ADL Homework #3 Report

b04902013 鄧逸軒

1. Basic Performance

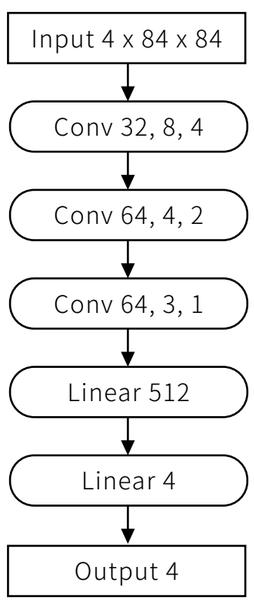
1.1 Network

網路架構與助教提供的模型相同。

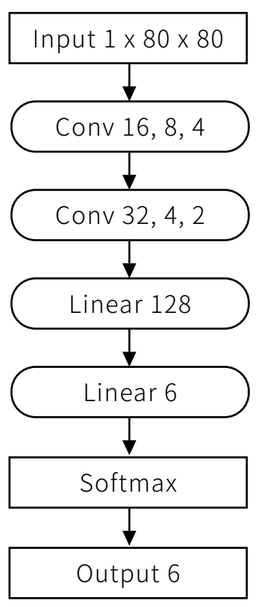
Network Structure

Deep Q Network

RMSProp lr = 1e-4
Gamma = 0.99
Eps decay = 1e-5
Eps start = 0.99
Eps end = 0.05
Memory Size = 16384
Online Net Freq = 4
Target Net Freq = 250
Death penalty = -10



Policy Gradient



RMSProp lr = 1e-4
Death penalty = -10

1.2 Deep Q Network

Epsilon greedy 用的下降方式是指數函數：

$$\epsilon = \text{start} + (\text{end} - \text{start})e^{-\frac{t}{\text{decay}}}$$

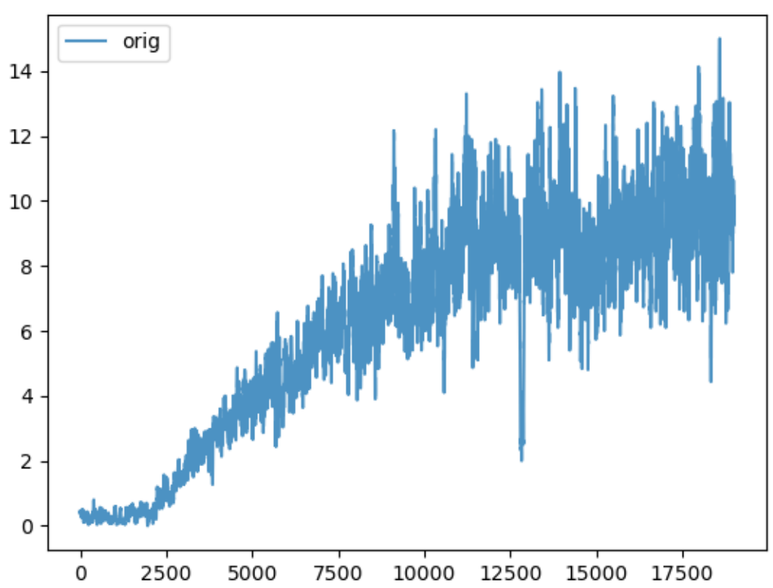
Death penalty 指的是最後一個 reward 改成 -10，一開始因為一直訓練不起來，想說讓他比較不會結束。

1.3 Policy Gradient

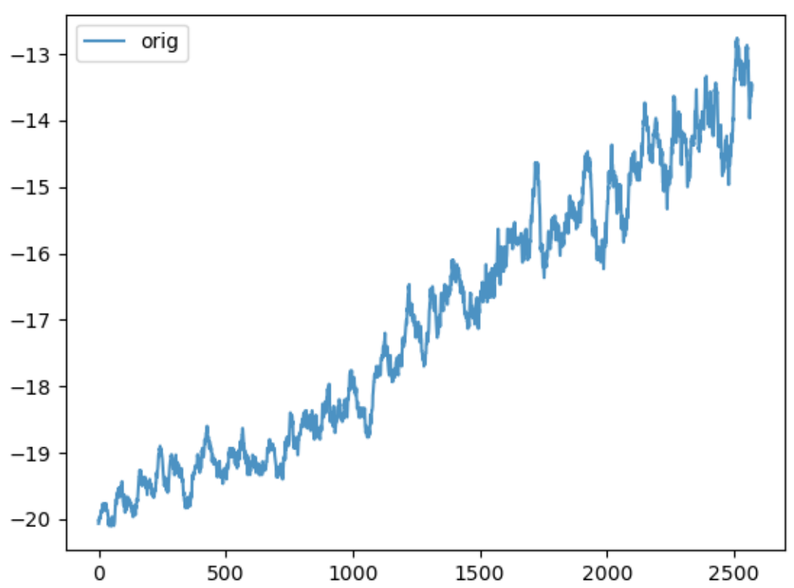
使用的是基本的 REINFORCE 算法。

輸入會先用助教提供的函式做前處裡，然後減去前一個輸入。每個回合的第一個動作是隨機決定。圖片看起來還在持續上升的階段，不過實在訓練過久，沒有在期限內跑完。

Deep Q Network Training Curve



Policy Gradient Training Curve



2. Experimental Results and Settings

2.1 Exploration Rate

調低 decay 的速度讓他可以多探索，不過看起來只是等到 epsilon 夠低才開始訓練起來。

2.2 Memory Size

試著降低 memory size 讓他比較能看到現在的參數玩出來的成果，可能是因為相關性太高效果比較差。

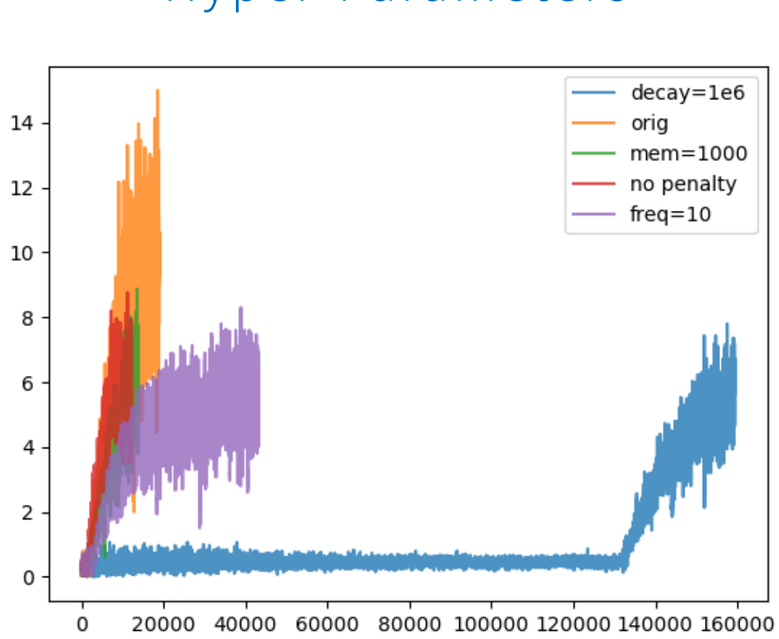
2.3 Target Network Update Frequency

可能是因為目標變動的太快，訓練的比較緩慢。

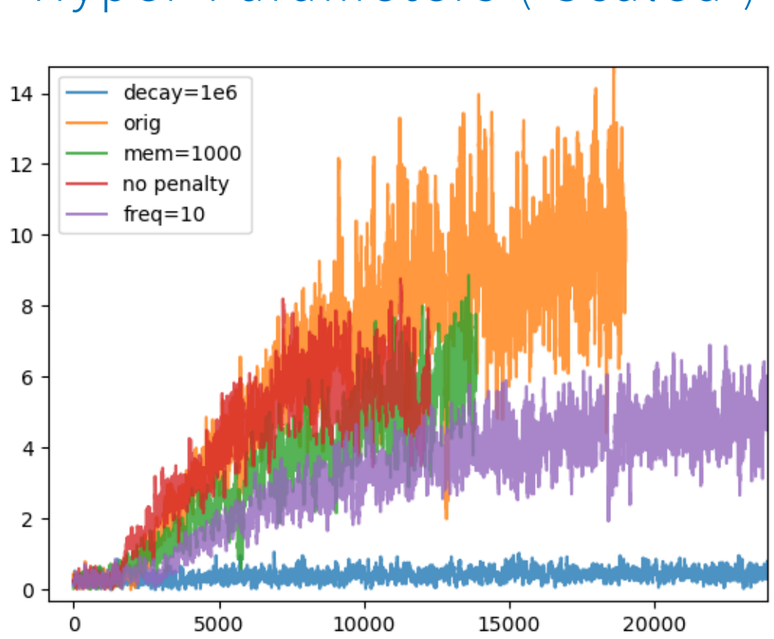
2.4 Death Penalty

如同前面說的，最初是因為訓練不起來，想讓他比較怕死而加上去的。實驗結果看起來前期沒有什麼差別，後期有比較好。

Hyper-Parameters



Hyper-Parameters (Scaled)



3. DQN Variations

3.1 Double DQN

效果看起來反而變差了，有可能是因為 target net 更新頻率太低，所以兩個網路差別過大，造成反效果。

3.2 Dueling DQN

效果看起來掉很多，可能是參數增加的原因，也有可能是 Network 最後面的 ReLU 造成 Advantage 沒辦法是負值。

3.3 Prioritized Replay

訓練前期 Variance 太大，算出來的誤差不太可信，遲遲訓練不起來。

3. DQN Variations

3.4 Training Curve