Machine Learning HW5 Report

學號：B06901063 系級：電機二 姓名：黃士豪

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)

   在 hw5_best.sh 中，我使用的是在 fgsm 中得到最好的 model resnet50，但跟 fgsm 不同的是，我將每一次進行加減的 learn rate 調低，卻在每次動作之後重新 predict，檢測是否跟最一開始的結果相同，若仍相同就再移動一步直到 model 做出錯誤的判斷。此作法雖然使得 L-inf. norm 較大，但在本機上能達到 100%的攻擊成功率，即使是重新存圖時因浮點數問題使得在 judge 上成功率下降，仍能達到極高的攻擊成功率。

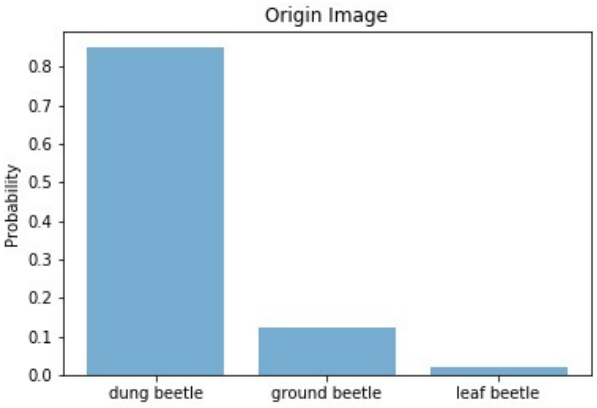
2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。

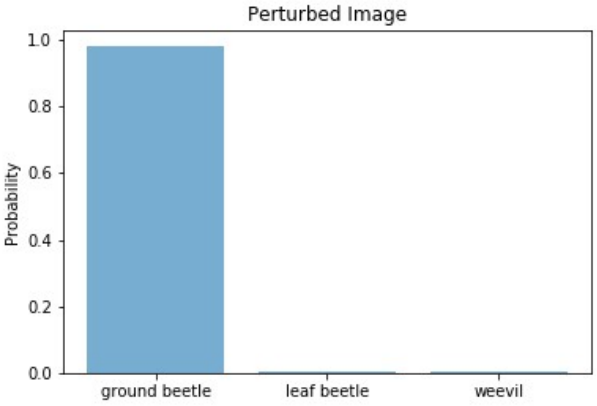
	hw5_fgsm.sh	hw5_best.sh
proxy model	resnet50	resnet50
success rate	0.910	0.975
L-inf. norm	2.7500	4.5200

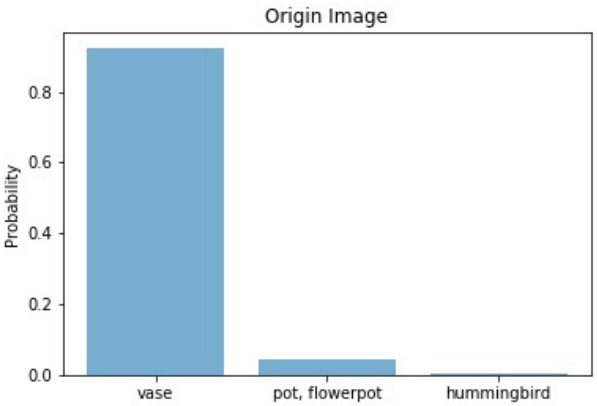
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

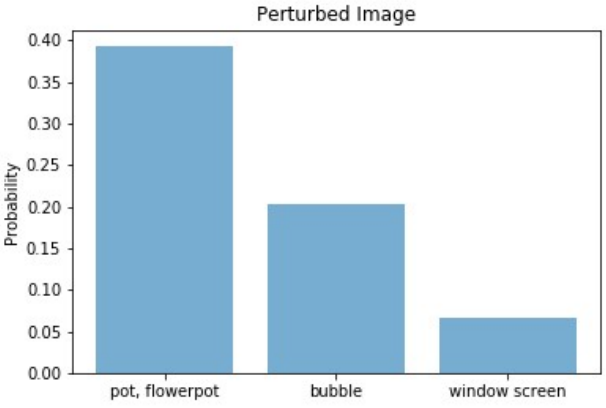
   我推測背後的 black box 就是 resnet50，因為在純 fgsm 的情況下，只有 resnet50 這個 proxy model 可以在本機和 judge 上都得到同樣的攻擊成功率，其他 model 都只能大概得到 60%的成功率。拿 vgg16 來說，我用來過 simple baseline 的 model 就是 vgg16，因為我那時候 epsilon 設極大(約 10)，因此在本機上的攻擊成功率達到了 100%，但丟上 judge 上後卻發現得到 0.635 的 success rate 和 10.9500 的 L-inf. norm，扣除浮點數的誤差，很明顯此 black box 的 model 並非 vgg16。一直試到 resnet50 才有明顯的進步。

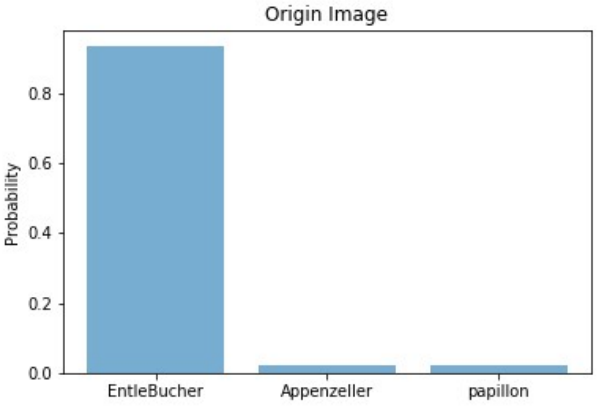
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

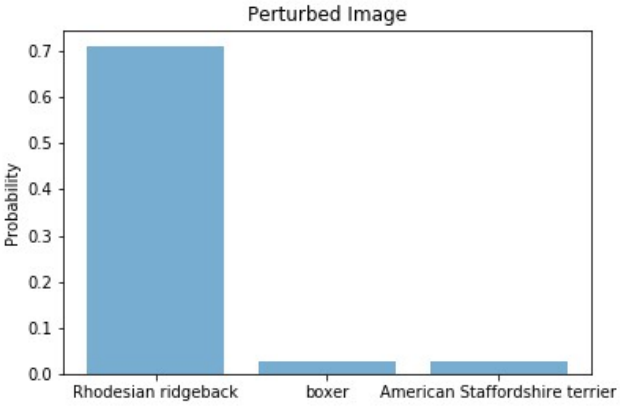







Perturbed Image
Probability
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
pot, flowerpot
bubble
window screen


Origin Image
Probability
0.8
0.6
0.4
0.2
0.0
EntleBucher
Appenzeller
papillon


Perturbed Image
Probability
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Rhodesian ridgeback
boxer
American Staffordshire terrier

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。

	origin	ImageFilter.SMOOTH
success rate	0.975	0.835
L-inf. norm	4.5200	77.1300

此 filter 可以突出圖片的大區域與低頻部分，主要減少圖片的高頻干擾，使圖片亮度變平緩，且梯度減少。用此功能可以大幅度減少 attack 造成的不均勻色塊，使得圖片更加不容易被辨認錯誤。可發現下圖比起前一題明顯模糊了許多。