Machine Learning HW5 Report

學號：B04901147　系級：電機四　　姓名：黃健祐

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)

   hw5_best.sh 使用的 proxy model 為 ResNet-50，方法則是 iterative FGSM，$\epsilon = 0.0015$，最大重複次數為 25。Iterative FGSM 是 FGSM 的進階版，對於一張圖片反覆進行 FGSM 直到成功使 proxy model 混淆（若已達到最大重複次數仍未成功則放棄），因此能夠使得 success rate 大幅提升，不過因為是反覆進行 FGSM，為了避免 L-inf norm 過大，參數也必須更謹慎地挑選。

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)。

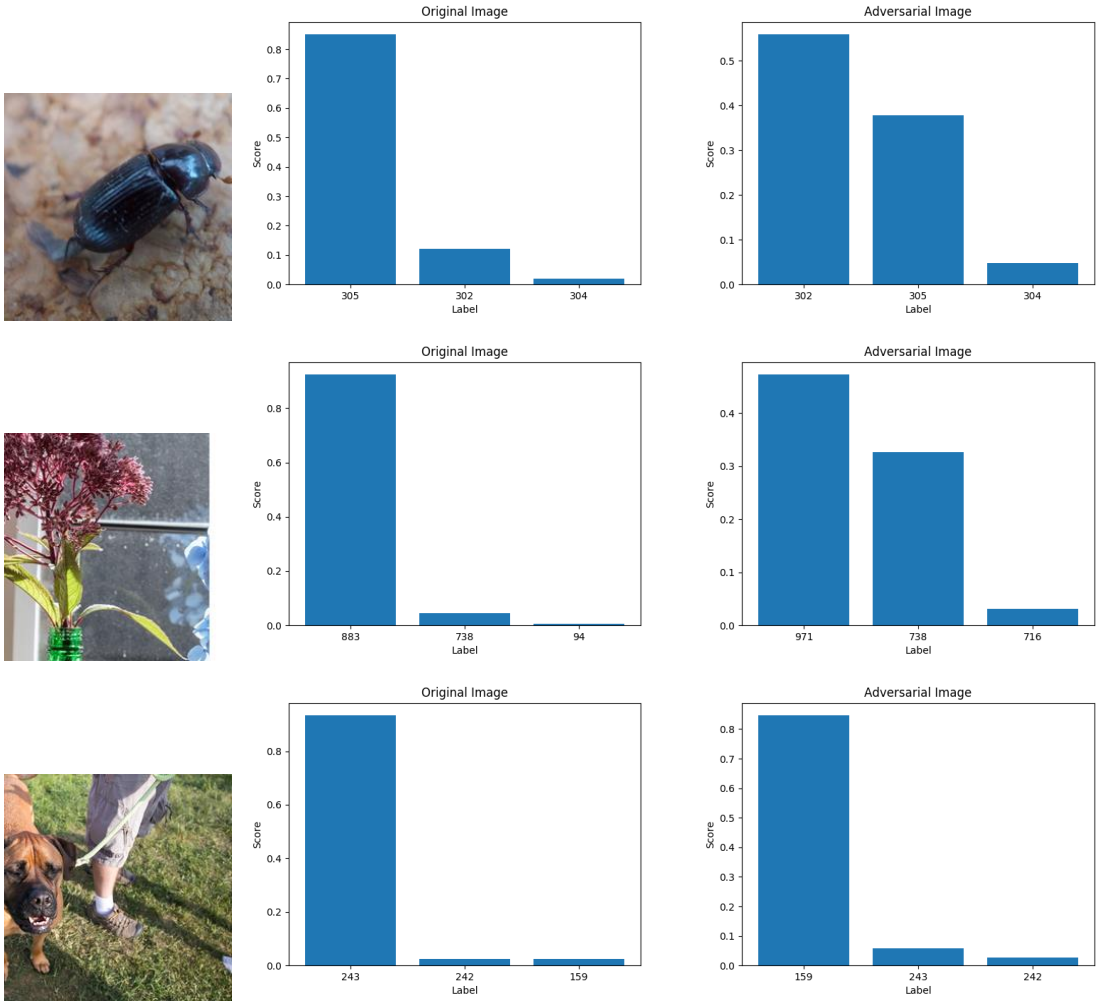
	Proxy Model	Success Rate	L-inf Norm
hw5_fgsm.sh	ResNet-50	0.730	1.0000
hw5_best.sh		0.975	2.2700

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。

Proxy Model	ResNet-50	ResNet-101	VGG-16
Success Rate	0.975	0.095	0.040
Proxy Model	VGG-19	DenseNet-121	DenseNet-169
Success Rate	0.040	0.070	0.055

由上表可以明顯看出 ResNet-50 的 success rate 遠大於其他的 proxy model，因此我推測本次作業所使用的 model 為 ResNet-50。

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)。



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。

本次實作中使用 median filter（大小為3×3）來進行 smoothing。經過 smoothing 後的 success rate 為 0.365，和原來的結果相比下降許多，說明了這種方法能夠達到相當程度上的防禦。然而，經過 median filter 處理過的 image 會變得較為模糊，是這種方法的一個缺點。