

## Machine Learning HW5 Report

學號：B05901003 系級：電機二 姓名：徐敏倩

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` 我使用 iterative FGSM，即每一次都只會更動圖片一小部分，而在多次更新之後，已經 attack 成功的圖片就不會再繼續變動，並繼續變更尚未成功 attack 的圖片。實際的參數設 Proxy Model 為 ResNet-50，epsilon 為 0.003，epoch 為 20。

不同於 FGSM 只更新一次，若想要提高 attack 的 success rate 就會需要較大的 epsilon，會導致較高的 L-infinity；而使用 iterative FGSM 時，欲達成相同 attack 的 success rate，epsilon 則不需要那麼高，雖然更新多次，但已經 attack 成功的圖片不再更新，因此，可以有效的降低 L-infinity。

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-infinity
<code>hw5_fgsm.sh</code>	ResNet-50	0.925	5.0000
<code>hw5_best.sh</code>	ResNet-50	1.000	4.0000

(`hw5_fgsm.sh` 使用 FGSM，設 Proxy Model 為 ResNet-50、epsilon 為 0.08；而 `hw5_best.sh` 使用第一題所述之攻擊方法)


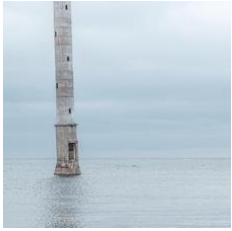

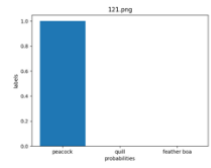
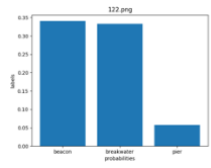
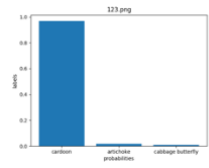
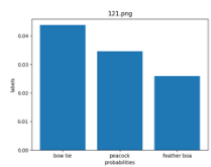
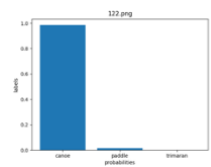
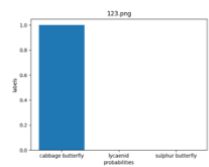
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

Proxy Model	Success Rate	L-infinity
VGG-16	0.185	4.0000
VGG-19	0.185	4.0000
ResNet-50	0.340	4.0000
ResNet-101	1.000	4.0000
DenseNet-121	0.240	4.0000
DenseNet-169	0.235	4.0000

(以上以第一題所述 `hw5_best.sh` 之攻擊方法，僅改變其 Proxy Model)

觀察不同 Proxy Model 之 Success Rate，可以發現只有在 Proxy Model 為 ResNet-50 的時候，有 1.000 的 Success Rate，而使用其他 Proxy Model 時，雖然直接以該 Proxy Model 預測時，Success Rate 都還有 0.9 左右，但上傳後的 Success Rate 都很低。因此，猜測背後的 black box 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)。

原圖			
攻擊前 機率圖			
攻擊後 機率圖			

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。

	攻擊前	攻擊前 + smoothing	攻擊後	攻擊後 + smoothing
Success Rate	0.000	0.160	1.000	0.570

(使用 `scipy.ndimage.gaussian_filter()` 對圖片以 Gaussian Filter 的方式做 smoothing)

從上表中可以看到不論是攻擊前或攻擊後的圖片，經過 `smoothing` 以後，其 `success rate` 都會受到影響。對於攻擊前的圖片，`smoothing` 的影響並不大，只有 16.0% 的圖片會被因此分到其他類別；但對於攻擊後的圖片而言，有將近一半 (57.0%) 的圖片，會因此被分回原本的類別，使攻擊失效。