學號：B06902042　系級：資工三　姓名：劉愷為

1. (2%) 試說明 hw6_best.sh 攻擊的方法，包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何？如何影響你的結果？請完整討論。

   我使用 densenet121 作為我的 proxy model, 將 epsilon 設成 0.1。用 cross_entropy 來算我的 loss。使用的攻擊方法依然是 fgsm, 在 Judge Boi 上的結果是 Success Rate 為 0.980, L -infinity 為 5.6250。

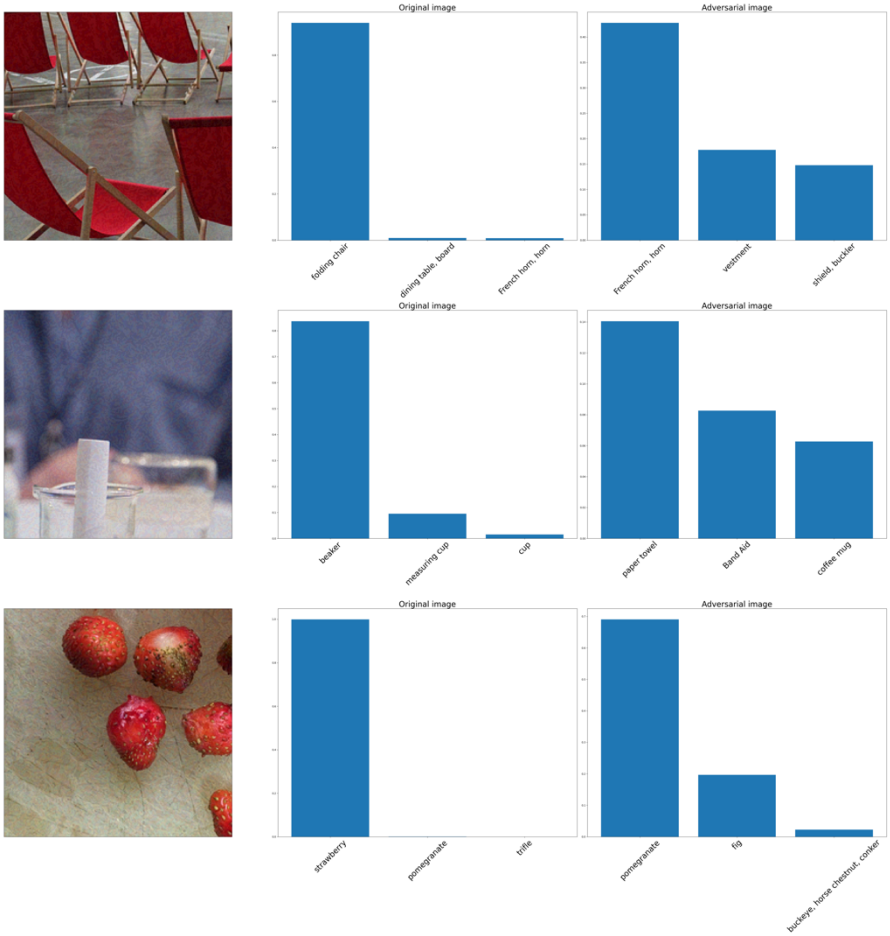
   實作過程當中，最後再輸出圖片時，我一開始沒有加上 img[img<0] = 0, img[img>1] = 1 這兩行。造成我的 L-infinity 達到 200 多。

2. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

   由於我太晚開始寫了只有上傳一點點的 proxy model

Proxy Model	Success Rate	L -infinity
Densenet-169	0.54	5.57
Densenet-121	0.98	5.65
Resnet-50	0.37	5.42

3. (1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分



別取前三高的機率)。

4. (2%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

我的方法是在Adverdatatset的__getitem__ 當中用ImageFilter加上GaussianBlur。而在上傳judge 之後可以發現success rate 從防禦前的0.98降到了0.54。GaussianBlur可以達到濾除雜訊、模糊化圖片的效果，因此可能就改變了原本圖片在被攻擊的gradiant的資訊，進而達到好的防禦效果。