ADL x MLDS assignment3 report

學號：B03901030 系級： 電機四 姓名：蕭晨豪

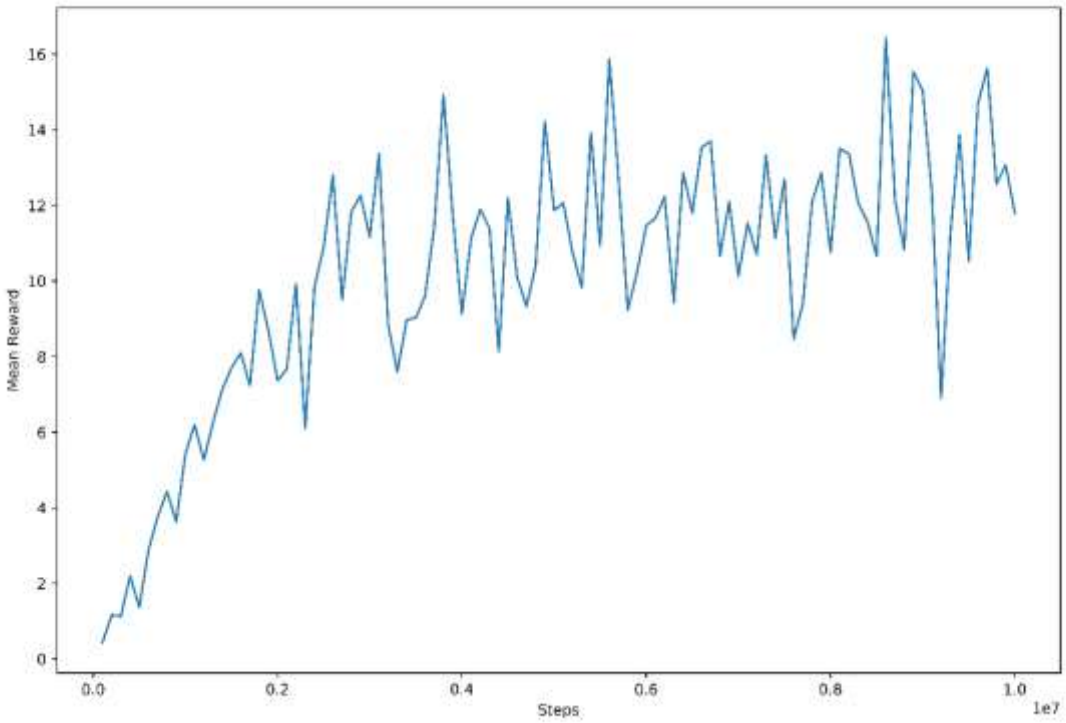
1. Basic Performance

DQN:

一開始過 baseline 的 DQN 我使用的是基礎的 DQN 演算法，主要描述的部分會著重在 model 的架構。由於 input 是 84x84x4 的圖片，我選擇使用 CNN 模型，架構如下:

Conv2d(4,32,kernel = 8x8, stride = 4)
ReLu
Conv2d(32,64,kernel = 4x4 stride = 2)
ReLu
Conv2d(64,64,kernel = 3x3, stride = 1)
ReLu
Linear(64x7x7, 512)
ReLu
Linear(512, num_of_actions)

用 DQN 在 Breakout 遊戲上做 Training 時的 learning curve 如下，x 軸為 step 數，y 軸為近 30 個 episode 的 mean reward(reward 有 clip)

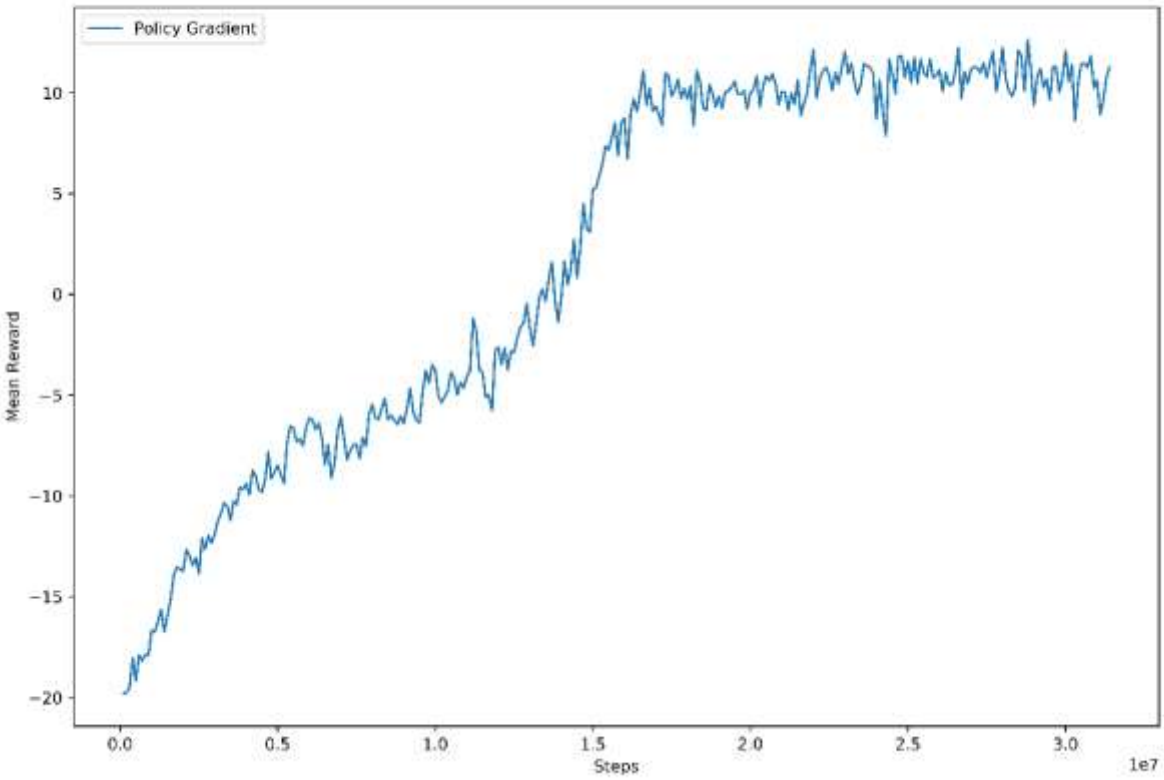


PG:

由於演算法也是使用一般的 PG 演算法，僅描述模型架構如下:

Conv2d(1,16,kernel = 8x8, stride = 4)
ReLu
Conv2d(16,32,kernel = 4x4 stride = 2)
ReLu
Linear(32x8x8, 128)
ReLu
Linear(128, num_of_actions)
Softmax()

注意到較特別的是 1.我使用的 input data 是兩個 frame 的 difference 2.原本 environment 中的動作有六種，但我發現其實真正的動作只有左右和停三種，因此我的 model 的 num_of_actions 是 3



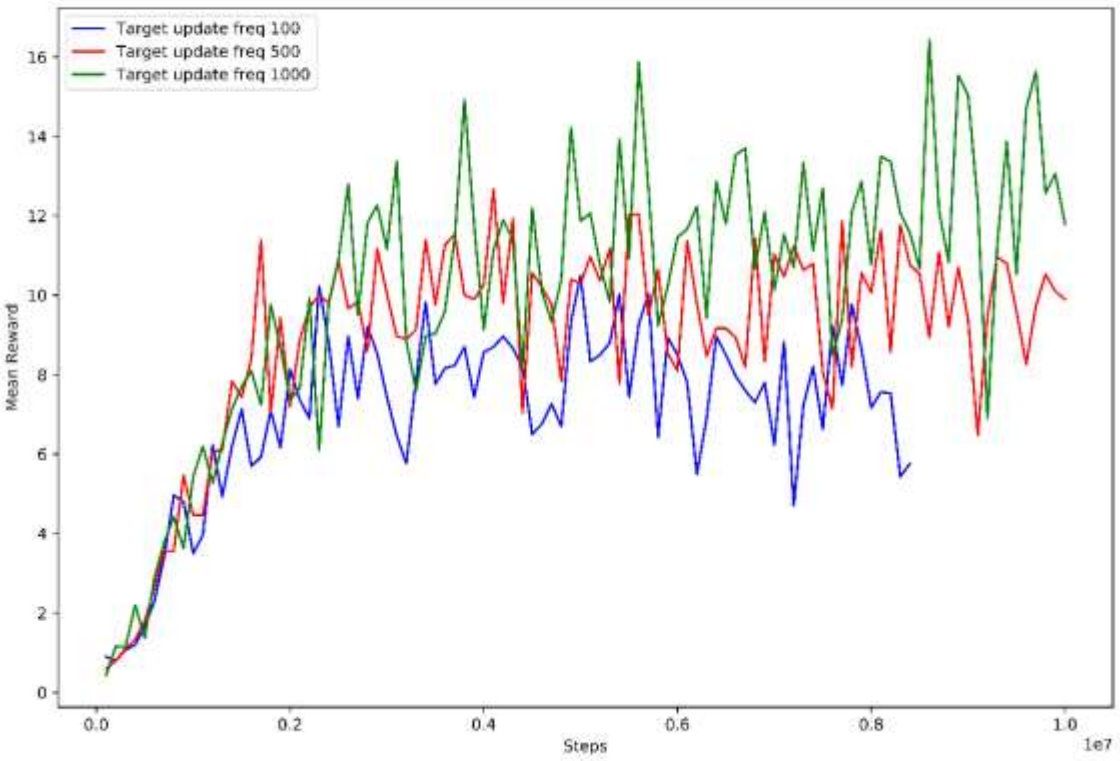
在 Pong 遊戲上做 Training 時的 learning curve 如下，x 軸為 step 數，y 軸為近 30 個 episode 的 mean reward

看起來比 DQN 穩定許多，可以看到在一千五百萬個 step 時 reward 突然上升，一千八百萬個 step 已經可以穩穩贏過電腦許多，之後 performance 便收斂

2.Experimenting with DQN hyperparameters

我選用的參數是 target network update frequency，在 DQN 的學習過程中，我們想要訓練一個好的 Q network，而其中拿來算 loss 的 target Q network 可以想像成是一個我們認為不錯的衡量方式，他是一段時間前儲存的 Q network，因此我們計算 Q(s,a)和 r+max(target_Q(s',a'))的差距。我認為這樣代表當 target update frequency 變高時 training 的速度會變快，因為 model 較快的去將 target Q network 改成對現在情況來說好的衡量方式。

原版 model 的 target update frequency 是 1000(每 1000 個 step update 一次 target Q)，我將其調成 500 和 100 來觀察出現的現象。
用 DQN 在 Breakout 遊戲上做 Training 時的 learning curve 如下，x 軸為 step 數，y 軸為近 30 個 episode 的 mean reward(reward 有 clip)，不同顏色的線代表不同的 target update frequency



出乎意料的，update 比較頻繁的 model 在前期的 reward 上升速度並沒有比較快，反而在中後期的 reward 上表現還較差，我猜測這是因為 update 的太頻繁會容易使 model 太快的尋找下一個要走的地方然後走到 local minimum，因此需要仔細衡量這個參數的意義才能獲得較好的成果，之後若有時間我也會嘗試 update 頻率更低的 model 看看。

3.Bonus on DQN

我實作了 Double DQN 和 Dueling DQN，

用 DQN 在 Breakout 遊戲上做 Training 時的 learning curve 如下，x 軸為 step 數，y 軸為近 30 個 episode 的 mean reward(reward 有 clip)，不同顏色的線代表不同的方法

可以看到 Double DQN 在前面的 step 就有明顯的 reward 上升，而 dueling 和一般的 DQN 的 learning curve 則一直都十分相近。由於 DQN 的訓練過程不太穩定，僅能看出 Double 隱約有較好的表現(大部分時間 reward 都在上面)但最後收斂的位置都差不多。