Machine Learning HW5 Report

學號：R07922108　系級：資工碩一　姓名：陳鎰龍

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 MI-FGSM 在 VGG16 上，他跟 FGSM 不同的地方在於，先另當前速度為 0，每次算出 Gradient 後，將 Gradient Normalized，每過一個 Epoch 將速度更新為之前速度 * decay + Normalized Gradient，接著對速度取 Sign 修改原圖，我使用的 decay 為 1。加上 momentum 後就比較不會卡在 local minimum，而結果也可以看出來，在 Epoch = 5 的情況下，利用 MI-FGSM 後從 0.485 上升到了 0.59。

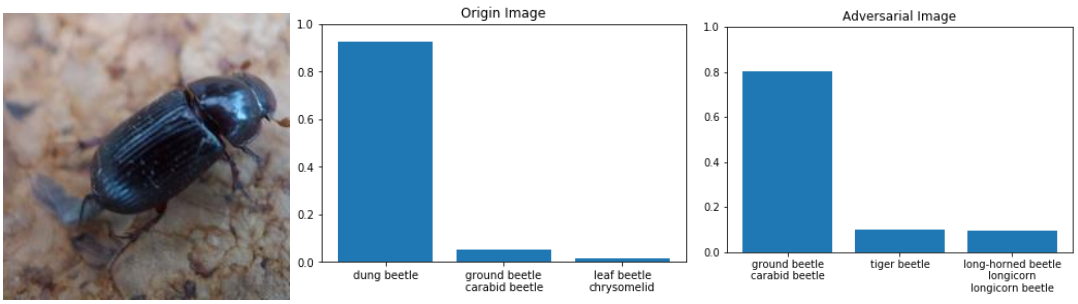
2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。

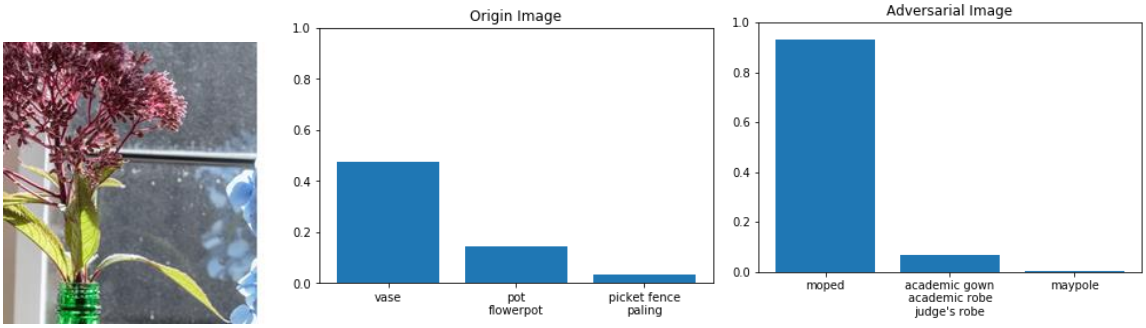
	Model	Success Rate	L-inf. norm
FGSM	VGG16	0.485	5
BEST	VGG16	0.59	5

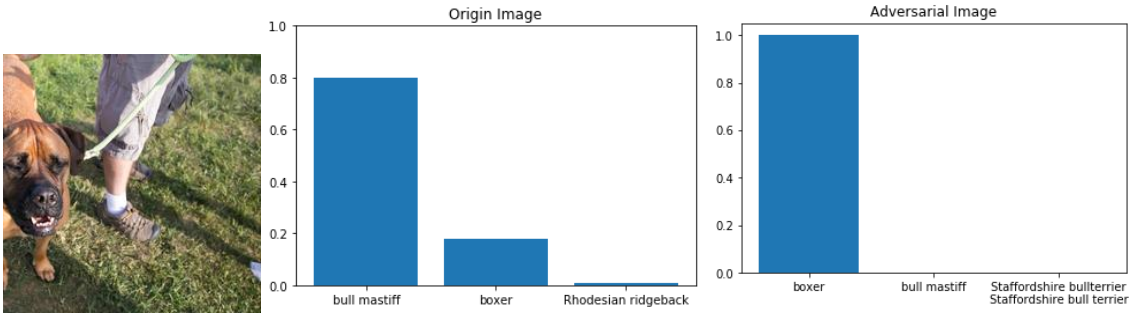
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

我用 L-inf Norm 最大為 10 的前提下，對每一個 model 用 MI-FGSM，發現 success rate 分別為 0.79(VGG16), 0.79(VGG19), 0.765(RESNET50), 0.69(RESNET101), 0.74(DENSENET121), 0.445(DENSENET169)，所以我猜測應該是 VGG16 或 VGG19，可是我 submit 後效果都不是很理想。

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)。







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，附上你攻擊有無的 success rate，並簡要說明你的觀察。

我使用 Median Filter，將圖片每一個 Pixel 與 Neighbor 8 個取 Median，原來用 MI-FGSM 的 success rate 為 0.59，而 smoothing 後 success rate 降為 0.545，有稍稍降低，而用在原圖上最後 predict accuracy 為 0.835(越高越好)。對於每一個 Pixel 最大差距 5，若假設因為 Spatial Locality，原圖 Pixel 與 Neighbor 很相近，而大部分 Pixel 經過 MI-FGSM 後，Pixel 差距會容易上升到 5，因為 Global Maximum 在範圍外(Pixel+-5)機率可能很高，則做 Median 會將 Pixel 差距為 5 的 Difference 縮小，進而提升 Softmax 值。