Lab6

機器人碩 309605014 王麒睿

1.

LunarLander-v2 (action離散) episode rewards



2.

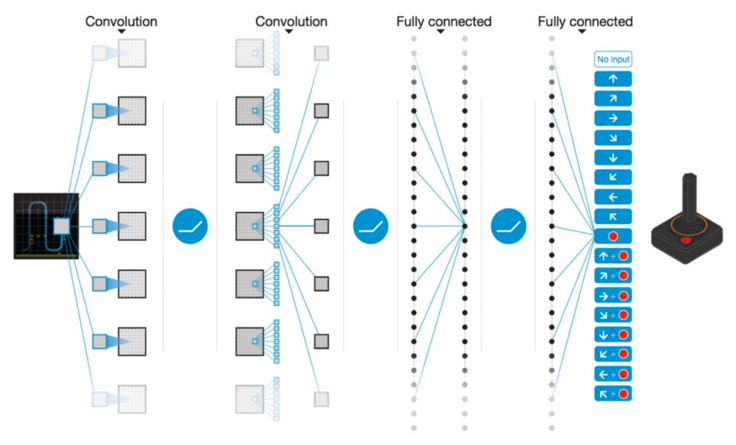
LunarLanderContinuous-v2 (action連續) episode rewards

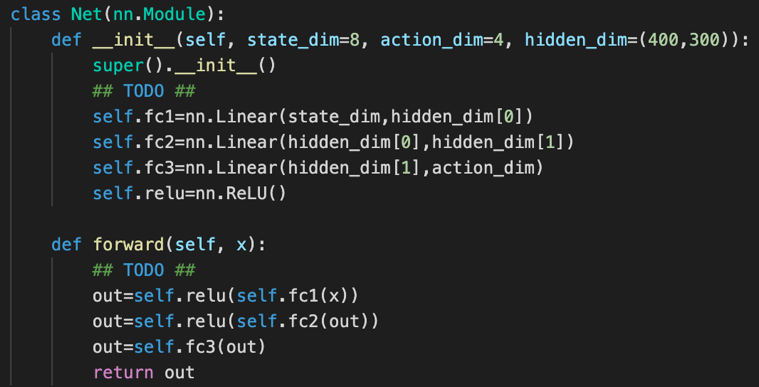


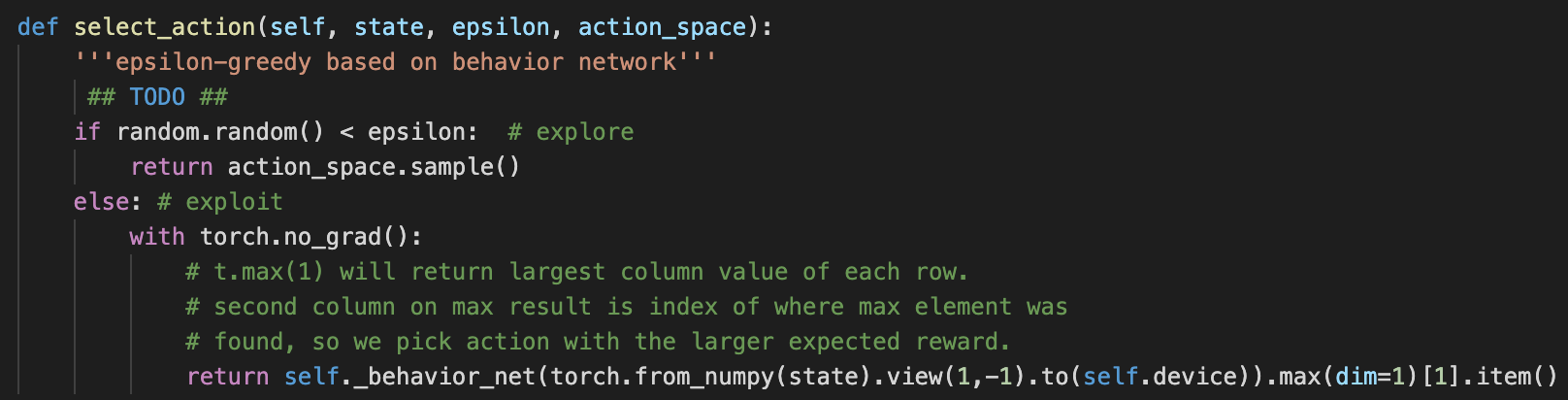
3. Describe your major implementation of both algorithms in detail

DQN:

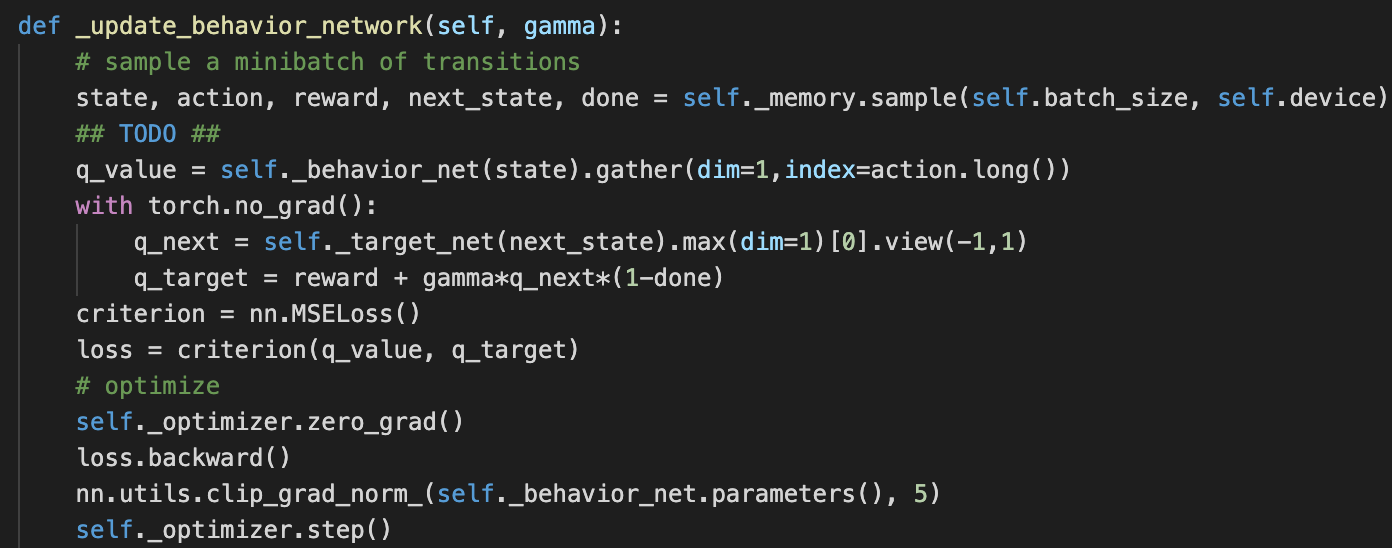
建立一個network來預測Q(s,a)的value，這裡的action4有種可能(No-op,Fire left engine,Fire main engine,Fire right engine)，所以網路最後一層為4個neuron



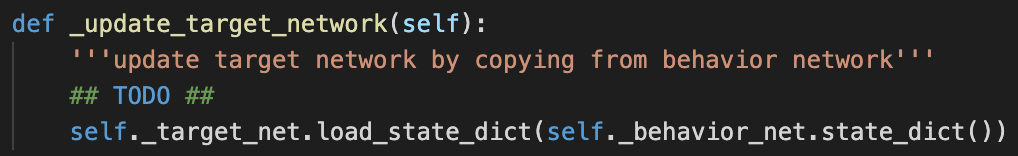


在episode中(玩遊戲的過程中)，選擇最大Q(s,a​i​)的a​i​或者有一定的機率ε隨機選擇action (called ε-greedy) 

update network的方法是由replay memory中sampling一些遊戲的過程: (state,action,reward,next\_state,done)來做td-learning，再用Q(s,a)與r + gamma \* max​a’​Q’(s’,a’)的差做MSELoss

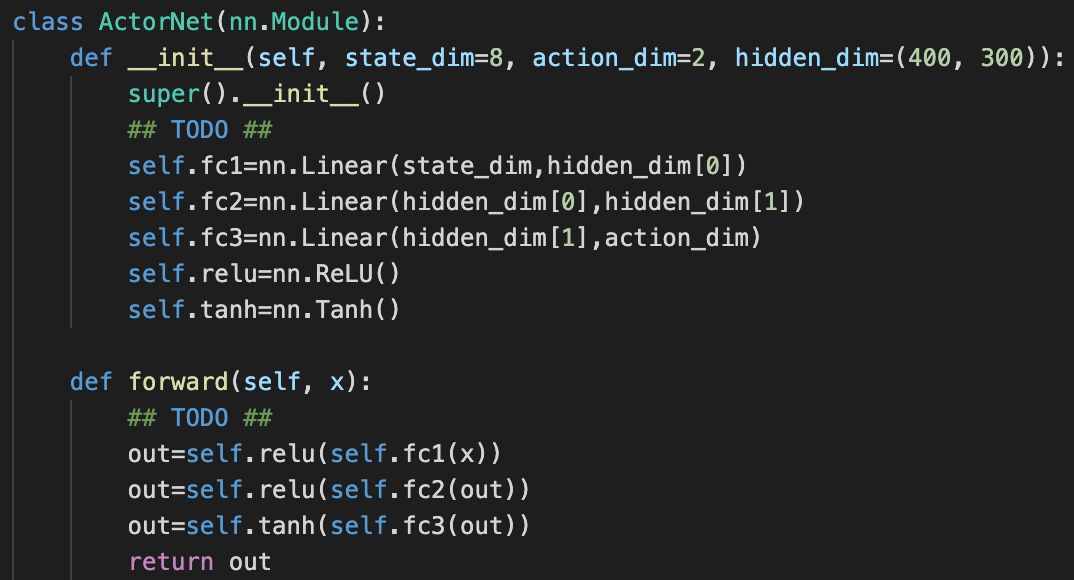


每隔一段時間，就用behavior network取代target network

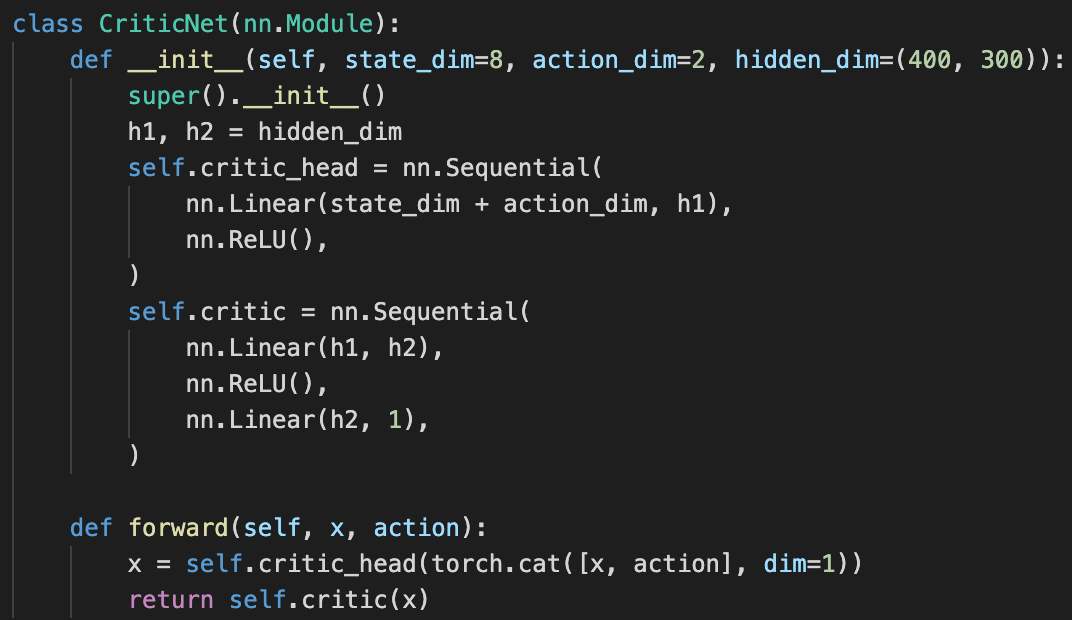


DDPG:  
建立一個可以依據目前state決定要執行哪個action的Actor Network, 由於action有2個(Main

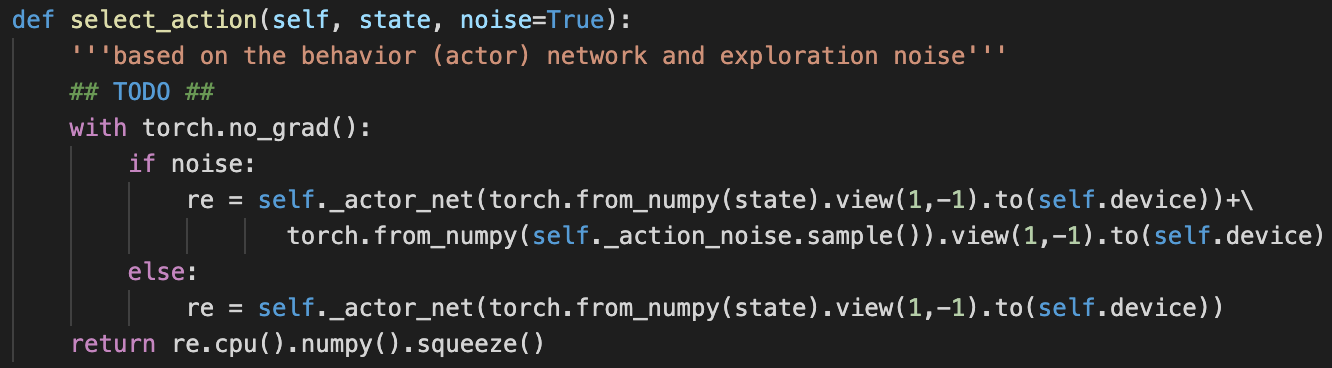
engine:-1~+1,Left-right-engine:-1~+1)，所以最後一層有2個neuron



建立一個可以預估Q(s,a)的Critic Network，由於輸出的是一個純量，所以最後一層neuron數 為1

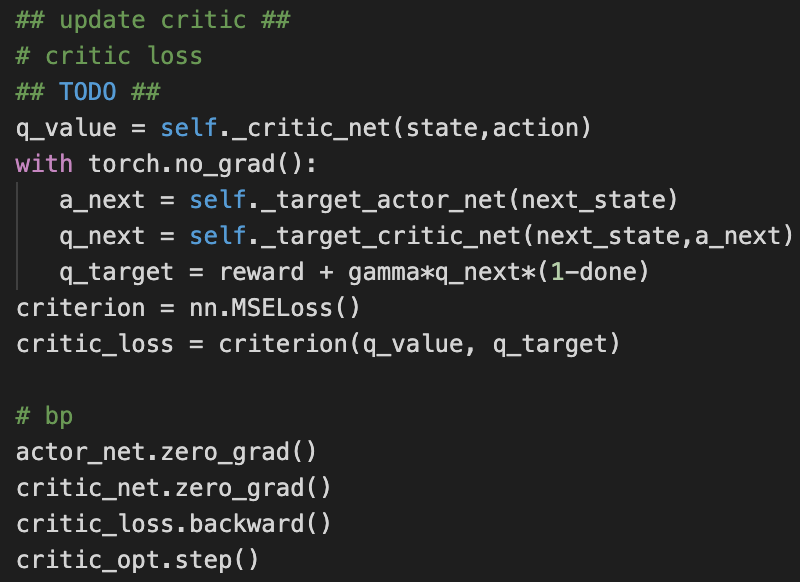


在episode中(玩遊戲的過程中)，由Actor Network選擇action並加上一noise



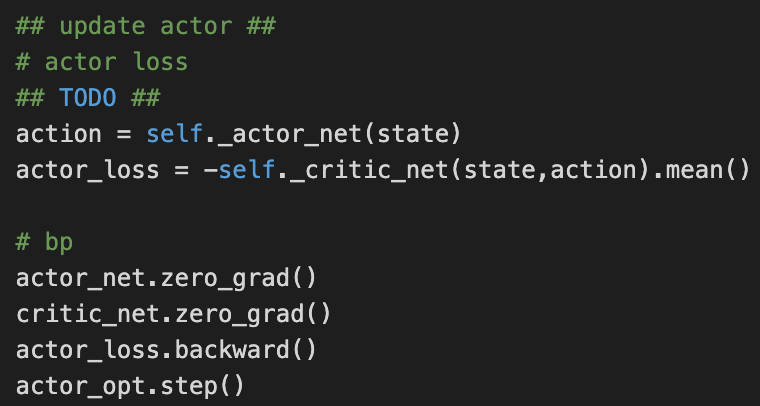
在episode中(玩遊戲的過程中)，也要更新Behavior的Actor Network μ, Critic Network Q, Target的Actor Network μ’, Critic Network Q’。

利用Target Network生出的q\_target與Behavior Network生出的q\_value做MSELoss更新Q。



利用Behavior Network的Actor Network μ與Critic Network Q可以求出Q(s,a)，我們想要更新μ 來使輸出的Q(s,a)越大越好，因此定義

Loss Value = E[-Q(s, μ(s))]，並透過backpropagation更新。



4.

Describe differences between your implementation and algorithms

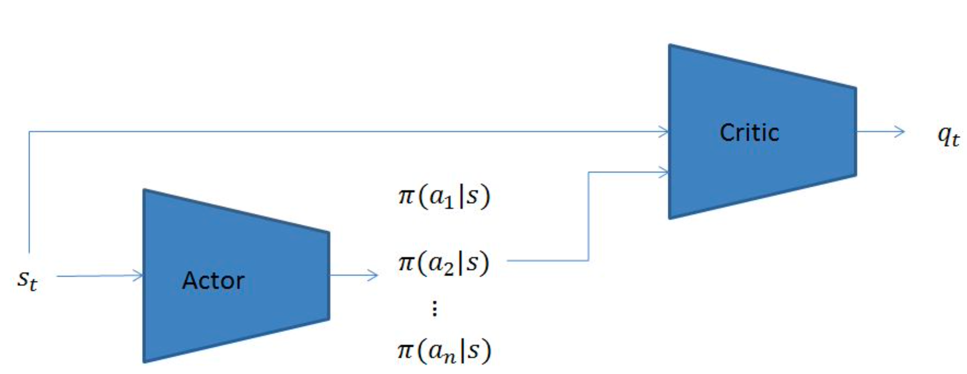
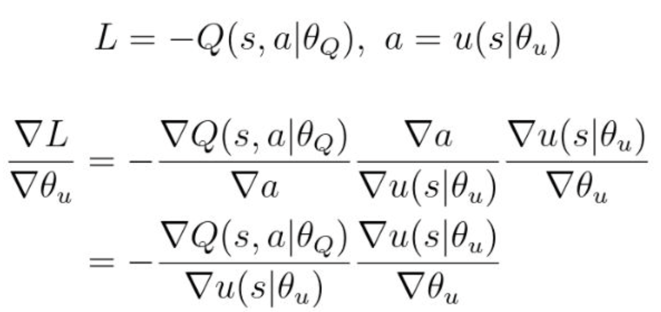
在training的時候，一開始會有一段warmup的時間，在這段時間中，不會去update network的 參數，只會隨便亂玩(隨機選擇action)，並把遊戲過儲存到replay memory裡。另外在DQN的 部份，並不是每個iteration都要更新Behavior Network，而是每隔一段時間(Ex: 4個iteration)才 會更新一次。

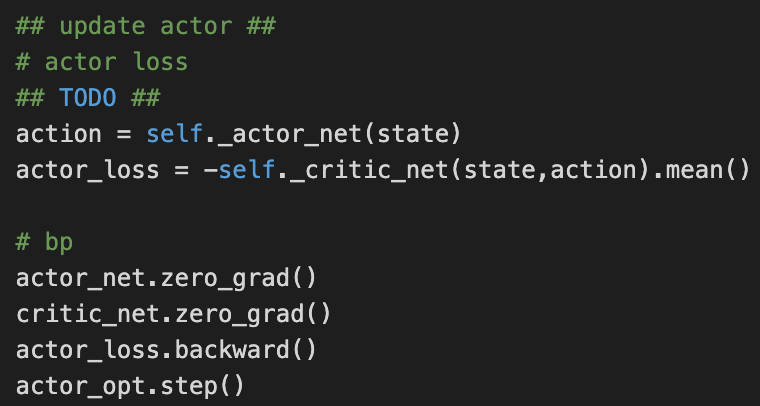
5.

Describe your implementation and the gradient of actor updating in DDPG

利用Behavior Network的Actor Network μ與Critic Network Q可以求出Q(s,a)，我們想更新 Actor Network μ來使輸出的Q(s,a)越大越好，

因此定義Loss Value = -Q(s, μ(s))，backpropagation的時候不更新Critic、只更新Actor。

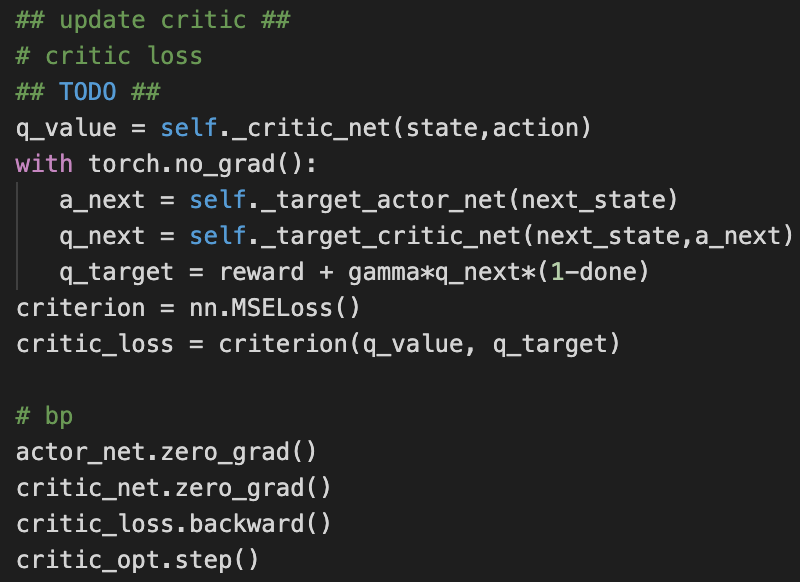


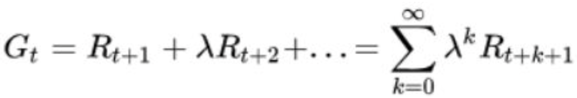


6.

Describe your implementation and the gradient of critic updating in DDPG

利用Target Network生出的Q​target​與Behavior Network生出的Q(s,a)做mean square error來更 新Q Network。



7. Explain effects of the discount factor 

λ就是discount factor，意思就是說，越是未來所給的reward影響是越來越小的，當下的 reward是最大的

8.

Explain benefits of epsilon-greedy in comparison to greedy action selection

我們必須在explore與exploit之間取得平衡，因此在greedily choosing action的基礎上，必須偶爾選擇其他的action來explore那些未知但可能是最佳的action。

9. Explain the necessity of the target network

有Target Network與Behavior Network的搭配可以使training的時候更穩定， 因為生出 Q\_target的Target Network每隔一段時間才會改變一次。  
10.

Explain the effect of replay buffer size in case of too large or

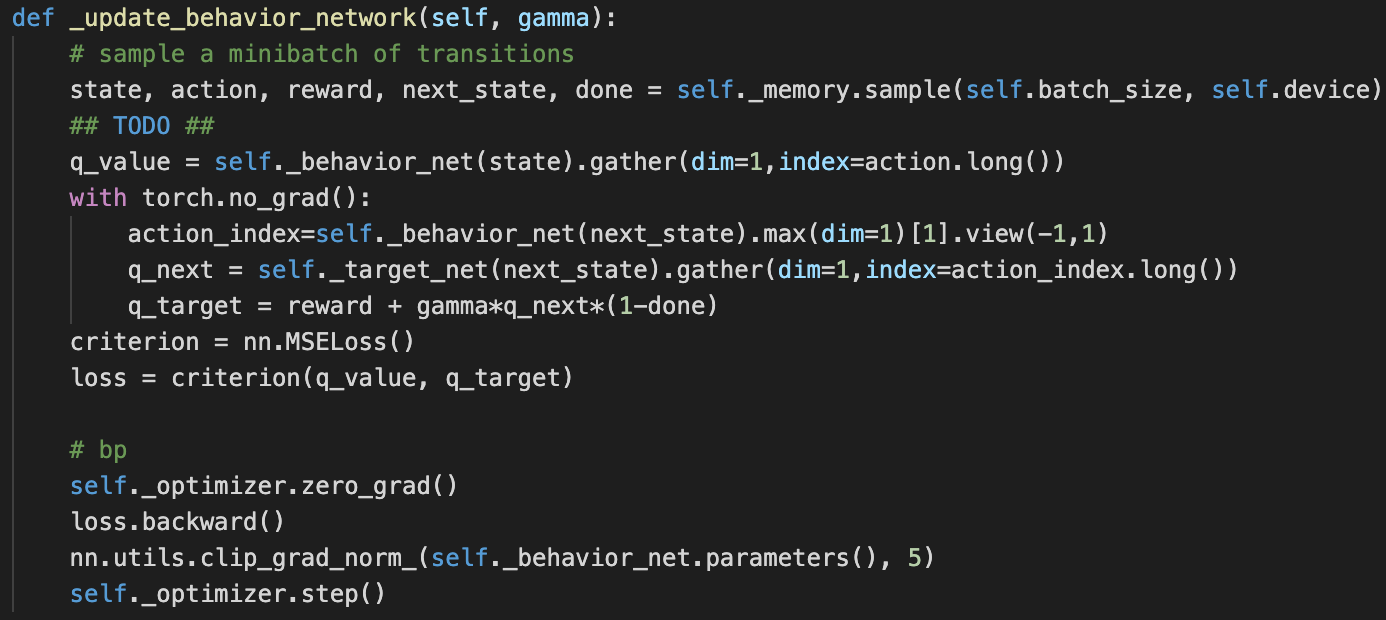
too small

如果replay buffer size越大，training過程可以更穩定，但會降低training的速度。如果replay buffer size越小，會一直著重於最近玩的episode的狀況，容易造成overfitting、甚至整個train 壞掉。

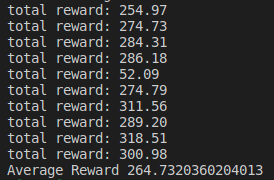
Bonus

1. Implement and experiment on Double-DQN

DDQN與DQN其實差不多，就只差在update Behavior Network時是如何決定q\_target的， DDQN在決定q\_target時，不是直接取max Q’(s,a​i​)，而是用Q(s,a​i​)中最大值的i作為t查找 Q’(s,a​i​)的index。



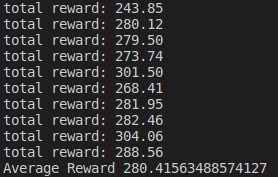
train出來的結果跟DQN差不多。(train 2000個episode)



**Performance**

1.

[LunarLander-v2] Average reward of 10 testing episodes: Average ÷ 30 DQN: 2000 episode



2.

[LunarLanderContinuous-v2] Average reward of 10 testing episodes: Average ÷ 30 DDPG: 2000 episode

