Machine Learning HW5 Report

學號：B05203050　系級：化學三　　姓名：陳品翰

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)

我以 resnet50 作為 proxy model，以 MI-FGSM (momentum iterative FGSM) 為方法攻擊。參數：ε = 5, μ = 1, iteration = 20，並先將所有圖片 normalize 至 -1 ~ 1 的區間 (ε 也隨之調整)，攻擊完後再 deprocess。

本方法與 FGSM 的差異為其使用了 iterative method，讓攻擊前後的圖片差值未必等於 ε，而是每個 iteration 都會根據 gradient 在該點為正或負，加或減 ε/iteration，使 gradient ascent 更容易到達高點。另外採用 momentum 的方法，概念與 adam 類似，可記憶前一步的 gradient，較不容易卡在平坦的 gradient 。此方法使 success rate 從 0.925 上升至 0.995 (L-inf. Norm = 5.0)。

實際攻擊的公式：

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^*, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^*, y)\|_1};$$

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* + \alpha \cdot \mathrm{sign}(\boldsymbol{g}_{t+1});$$

ref: Y. Dong et al., Boosting adversarial attacks with momentum, 2017

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: model = resnet50, success rate = 0.925, L-inf. Norm = 5.0

hw5_best.sh: model = resnet50, success rate = 0.995, L-inf. Norm = 5.0

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。

以下為不同 model 用 FGSM attack 之 success rate：

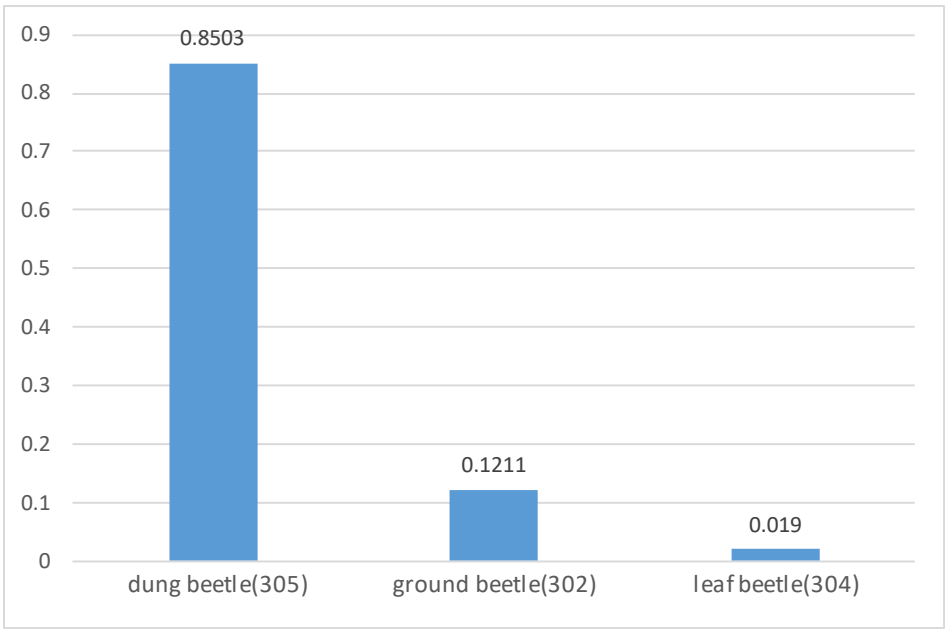
model	Success rate
VGG16	0.245
VGG19	0.245
Resnet50	0.925
Resnet101	0.415
Densenet121	0.340
Densenet169	0.355

Resnet50 之 success rate 顯然遠高於其他，故背後之模型應為 Resnet50。

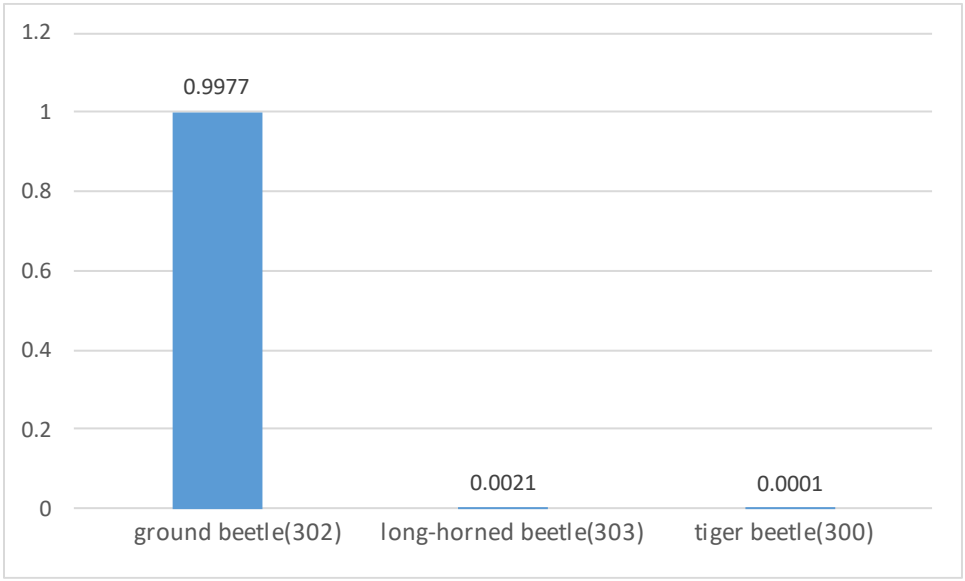
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

1.原圖：

攻擊前：

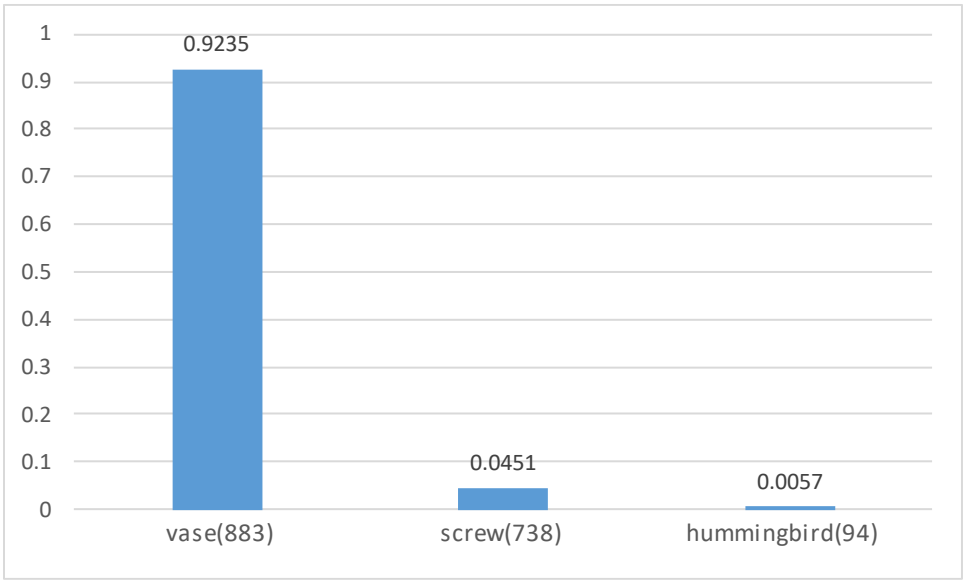

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.8503
0.1211
0.019
dung beetle(305)
ground beetle(302)
leaf beetle(304)

攻擊後：

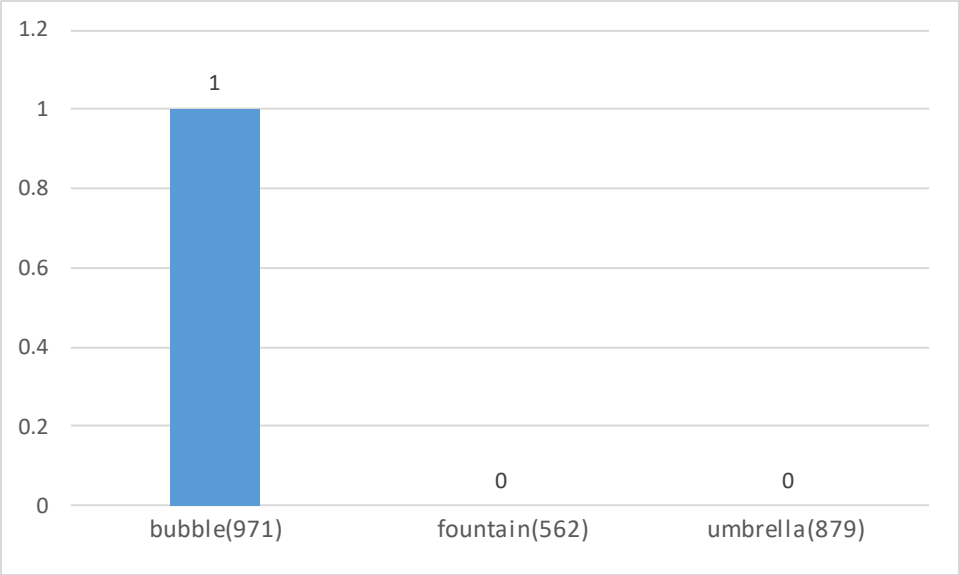

1.2
1
0.8
0.6
0.4
0.2
0
0.9977
0.0021
0.0001
ground beetle(302)
long-horned beetle(303)
tiger beetle(300)

2.原圖：

攻擊前：

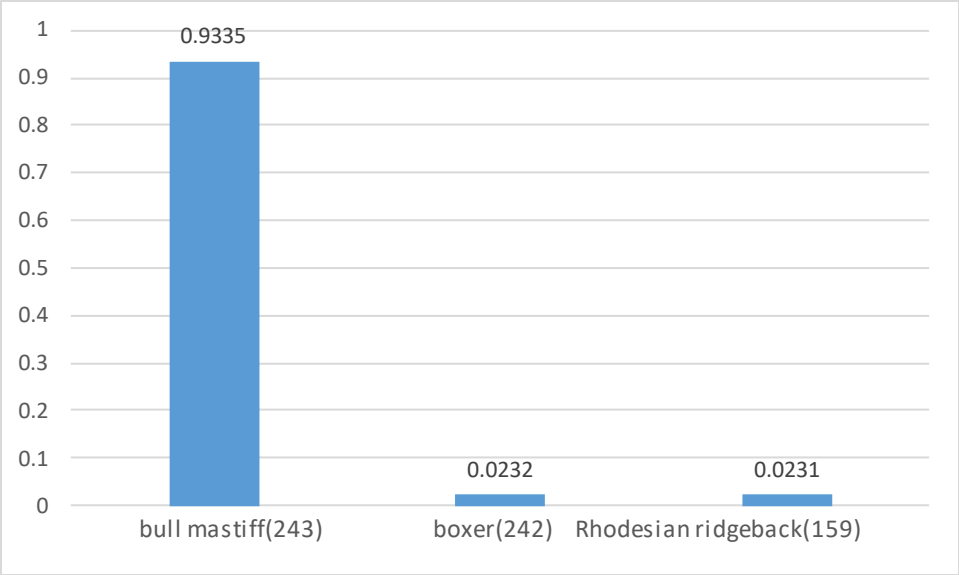

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.9235
0.0451
0.0057
vase(883)
screw(738)
hummingbird(94)

攻擊後：

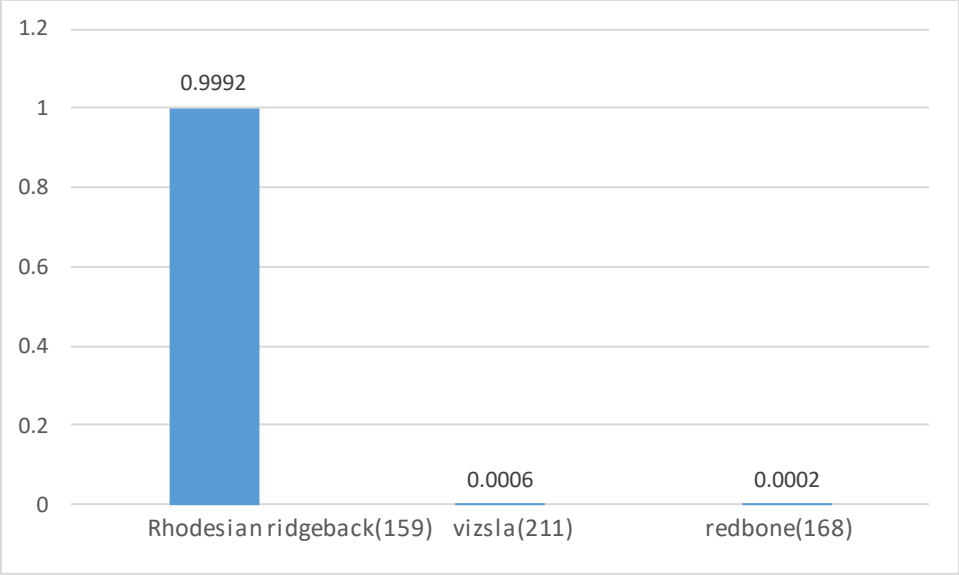


3.原圖：

攻擊前：



攻擊後：



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。

以 5*5 的 average filter 處理 hw5_best 攻擊後的圖片，success rate 由 0.995 降至 0.610，可見 average filter 有效降低模型誤判的比例。但 L-inf. Norm 也從 5.0 升至 147.405，這是因為以 average filter 處理，5*5 的平均有可能跟原本有很大的差距，

例如以下狀況，會得到 245 - 0 = 245 的 L-inf. Norm：

255	255	255	255	255
255	255	255	255	255
255	255	255	255	255
255	255	255	255	255
255	255	255	255	0

↓

245	245	245	245	245
245	245	245	245	245
245	245	245	245	245
245	245	245	245	245
245	245	245	245	245

但肉眼看起來差異不大，只會讓原始圖片模糊化：

攻擊後

filter 後