Machine Learning HW5 Report

學號：B04901069 系級：電機四　姓名：林志皓

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)

我使用的 proxy model 為 ResNet50。我的方法為對於任一張輸入圖片，找到模型輸出中「第二可能的類別」，也就是 1000 個輸出分數中第二高的，作為我的 attack target，以此為目標的 Cross Entropy loss 進行 backpropagation 之後，圖片每個 pixel 獲得各自的 gradient，我取 sign 值（類似 FGSM）進行 gradient decent，使的該圖片會更像第二可能的類別，誤導模型誤判。而對於每一張圖片我都做足夠的 iteration 直到模型誤判，使成功攻擊的機率為 100%，即使看起來有點報暴力，但位於大多數的圖片只須一次即成功，最多的也只要 5 次，最後平均的 L-infinity 為 1.12 。

FGSM 的方法，是使圖片變「不像」原本的類別，如果對於每一圖片也做足夠的 iteration 直到成功，那平均 L-infinity 為 2.6，有些圖片最多須做 60iteration。而我使用的方法，是使圖片除了不像原本類別外，也去像另一個可能的目標，都包含在 Cross Entropy loss 中，使圖片能更有目標，更有效率的找到 classifier 的邊界，提高成功的機率。

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

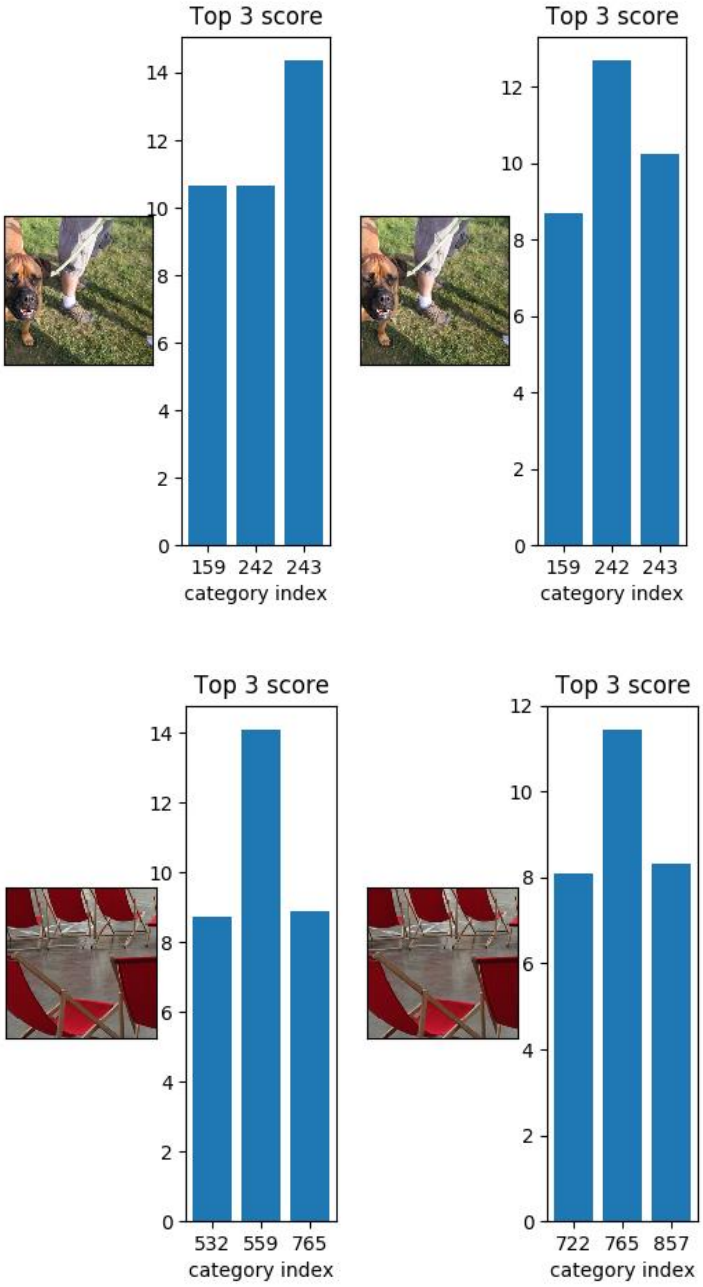
Method	**proxy model**	**sucess rate**	**L-inf norm**
hw5_fgsm.sh	ResNet50	0.73	1.0
hw5_best.sh	ResNet50	1.0	1.12

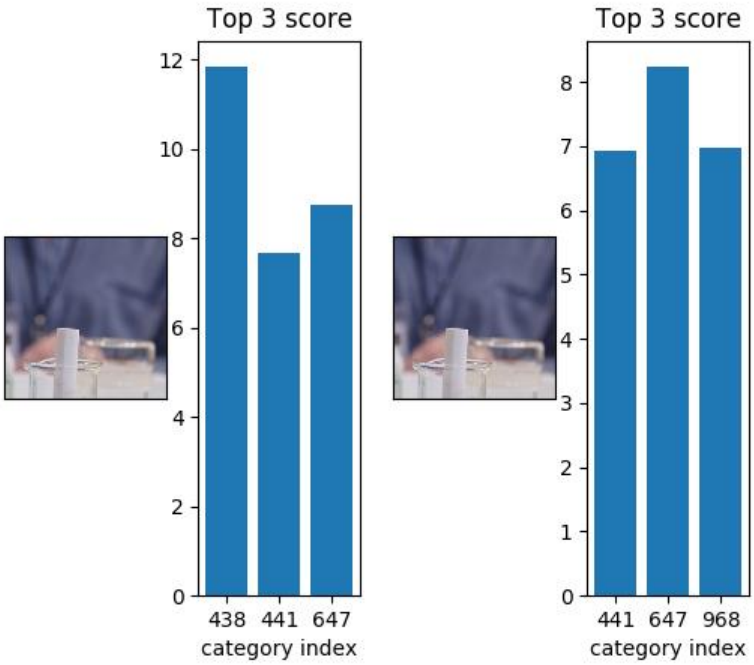
2. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

對於 6 種可能的 model，我分別進行 FGSM 攻擊，最後 success rate 與 L-inf norm 如下表所示，可以很明顯的看出以 ResNet50 作為攻擊對象，遠比其他模型還要容易成功，因此我推論 proxy model 即為 ResNet50

Model	Vgg16	Vgg19	ResNet50	ResNet101	DenseNet121	DenseNet169
success rate	0.04	0.04	0.73	0.1	0.075	0.055
L-inf norm	1.0	1.0	1.0	1.0	1.0	1.0

3. (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)。





4. (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。

我實做的是以 gaussian filter 進行 smoothing (sigma = 1)，攻擊產生的影像，success rate = 1.0, L-inf norm = 1.12，而經過 smoothing，success rate = 0.43, L-inf norm = 113.25。可以明顯看出 smoothing 確實可以做到 defense 的效果，因為 gaussian 可以將攻擊的訊號藉由平均而弱化，同時 L-inf norm 也因為 filter 的效果使每個 pixel 與原來的值不同而變大不少。