了讓訓練速度增快，我也比較貪心，所以我將三者機率設為[NOOP (0.25), UP (0.65), DOWN (0.1)]。前 20 個 epoch 如下圖，可以看出一開始 training reward 就有大概 12 分的成績，然後每 10 個 epoch 記錄一次 evaluation reward 一開始當然是 0 分：

```
Episode:      1, interaction_steps:   2048, reward: 12, epsilon: 0.998054
[Info] Save model at './model' !
Evaluation: True, episode:      1, interaction_steps:   2048, evaluate reward:  0
Episode:      2, interaction_steps:   4096, reward: 12, epsilon: 0.996109
Episode:      3, interaction_steps:   6144, reward: 11, epsilon: 0.994163
Episode:      4, interaction_steps:   8192, reward: 13, epsilon: 0.992218
Episode:      5, interaction_steps:  10240, reward: 11, epsilon: 0.990272
Episode:      6, interaction_steps:  12288, reward: 12, epsilon: 0.988326
Episode:      7, interaction_steps:  14336, reward: 13, epsilon: 0.986381
Episode:      8, interaction_steps:  16384, reward: 12, epsilon: 0.984435
Episode:      9, interaction_steps:  18432, reward: 16, epsilon: 0.982490
Episode:     10, interaction_steps:  20480, reward: 10, epsilon: 0.980544
Episode:     11, interaction_steps:  22528, reward: 13, epsilon: 0.978598
Evaluation: True, episode:     11, interaction_steps:  22528, evaluate reward:  0
Episode:     12, interaction_steps:  24576, reward: 11, epsilon: 0.976653
Episode:     13, interaction_steps:  26624, reward: 13, epsilon: 0.974707
Episode:     14, interaction_steps:  28672, reward: 13, epsilon: 0.972762
Episode:     15, interaction_steps:  30720, reward: 14, epsilon: 0.970816
Episode:     16, interaction_steps:  32768, reward: 12, epsilon: 0.968870
Episode:     17, interaction_steps:  34816, reward: 12, epsilon: 0.966925
Episode:     18, interaction_steps:  36864, reward: 13, epsilon: 0.964979
Episode:     19, interaction_steps:  38912, reward: 10, epsilon: 0.963034
Episode:     20, interaction_steps:  40960, reward: 10, epsilon: 0.961088
```

3. **Plot the episode reward in learning time and evaluation time. Show your configuration and discuss what you find in training phase.**

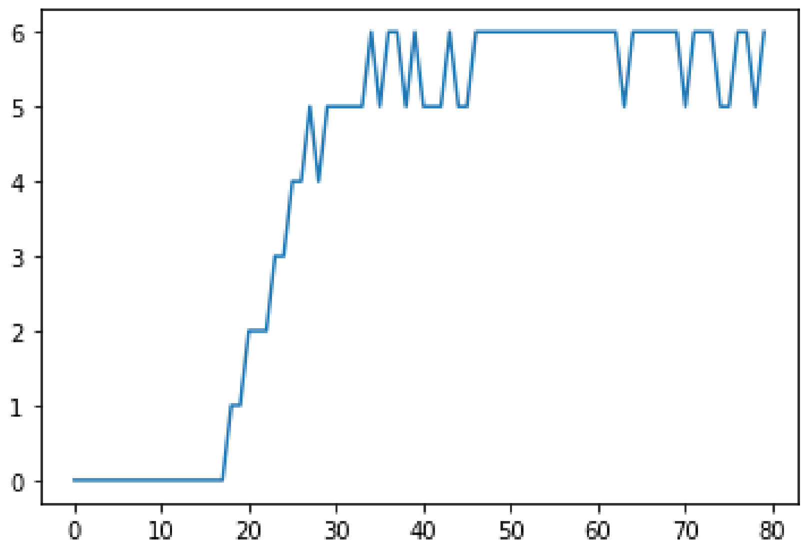
這些 reward 都已經放入 log 中的 txt 檔中，也可以藉由 plot.py 程式畫出如下圖的圖片。

Reward in training：



可以看出跑了 200 個 epoch 之後開始穩定上升，到 500 個 epoch 開始收斂，最後的 reward 大概都有 32 分，共有 800 筆資料。

Reward in evaluation：



每跑 10 次 training 就記錄一次的 evaluation reward 也和 training reward 的趨勢一樣，隨之上升。

4. After training, you will obtain the model parameters for the agent.

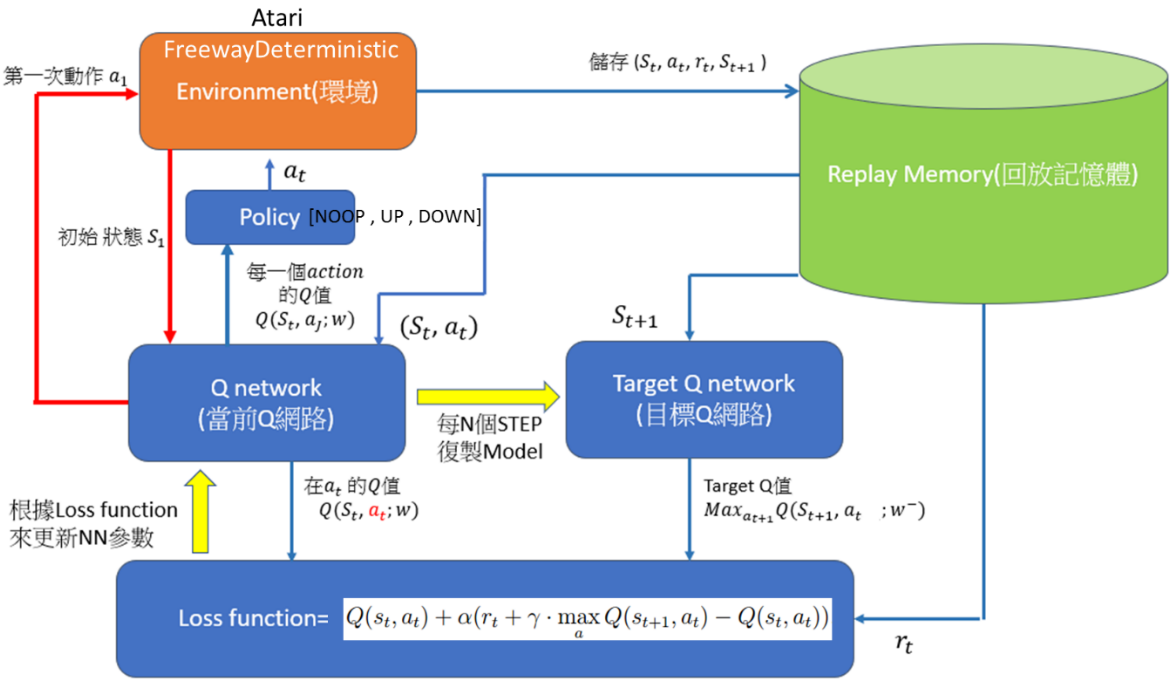
Show total reward inn some episodes for deep Q-network agent.

這是最後 10 次訓練的結果，可以看到 training reward 已經穩定到 32 分了。Model 也已經放入 model 的資料夾中！

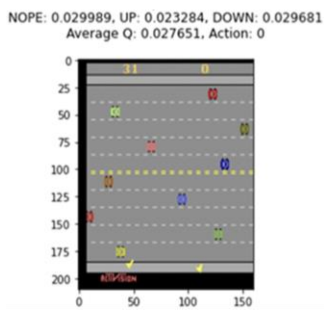
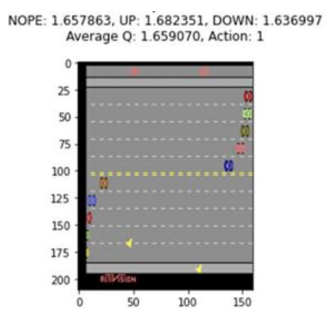
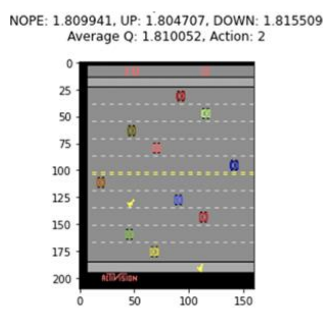
```
Episode:    791, interaction_steps: 1619968, reward: 31, epsilon: 0.050000
Evaluation: True, episode:    791, interaction_steps: 1619968, evaluate reward:
Episode:    792, interaction_steps: 1622016, reward: 33, epsilon: 0.050000
Episode:    793, interaction_steps: 1624064, reward: 33, epsilon: 0.050000
Episode:    794, interaction_steps: 1626112, reward: 32, epsilon: 0.050000
Episode:    795, interaction_steps: 1628160, reward: 32, epsilon: 0.050000
Episode:    796, interaction_steps: 1630208, reward: 34, epsilon: 0.050000
Episode:    797, interaction_steps: 1632256, reward: 32, epsilon: 0.050000
Episode:    798, interaction_steps: 1634304, reward: 32, epsilon: 0.050000
Episode:    799, interaction_steps: 1636352, reward: 34, epsilon: 0.050000
Episode:    800, interaction_steps: 1638400, reward: 31, epsilon: 0.050000
```

5. **Sample some states, show the Q values for each action, analyze the results, and answer**

DQN 的架構如下圖設計：



所用的參數調整： batch_size =128、lr =0.0005、gamma= 0.999、 epsilon_start =1.0、epsilon_final =0.05。

NOOP	UP	DOWN
		

a. **Is DQN decision in the game the same as yours? Any good or bad move?**

並不一定像我的遊戲策略，有時也會有意想不到的步伐，並不能說是好或者不好，可能 DQN 會對未來做一些預測，所以才會有和我不太一樣的步伐。

b. **Why the averaged Q-value of three actions in some state is larger or less than those of the other states?**

averaged Q-value 越高代表往那個策略走的 reward 可能較高，所以就是一個決策的選擇，往越高的就是可能的 reward 會較高。