

Machine Learning HW5 Report

學號：r07942092 系級：電信碩一 姓名：白佳灝

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 為 resNet50，best 使用的方法為做多次的 FGSM，與原本 FGSM 的差異為，原本只對一張圖做一次 FGSM，best 做了三次。兩個攻擊方式同樣選用 epsilon 為 0.01，，雖然 best 的 L-inf(原本為 1.0，現在 2.0)比原本大了些，但在攻擊程度上明顯提升至 0.995(只用一次 FGSM 只有 0.865)。

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)。


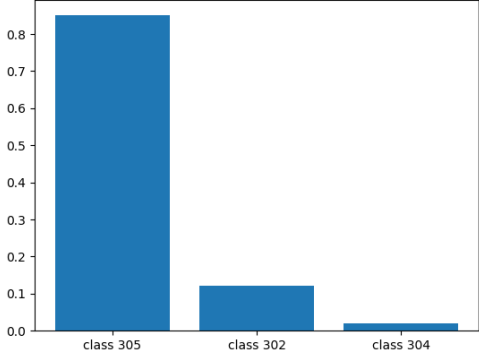
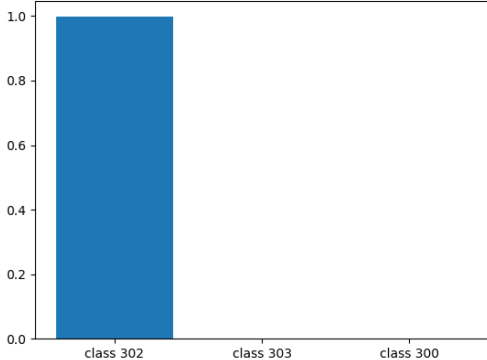

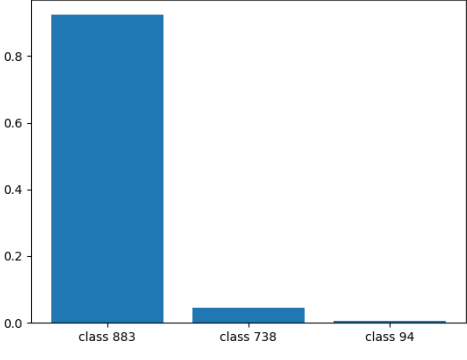
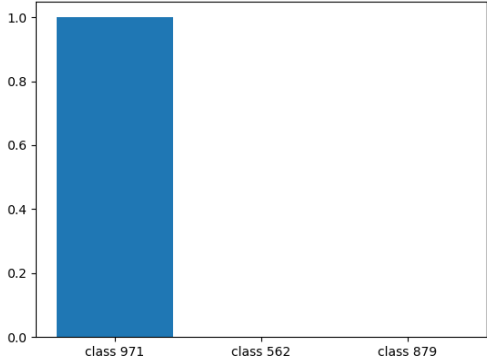

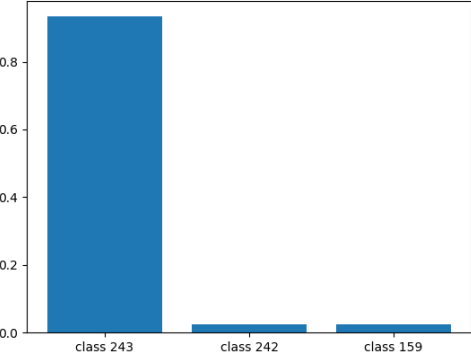
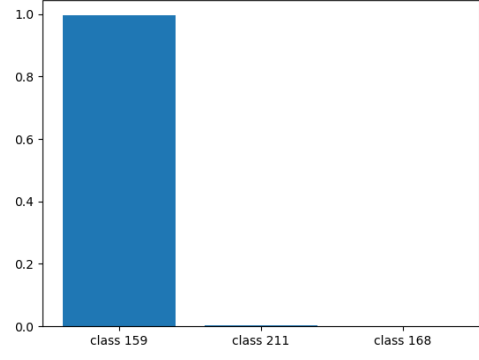
	hw5_fgsm.sh	hw5_best.sh
proxy model	resNet50	resNet50
success rate	0.865	0.995
L-inf. norm	1.0	2.0

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

使用 FGSM，將 epsilon 皆設為 0.01，以下使用不同的 proxy model，觀察它們的 success rate，發現同樣參數下，攻擊成功最高的那個 model 為 resnet50，故猜測 black box 最有可能是 resnet50。(以下為各 model 的 success rate)

Vgg16: 0.1	Vgg19: 0.110	ResNet-50: 0.865
ResNet-101: 0.175	DenseNet-121: 0.135	DenseNet-169: 0.130

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	Original image	Adversarial image
 class 305	 Class305 85.03%	 Class302 99.99%
 class 883	 Class883 92.35%	 Class971 99.93%
 class 234	 Class234 93.35%	 Class159 99.57%

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。

使用 **gaussian filter** 當作被動防禦(採用 **kernel size = 5, sigma = 1** 的 **filter**)，若將攻擊的圖片(圖 a)加上 **gaussian filter** 得到圖 b，發現圖片會相較原圖模糊些，因為 **gaussian filter** 的功用即是將圖片模糊化並且濾掉一些雜訊。



因此我們原本對圖片攻擊的雜訊可能會被 **gaussian filter** 濾掉一些，所以攻擊成功率也明顯下降許多，從原本 **99.5%**掉到只剩下 **55%**。

	success rate
adversarial img	0.995
adversarial img + gaussian filter	0.55