Machine Learning HW5 Report

學號：B06901087　系級：電機二　姓名：翁瑋襄

1. (1%) 試說明 hw5_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。(依內容完整度給分)

   我的hw5_best.sh寫的是non_targeted iterative FGSM attack，使用的proxy model是resnet50，我設的epsilon = 0.4/255/0.23 、 epoch = 5，success rate = 0.995，L-inf = 2.0000，我認為跟hw5_fgsm.sh兩個最大的差異點在hw5_fgsm.sh只有一個epoch，調的L-inf = 10.0000，用的proxy model是vgg16，success rate = 0.405。

2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。

	hw5_fgsm.sh	hw5_best.sh
proxy model	vgg16	resnet50
success rate	0.405	0.995
L-inf.	10.0000	2.0000

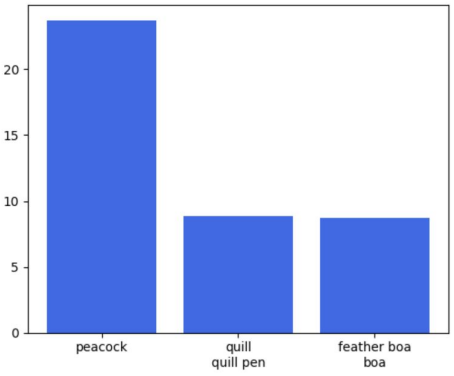
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

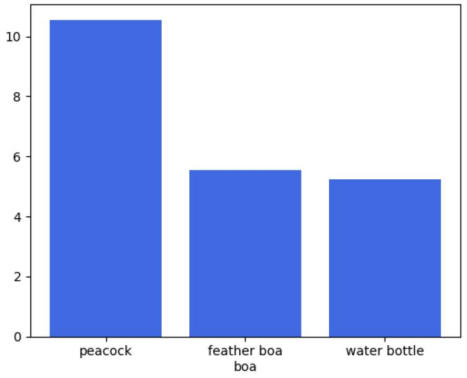
model	sucess rate	L-infinity
vgg16	0.095	2.0000
vgg19	0.105	2.0000
resnet50	0.995	2.0000
resnet101	0.260	2.0000
densenet121	0.170	2.0000
densenet169	0.145	2.0000

   我認為是resnet50，因為其他的model success rate都沒有超過0.5，唯獨resnet50一支獨秀，達到0.995的success rate。

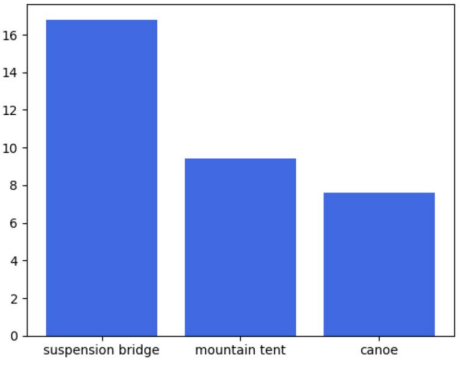
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分

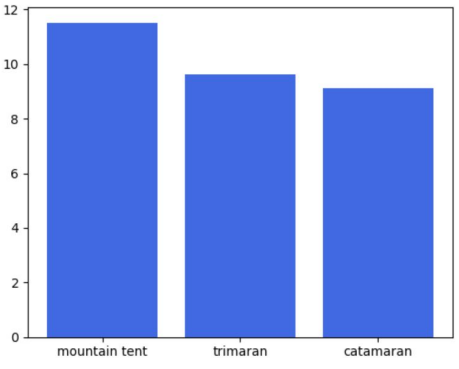
別取前三高的機率)。



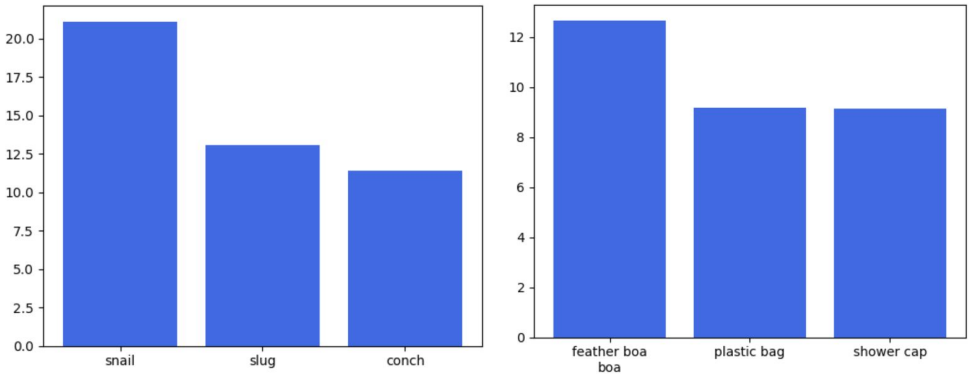


這張孔雀的圖是我準確率0.995唯一attack不成功的一張圖，即便L-inf已經達到2.0000，仍然因為孔雀的羽毛特色太鮮明導致還是很難讓機器誤判。





第二章圖是吊橋，經過attack後，成功讓機器誤判為第二高的類別，由於本身是non-targeted attack，吊橋被預測的機率甚至掉出前三名。



上圖原本被歸類的是蝸牛，經過attack後，竟然被誤判為羽絨披肩，這個是最令我訝異的一個，竟然能讓每個pixel都改2以內就讓機器誤判為另一個完全不相關的物品種類。

5. (1%) 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，附上你攻擊有無的 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。

我使用的是PIL.ImageFilter.SMOOTH()，實作的原理好像是類似lowpass filter，把高頻的雜訊濾掉，達到smooth的效果。

	沒有smooth	有smooth
origin image	0.00	0.08
attack image	0.995	0.9

我發現加上SMOOTH後，圖片的邊緣有稍微變模糊的趨勢，可能也是因為這個原因，能稍微把attack加入的雜訊稍微去除掉。