學號：B06902006　　系級：資工三　　姓名：王俊翔

1. 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。

   我的 hw6_best.sh 使用的 proxy model 為 DenseNet121，方法類似原本的 FGSM，不過我將 epsilon 調小(0.02)，加上我算出其 gradient 後，iterate 至多 30 次 (直到 predict 出不同的 label，通常兩三次就成功了，仍會出現錯誤的原因是有些值轉成 image 後 <0 或 >1 我做完後才調整)，也就是老師投影片上面的 basic iterative method (https://arxiv.org/abs/1607.02533)中的第二個方法，另外 loss 使用的是 cross entropy(雖然是沒差，只看 sign)，效果好非常多，尤其是可以大幅減少 L-inf norm 的值。

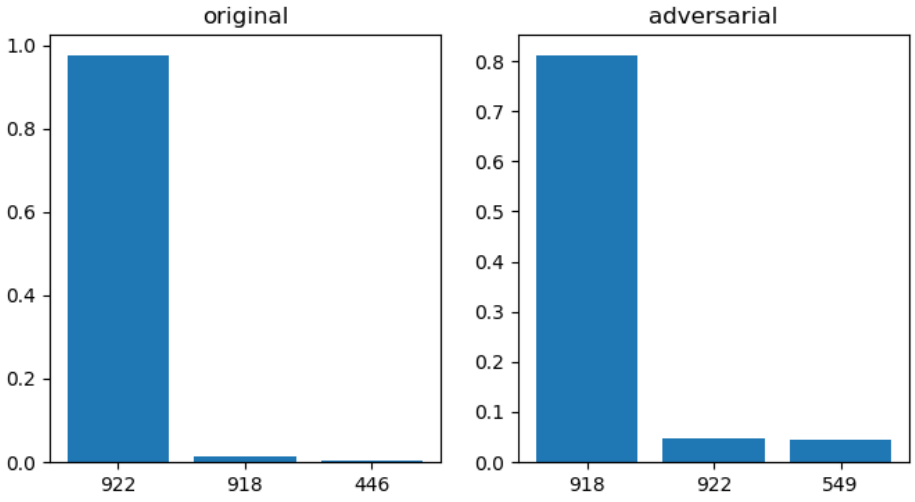
2. 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

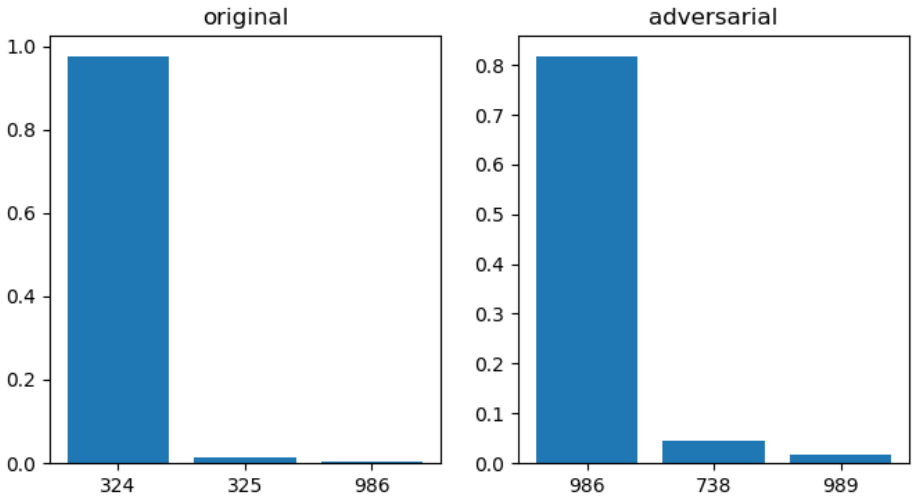
   背後的 black box 最有可能是 DenseNet121。

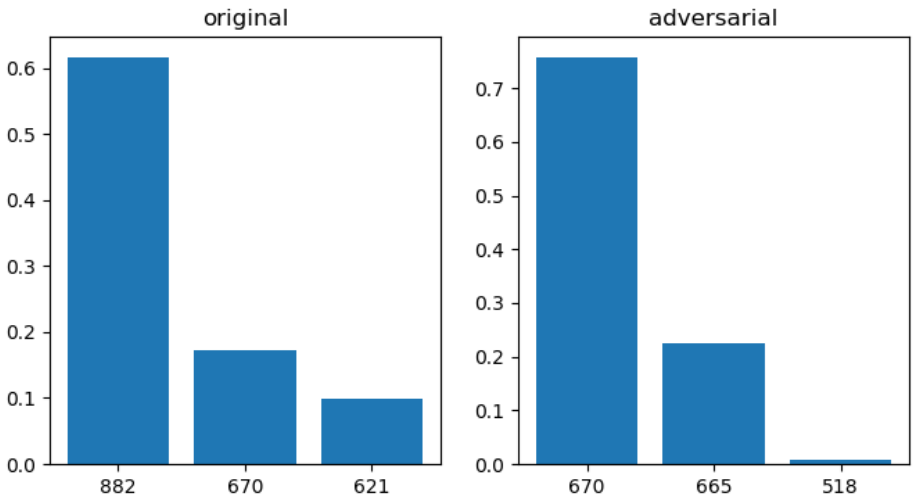
	Success Rate A	Success Rate B
Actually on Website	0.4	0.785
VGG16	0.36	X
VGG19	0.4	0.925
ResNet50	0.875	X
ResNet101	0.375	X
DenseNet121	0.4	0.785
DenseNet169	0.345	X

   我丟了兩次上去，發現成功率為 40%及 78.5%，代表使用某個 model 預測的 label 與原 label 不同的有 40%及 78.5%，又投影片說只會是上述六種 pretrained model，所以很明顯答案就是兩次都 match 的 DenseNet121。

3. 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)。







第一張圖 original 922 的機率為 0.9776，adversarial 918 的機率為 0.8122；第二張圖 original 324 的機率為 0.9766，adversarial 986 的機

率為 0.8181；第三張圖 original 882 的機率為 0.6167，adversarial 670 的機率為 0.7583。(皆過 softmax 後得出)

4. 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。

   這裡我使用的是 gaussian filtering 實作為使用 opencv 的 GaussianBlur，mask 設為 5*5，防禦前的 success rate 為 100%，防禦後為 74%，且 L-inf norm 衝到 100 多，原本為 2.92。我認為成效算是還可以，gaussian filter 藉由平均(可能有加權)附近的 pixel 來得到後來的 pixel，可以改變每個 pixel (feature)的值讓他們表現得更平滑不會突兀，不過因為增加了奇奇怪怪的雜訊 l-inf 衝高也是正常的。