Homework 4 Report

Questions

1. What kind of RL algorithms did you use? value-based, policy-based, model-based? why?
   - DDPG (Deep Deterministic Policy Gradient)

   我使用 DDPG (Deep Deterministic Policy Gradient)。是 actor-critic 的強化版，並結合 DQN 的演算法，是種結合了 value-based 和 policy-based 方法，在選擇 action 時 Deterministic 改變了原本 Policy gradient 選擇動作的過程，Deterministic 只在連續動作上選定了一個輸出當作 action。policy 網路是 actor(演員)，輸出動作（action-selection）。value 網路是 critic(評價家)，用來評價 actor 網路所選動作的好壞（action value estimated）。

2. This algorithm is off-policy or on-policy? why?

   DDPG 這個演算法是 off-policy，有通過其他新的策略來更新當前的 Q 值，會參考到歷史紀錄，不一定使用當前策略所產生的樣本更新當前的 Q 值，可以隨機抽取一些之前的經歷進行學習. 隨機抽取這種做法打亂了經歷之間的相關性, 也使得神經網絡更新更有效率。

3. How does your algorithm solve the correlation problem in the same MDP?

   MDP： 在當前狀態 S_t 採取動作 a_t 後的狀態 S_{t+1} 和 reward r_{t+1} 只和當前的狀態與動作有關，與歷史狀態並無關係。

   而 DDPG 這個演算法在當前動作採取每一個動作時，都會參考到過去歷史的狀態。

Analysis of algorithm

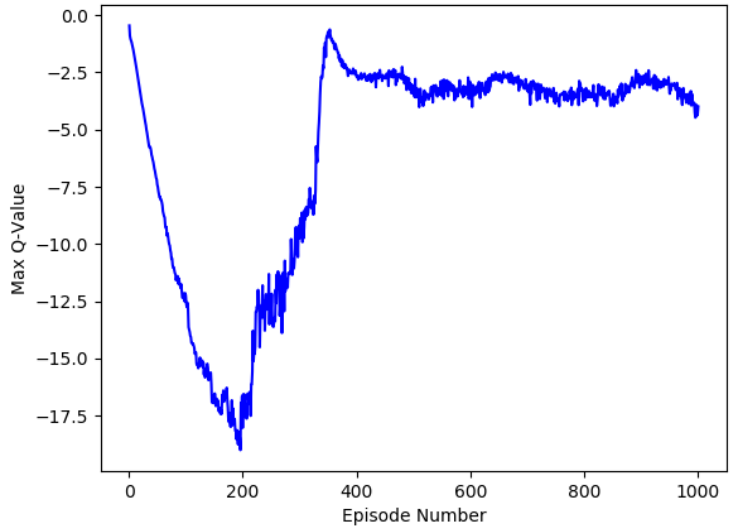
利用每一次訓練的 Max Q-value 和 Reward 來評估

- GAMMA = 0.99, TAU = 0.001, BATCH SIZE = 64, episode = 1000, step = 10000

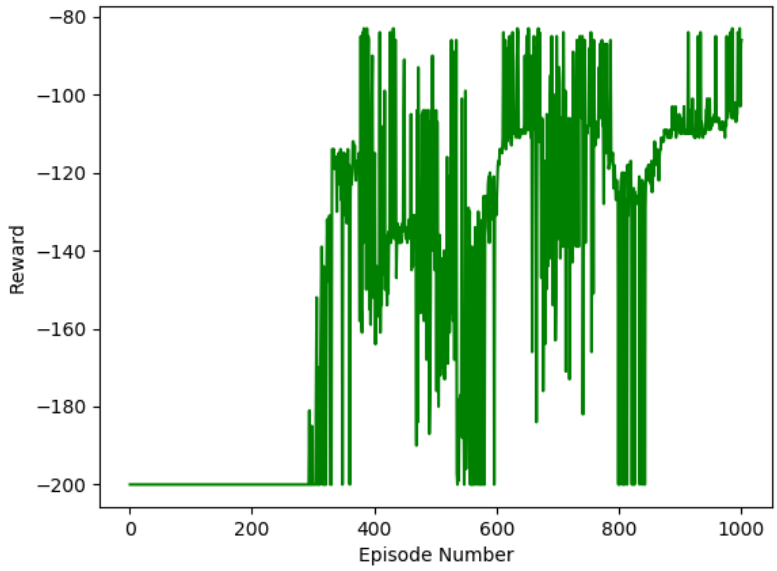
一開始的 Q value 會往下掉，表示模型正在探索甚麼樣的動作會比較好，在訓練到約第 200 次時，有一個轉折點 Q 值會越來越高，表示模型有學到什麼樣 action 友好的結果可以到達終點。

在 Reward 的圖中，reward 最小值為-200，表示沒有訓練時車子沒有到達終點，在大約 250 次訓練的時候 reward 才會開始不是-200，接下來 reward 會一直震盪，因為用的是 value approximator，Neural Network 本身跟環境都帶有隨機性，所以 reward 才會不穩定。

Max Q-value



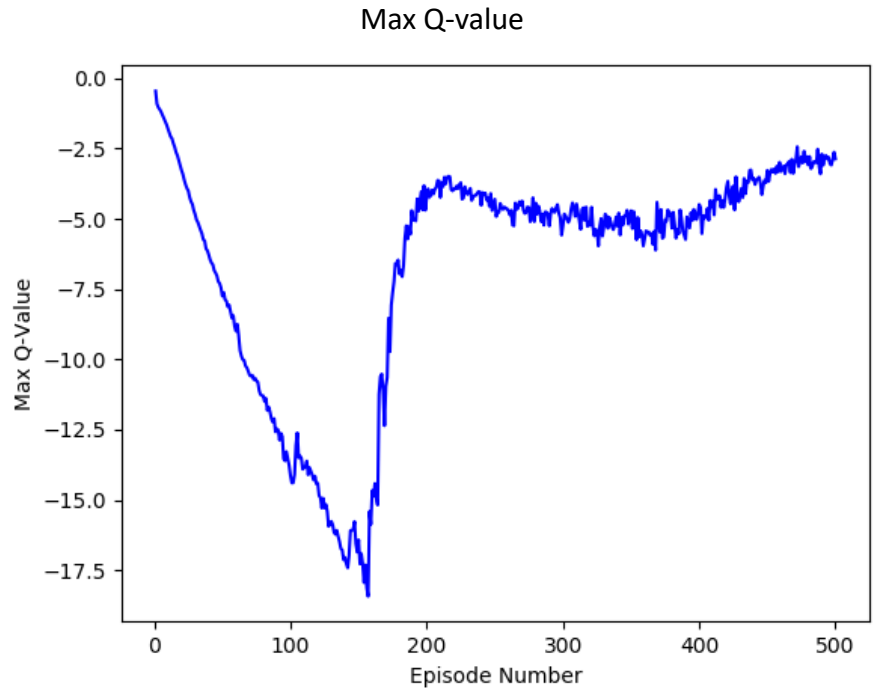
Reward

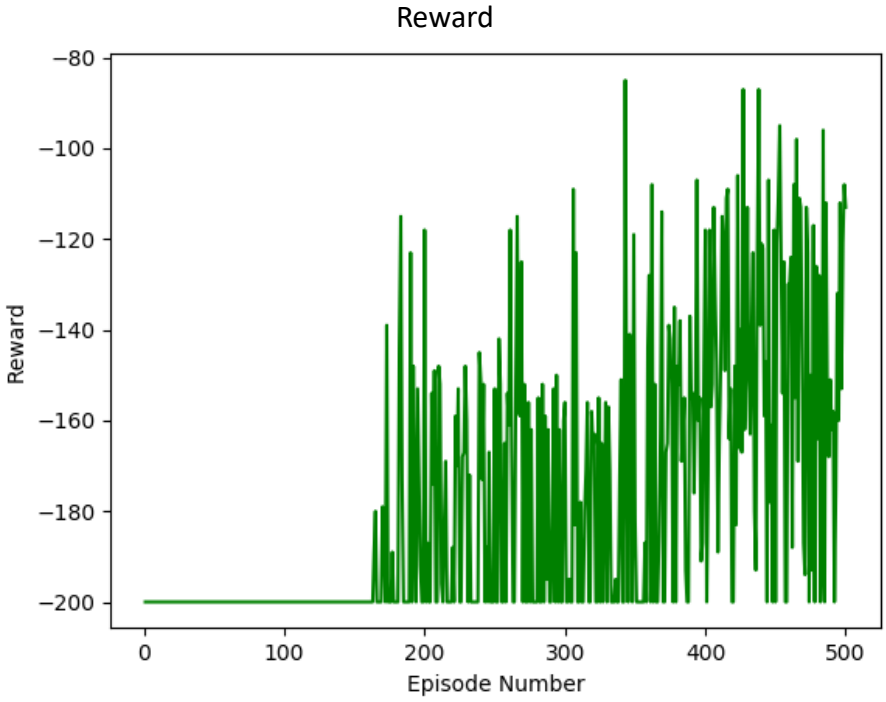


調整 Episode

- GAMMA = 0.99, TAU = 0.001, BATCH SIZE = 64, episode = 500, step = 1000

在訓練到約第 150 次時，有一個轉折點 Q 值會越來越高。在大約 150 次訓練的時候 reward 才會開始不是-200，接下來 reward 會一直震盪，可以看到訓練 500 是不夠的，整個 reward 和 Q 值還在上升，還沒有穩定。



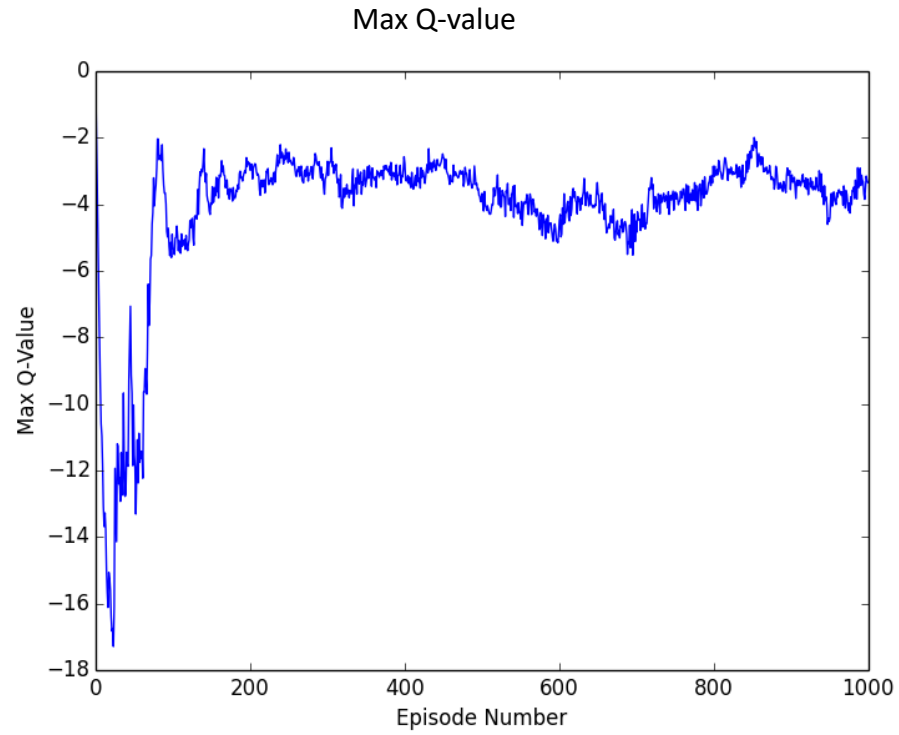


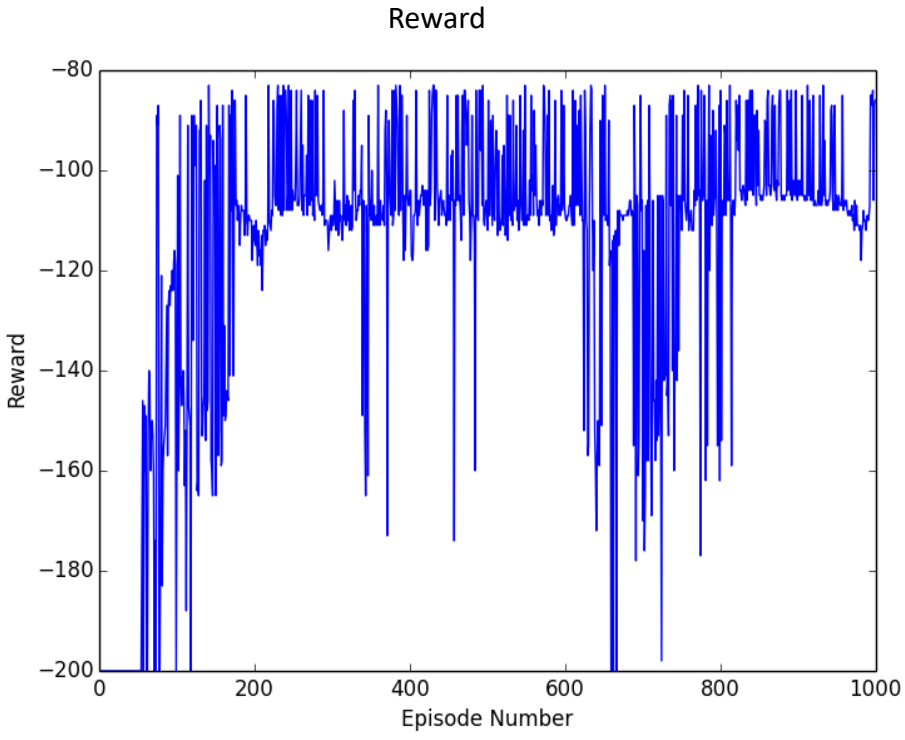
調整 TAU

- GAMMA = 0.99, TAU = 0.01, BATCH SIZE = 64, episode = 1000, step = 10000

讓 Tau 變大 10 倍來讓訓練的速度變快。大約在訓練第 30 次時，就有轉折點，Q 值會提升的較快，在第 200 次時 Q 值就不會有太大的改變。

從 reward 的圖來看，第 50 次時 reward 就不是-200，reward 會開始變大。



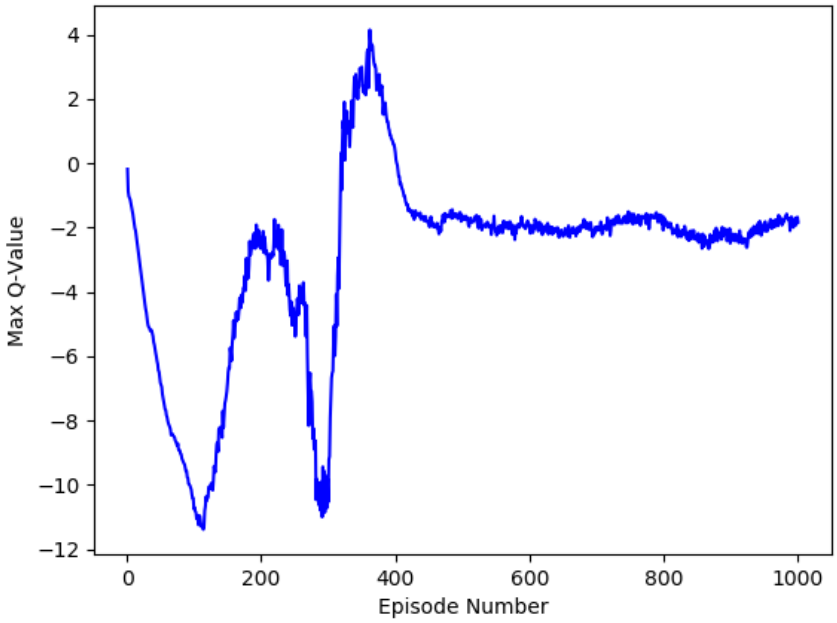


調整 Batch size

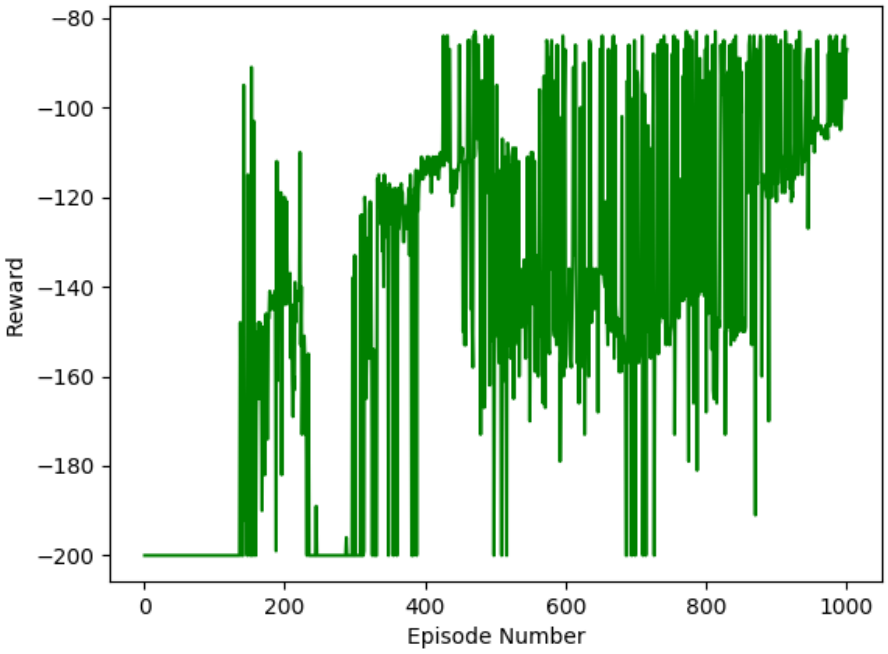
- GAMMA = 0.99, TAU = 0.01, BATCH SIZE = 128, episode = 1000, step = 10000

讓 Tau 變大 10 倍來讓訓練的速度變快。大約在訓練第 30 次時，就有轉折點，Q 值會提升的較快，在第 200 次時 Q 值就不會有太大的改變。
從 reward 的圖來看，第 50 次時 reward 就不是-200，reward 會開始變大。

Max Q-value



Reward



調整 GAMMA

- GAMMA = 0.5, TAU = 0.01, BATCH SIZE = 128, episode = 1000, step = 10000

將 gamma 衰減率調整為 0.5，會使得在訓練的過程變得較慢趨於收斂，訓練變得比較久，Q 值一直下降，reward 也到了約 900 次才開始不為 -200。表示 gamma 越大訓練的速度可以更快。