Machine Learning HW5 Report

學號：R07922104　　系級：資工碩一　　姓名：林傳祐

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)

   1. Resnet50
   2. Least-Likely-Class Iterativa Method (LLCIM)

   $$\mathbf{I}_{\boldsymbol{\rho}}^{i+1} = \mathrm{Clip}_{\epsilon}\left\{\mathbf{I}_{\boldsymbol{\rho}}^{i} + \alpha\ \mathrm{sign}(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_{\boldsymbol{\rho}}^{i}, \ell)\right\}$$

   3. Alpha=1,epsilon=0.08
   4. FGSM 是一次走一大步。而 LLCIM 是 Basic Iterative Method(BIM)的一種，BIM 就是將 FGSM 的一大步，改為每次走一小步，然後 iterative 多次來達到同樣的效果。LLCIM 顧名思義就是往機率最低的 Class 梯度方向走。在 Pytorch 中只要 torch.min(output, 1)[1]就能完成。
   5. 影響我認為分為兩個部分。第一個是，透過多次走一小步，會比一次走一大步增加更多過程中的變異。第二個是，選擇往機率最小的類別走，能與原本預測有更大的差別。

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)。

	normalization	Proxy model	Success rate	L-inf. norm
FGSM	Mod 255	Resnet50	0.475	2.0000
BEST	同 Resnet50	Resnet50	0.990	5.0000

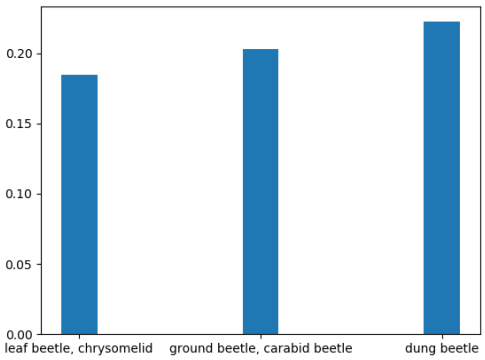
註:Resnet50 的 normalization 方法 ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]) (Mean 與 Variance)，這個也會影響結果很多!

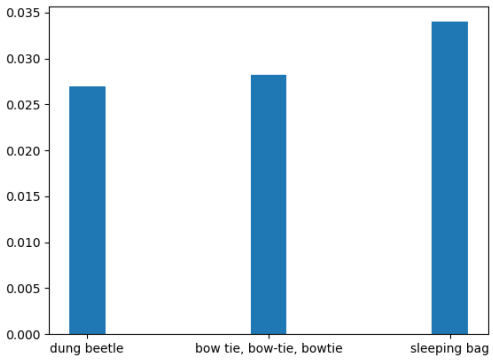
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

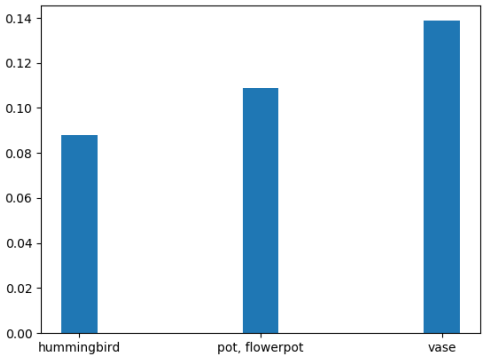
	Resnet101	DenseNet121	DenseNet169
Success rate	0.160	0.150	0.145

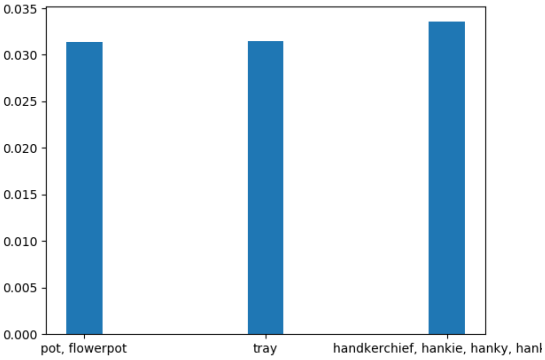
這是與我的第一個方法比較(FGSM,Resnet50,SR=0.475)，可以看出上面三個 SR 都差不多且很低，因此我猜測 Black Box 應該是 Resnet50。另外一點是，在 Training data 上觀察，FGSM 都能使每個方法 SR 很高，但只有 Resnet50 在 Testing data 能維持，更讓我確定 Black Box 就是 Resnet50。

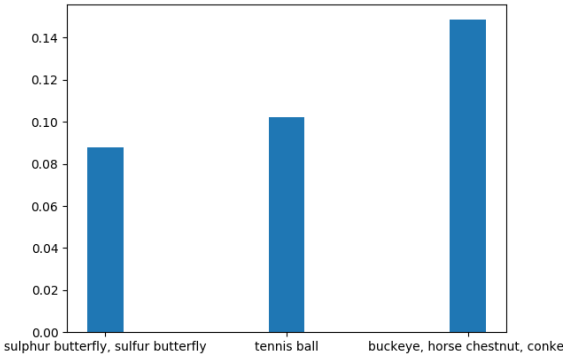
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

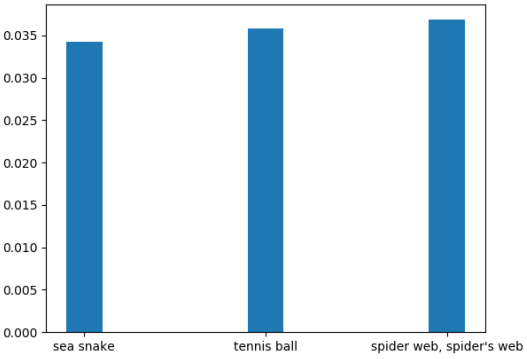





0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
hummingbird
pot, flowerpot
vase


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
pot, flowerpot
tray
handkerchief, hankie, hanky, hanl


0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
sulphur butterfly, sulfur butterfly
tennis ball
buckeye, horse chestnut, conke


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
sea snake
tennis ball
spider web, spider's web

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。

Smoothing method : Median Filter

Success rate 變化 : 0.990 → 0.555

Median Filter 是將 filter(3*3)放在(x,y)上，算出 filter 內所有值的中位數，來取代原本(x,y)的值，這個方法可以去除原本圖像中較為極端的雜訊

以 Success Rate 來看，防禦了約 45%的攻擊圖片，但仍舊有超過 50%攻擊成功，可能只能算是中等能力的防禦方式