學號：R08922167 系級：資工碩一 姓名：曾民君

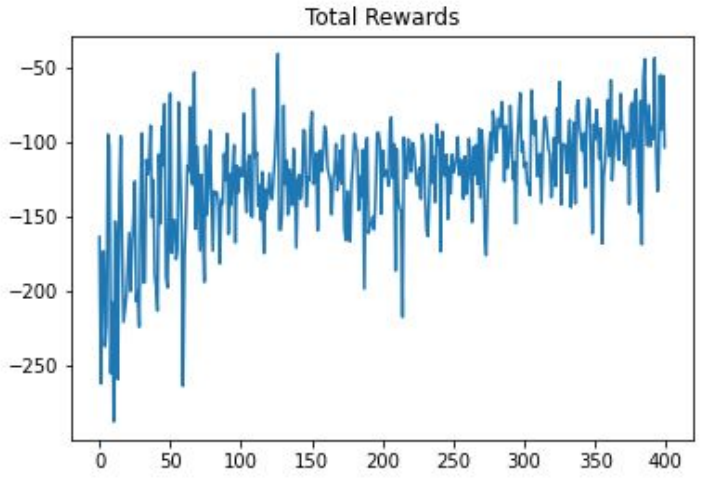
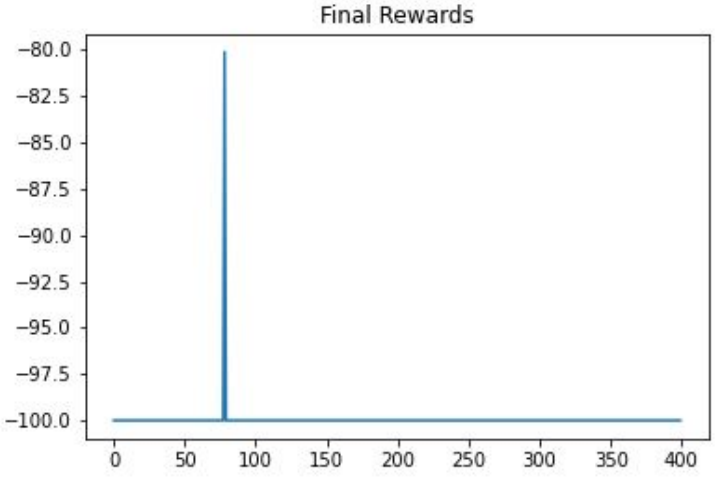
1. (20%) Policy Gradient 方法

   a. 請閱讀及跑過範例程式，並試著改進 reward 計算的方式。

   b. 請說明你如何改進 reward 的算法，而不同的算法又如何影響訓練結果？

Ans:

a. 原始的 code 表現結果

Total Rewards	
Final Rewards	

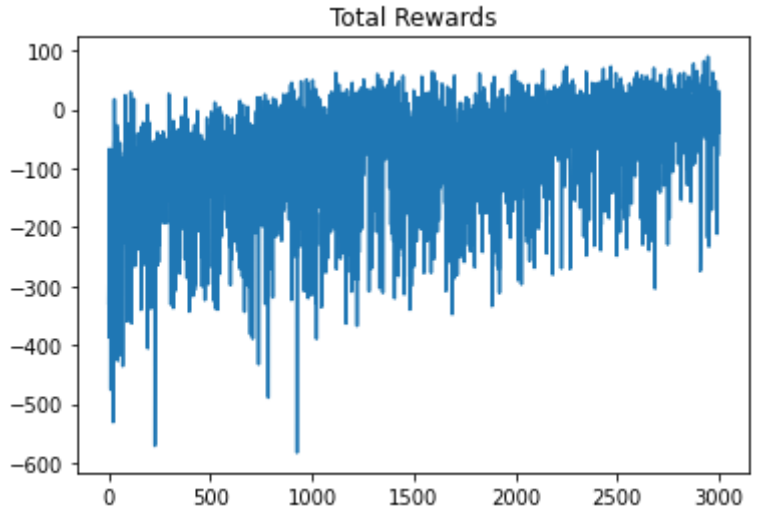
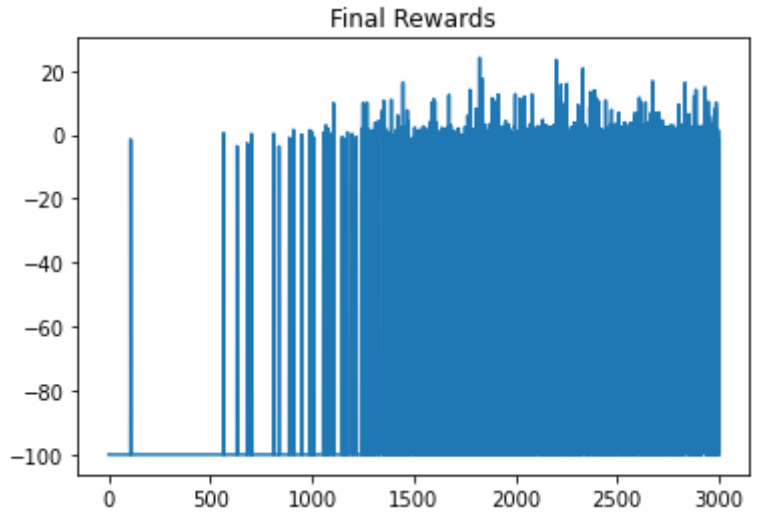
b. 這邊使用 Q Actor-Critic 方式進行訓練，其中 actor 與 critic model 皆如同助教 baseline 的 actor model, 兩著 optimizers 也如同原本 baseline 的 SGD， 訓練時演算法部份如下：

```
# Initialize here

for batch in n_batches:
      for i in count():

            1. Get action from actor and value

            2. Update env and get reward

            3. Record infors: log_porbs, values, rewards, masks

            4. Update state

            5. Break when finished


   # Compute loss for critic and actor

   # Update models
```

其中 critic 的 loss 為 TD_loss 的平方，只是每一場遊戲更新一次。而 actor 的 loss 為 TD_loss * log_probs，然後NUM_BATCH 設為3000。

已結果來看，在 Total rewards 可以比較好一點，另外 Final rewards 則可以 穩定在 0 附近徘徊，偶爾會跌到 - 170 附近。

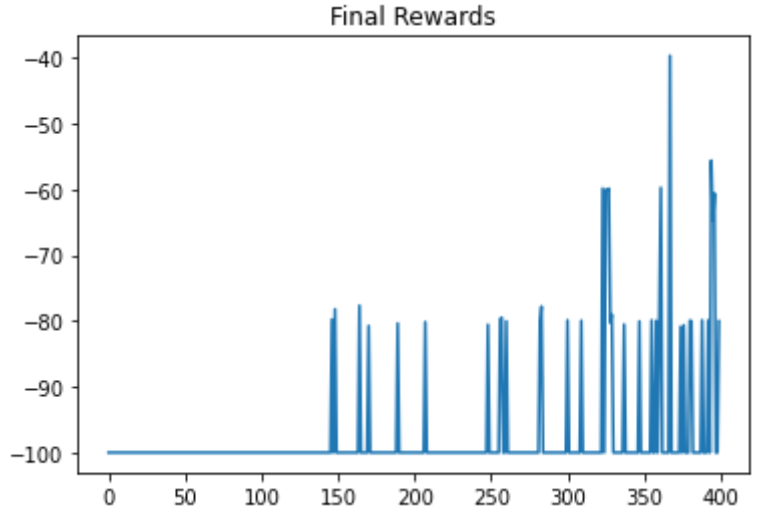
參考： https://github.com/yc930401/Actor-Critic-pytorch

Total Rewards	
Final Rewards	

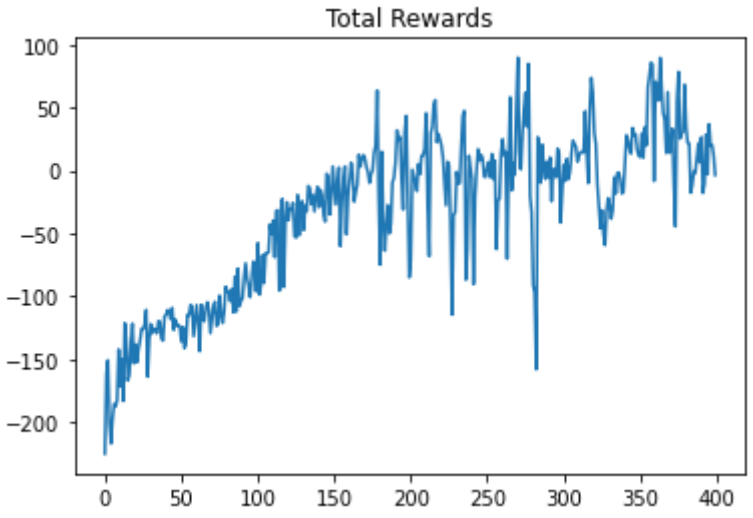
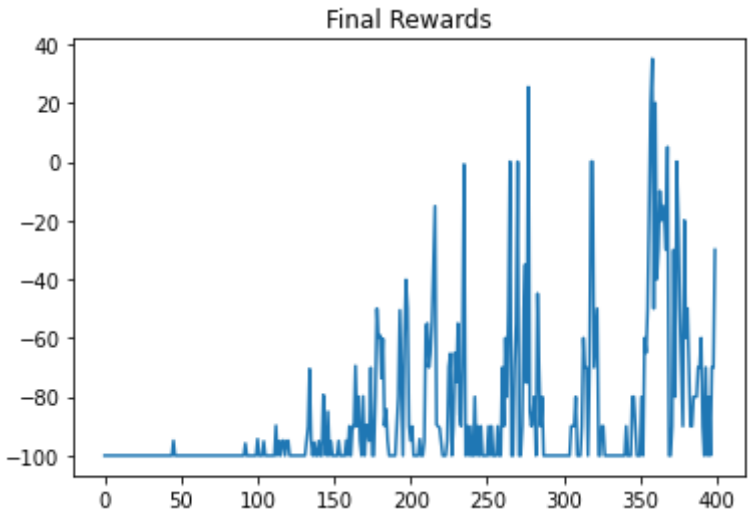
2. (30%) 試著修改與比較至少三項超參數（神經網路大小、一個 batch 中的回合數等），並說明你觀察到什麼。

   Ans: baseline model (比較對象)，結果同 1-a 的圖，為助教原本的程式，而分別改動以下幾個參數後的結果：

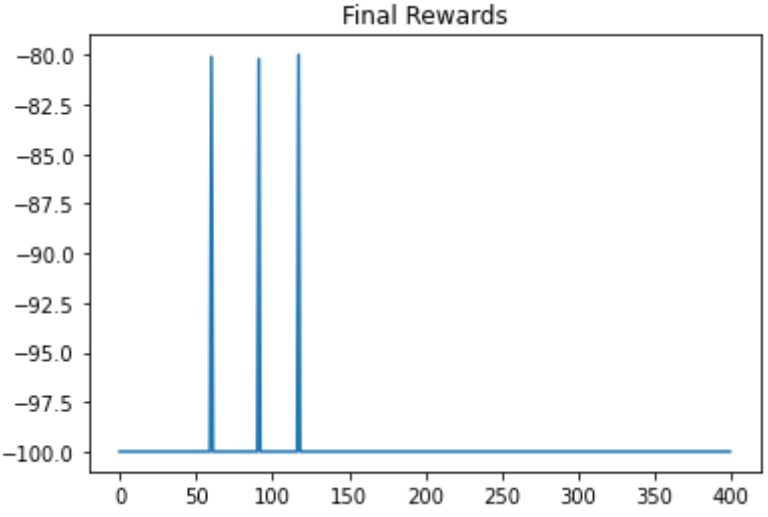
   a. model 原本 3 層的 linear layers 的 input size: 8, 16, 16，改成 8, 128, 256，已結果看來增大 model 是會有比較好的表現，但是 Final Rewards 依然是負的，可能原因是每一次遊玩都更新可能會往不好的地方更新。

Total Rewards	
Final Rewards	

b. batch 中的回合數 從原本 5 改成 20，由於改動 model 依然表現不足，所以第二個變數想要比較 batch 中的回合數，若提升之後會不會使得每次更新都會比較有效率，而已結果來看，的確能夠改善 只改 model 無法解決的問題

Total Rewards	
Final Rewards	

c. Optimizer 由 SGD 改成 ADAM，想看說同樣在 400 次更新情況下，這兩個 optimizer 收斂程度的差別

Total Rewards	
Final Rewards	

另外有嘗試將 model 中的 tanh 改成 relu，但train 到一個程度後，表現會突然爛掉，然後就爬不起來。

3. (20%) Actor-Critic 方法

    a. 請同學們從 REINFORCE with baseline、Q Actor-Critic、A2C 等眾多方法中擇一實作。

    b. 請說明你的實做與前者（Policy Gradient）的差異。

Ans: 選擇實做 Q Actor-Critic，實做的更新策略是以 Temporal-Difference 為更新方式，是以每一次遊戲後再一起稱新 actor 與 critic，但這部份有個缺點是不太能保證每次更新為有效更新，從訓練過程中的數據圖來看來回震盪幅度非常大，且所需運算量會是只有 Actor 的兩倍多。其餘訓練中的細節與 1-b 那邊的敘述相同。

4. (30%) 具體比較（數據、作圖）以上幾種方法有何差異，也請說明其各自的優缺點為何。

Ans: 數據與作圖部份，由於前面幾題都已經有附上訓練過程數據圖，所以這邊就直接已文字表格形式列舉出以上方法的比較 ~

Model	Total reward	Final reward	優點	缺點
Actor	最佳只能到 - 40左右	幾乎都是在 -100	Train 最快	整體表現最差
Q Actor-Critic	最佳只能在 90 ~ 100 之間	會維持在 0 附近	Total reward 表現最好	Train 最久， 表現振幅較大
Actor改	最佳能到 30 左右	比 baseline model 有更多 分數高於 -100	表現比 baseline 好許 多	表現不足保證 遊戲 reward 能高於 0
Actor改 搭配 20 場平均	最佳能到 90 左右	分數表現Actor 改還要好	每次更新最有 效率，表現振 幅較小	Train 最久

另外有嘗試將 Q Actor-Critic 搭配20 場平均更新一次，但尚未釐清 train 失敗的原因