Machine Learning HW5 Report

學號：R07942091　系級：電信碩一　　姓名：許博閔

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)

使用的 proxy model : ResNet50
方法 : 基本上和 FGSM 相同，每次對圖片加上 noise 後，會測試是否攻擊成功，若不成功則會再加 noise，直到成功為止，以達到 100%success rate
參數 : 第一次加 noise 的時候，ε = 0.005，往後每次加 noise 時的 ε = 0.001
和 FGSM 的差異 : 會加 noise 不只一次，ε 的值會改變
如何影響結果 : 因為 ε 的值比較小，因此需要花的時間會比較久，但在同樣達到 100%success rate 的狀況下，因為每次變動的 pixel 值較小，最終可以達到較小的 L-inf. Norm 。

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	ResNet101	0.615	11.9800
hw5_best.sh	ResNet50	1.000	1.0100

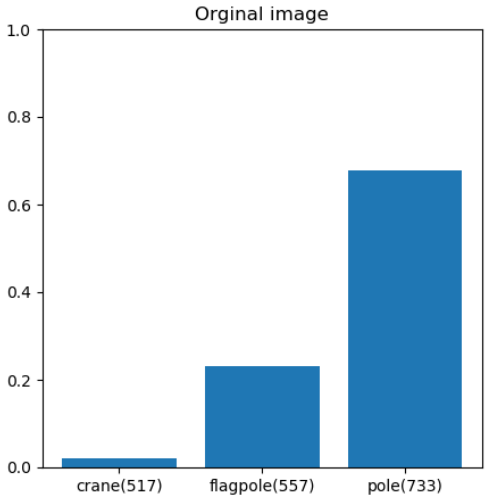
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

我用相同的 code，只變動 proxy model 的方法，看哪個 model 可以得到最高的 success rate 來判斷，以下是 ε = 0.01，用 FGSM 加 noise 20 次後的結果，因為 ResNet-50 的結果明顯高於其他 model，因此 black box 最有可能是 ResNet-50。

	Success rate	L-inf. norm
Vgg-16	0.465	12.0000
Vgg-19	0.440	11.9700
Resnet-50	1.000	11.9600
Resnet-101	0.620	11.9850
DenseNet-121	0.550	11.9650
DenseNet-169	0.525	11.9450

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)。

image 056



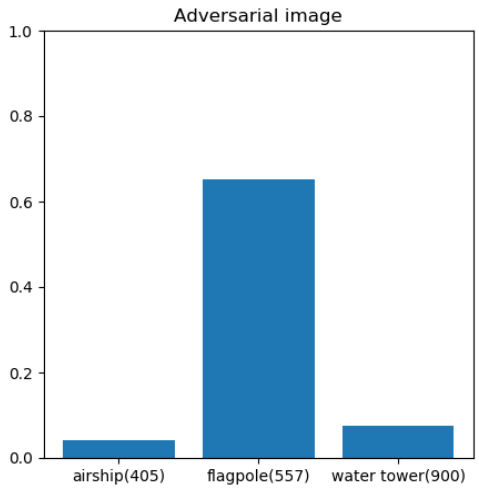
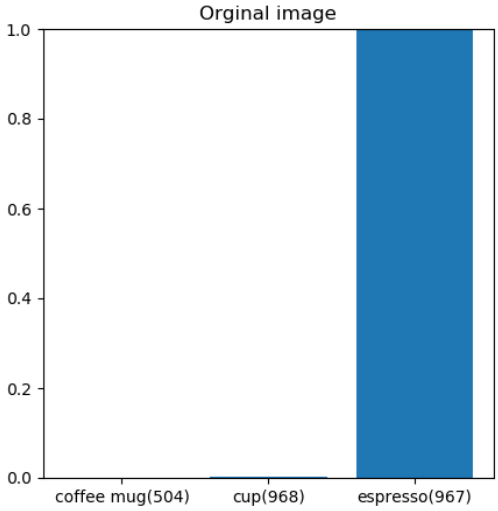
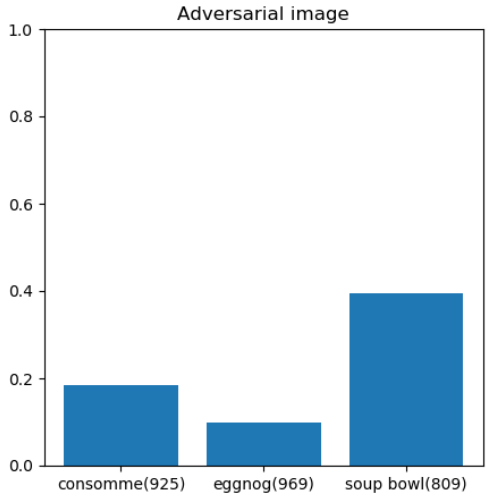


image 103





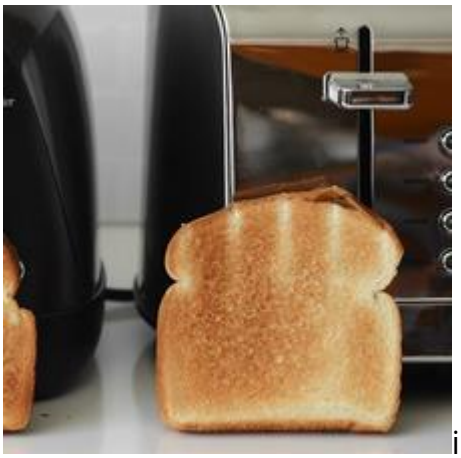
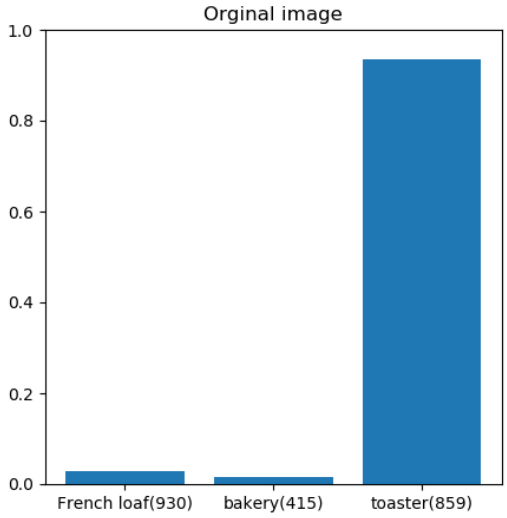
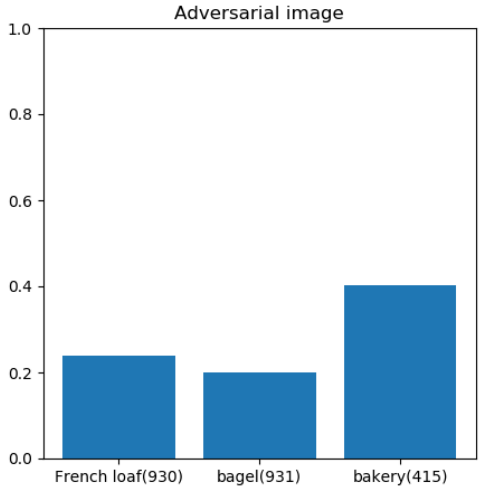


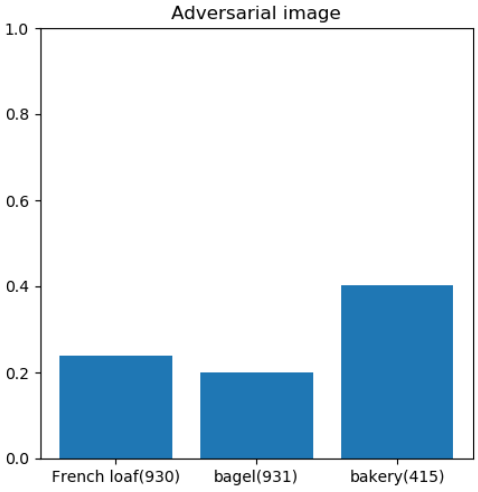
image 183

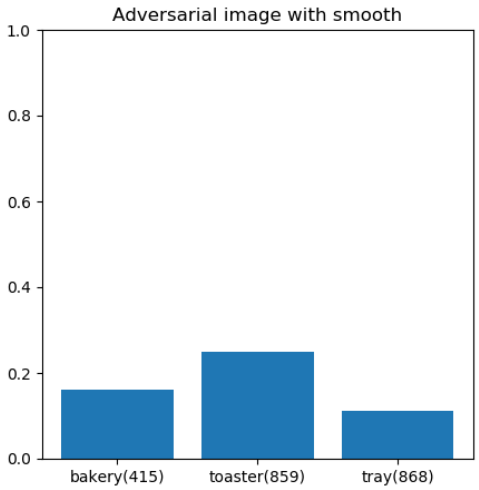




5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

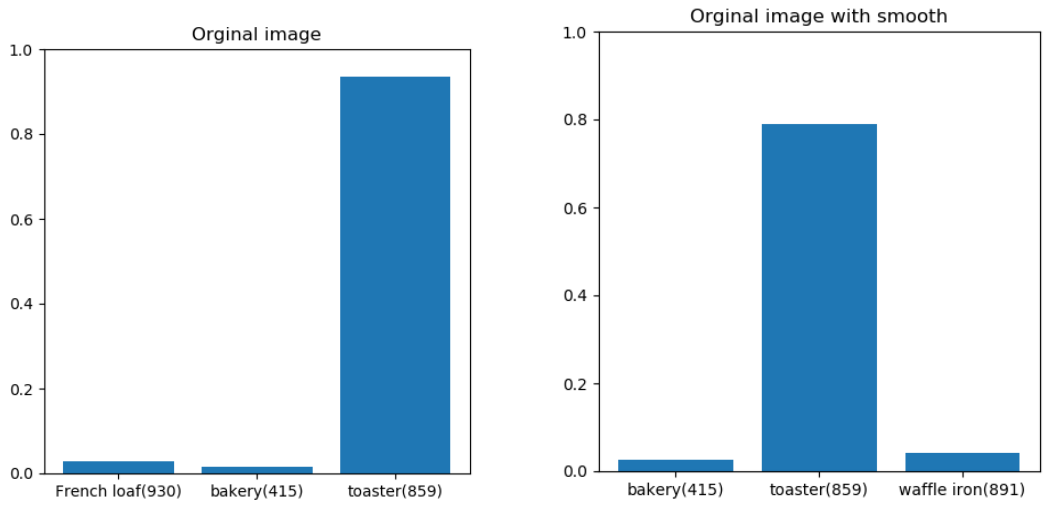
我用 Median filter 來被動防禦，使用 Pillow 中的 ImageFilter.MedianFilter 這個 function 來達到 Median filter 的效果。
防禦前 success rate : 1.00 (200/200)
防禦後 success rate : 0.45 (90/200)
觀察 : 以 image 183 為例





上面兩張圖分別是 image183(true label toaster) 有無防禦的機率圖，加了防禦後確實讓攻擊失敗了，使 model 正確判斷，但若是和沒有被攻擊的圖相比，true label 的機率仍是降低了不少。

對原始圖片的影響
防禦前 model 正確率 : 1.00
防禦後 model 正確率 : 0.90
觀察 : 以 image 183 為例



防禦會使原始圖片在 true label 的機率下降，甚至使 model 做出錯誤的判斷，因此我認為要謹慎選擇防禦的方法、參數，或是在訓練 model 的時候就先讓圖片經過 smoothing，或許可以避免 model 判斷錯誤的機率下降。