Machine Learning HW5 Report

學號：b05502145 系級：電機三 姓名：林禹丞

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)

(proxy model)我的 proxy network 採用 resnet50。

(方法、參數)我分成 4 組參數來實行 iterate fgsm (ifgsm)：

5 個 epoch & learning rate 0.0025

25 個 epoch & learning rate 0.0005

20 個 epoch & learning rate 0.001

10 個 epoch & learning rate 0.007

如果第一組參數 train 出來的圖片對 proxy network 攻擊無效，就採用第二組參數，如果第二組也不行就採用第三組，依此類推。

ifgsm 和 fgsm 的差異只在於，fgsm 只 train 一個 epoch。所以 ifgsm 可以經過多次 iterate 達到更好的攻擊效果。我把參數分成 4 層是為了讓所有的圖片對 proxy network 都攻擊成功。

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)。

兩個都用 resnet50 作為 proxy network。

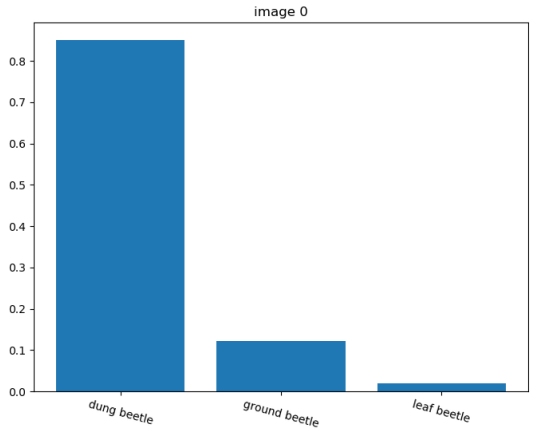
hw5_fgsm 的攻擊正確率為 0.895，infinity norm 為 18.0。

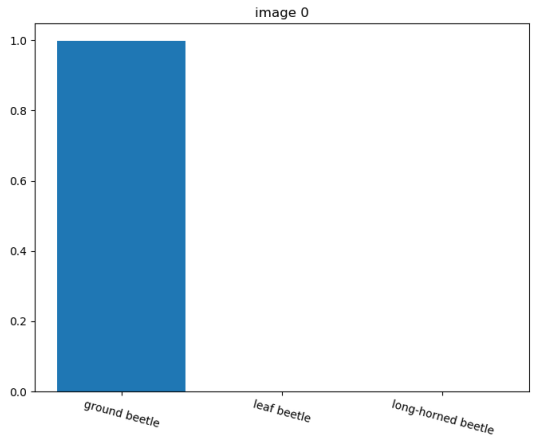
hw5_best 的攻擊正確率為 1.0，infinity norm 為 1.015。

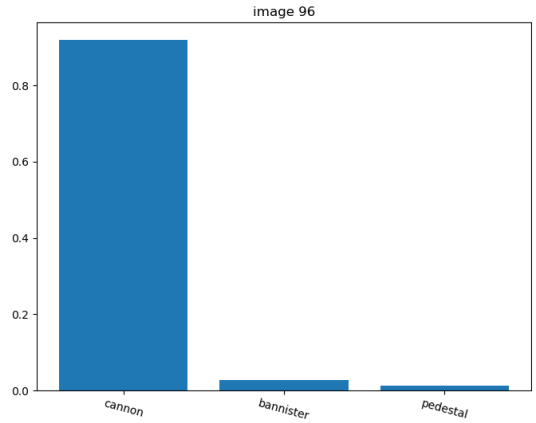
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

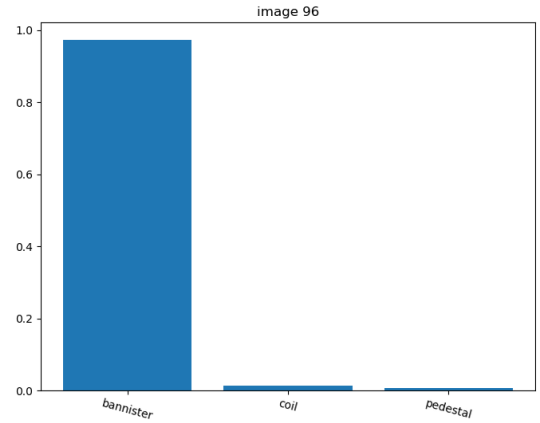
在相同參數下去 train 6 種 model，發現 resnet50 的攻擊正確率最高，且比其他 5 個都高的許多。所以我認為背後的 black box 就是 resnet50。

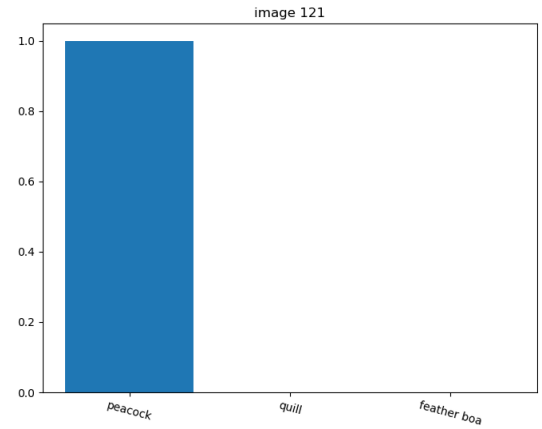
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

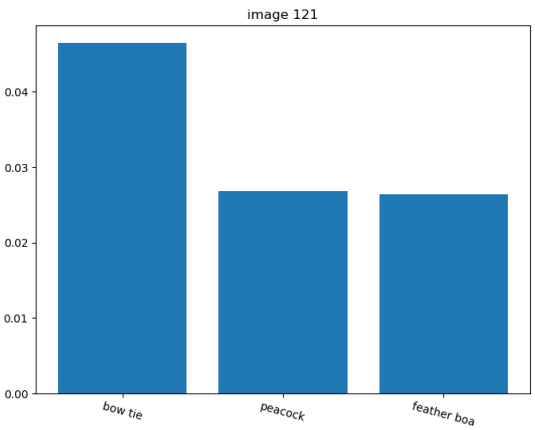












由上到下分別是圖片 0、95、121。而左側的三張為原本的圖片前三高的機率，右邊三張為 best 的圖片前三高的機率。

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。

   Best 的準確率：100%

   Filter 後的準確率：53.5%

   Fgsm 的準確率：89.5%

   Fgsm 做 filter 後的準確率：88.5%

   我對 best 及 fgsm 做出來的圖片都進行 gaussian filtering，可以從結果發現：對 best 可以很有效的降低準確率，而 fgsm 則沒有很明顯的下降。我認為原因在於 best 的 infinity norm 太小，以至於在做 gaussian filter 後，對整個圖片 train 出來的 gradient 有很大的改變，而 fgsm 在做 gaussian filtering 後還是依然保有原本 train 出來的很大的 gradient。