

HW5 REPORT
黃柏睿 R04725040

1. Experience Replay:

先把 episode 儲存起來，等儲存到一定的數目後，在隨機從中抽樣出來訓練。
做法就是設定一個固定的 frame number (設定成 250000)，如果當前遊戲的
number mod 設定好的 frame number = 0，就從過程中儲的抽出來訓練。

2. Target Network:

使用兩個 network，一個用來產生策略，另一個(target network)產生 Q function
的估計值，target network 通常用舊的參數，然後在經過較多的 episode 之後
才更新。

3. 加入一個 epsilon 值，代表我們下一個 action 是隨機決定的機率，這樣可以
讓 network 學到更多資訊。

程式的初始值是 1，然後會越來越小達到 min(我設為 0.1)不動，如果 random
產生的數字 <epsilon，就從 action_set 裡面隨便選一個 action 來用。

4. Clip Reward:

固定每次的 reward 只會出現 [-1, +1] 來讓 Network 更穩定。

ALE 直接給的就是 +1 了。

改為每次從 action set 裡面選出機率最大的兩個 action，然後在更新 network 時
把她轉乘 sequential 的。例如說原本是 ABACCD 三次 action，就轉成 ABACCD
六次 action，但每兩次 action 之後的 state 跟 action 是依樣的，並且轉換時誰前
誰後是用隨機決定。