Machine Learning HW5 Report

學號：B06507007　系級：材料二　姓名:王致雄

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)

   我使用的 proxy model 為 pytorch 上 pretrain 的 RES50。我的 hw5_best 單純將一次 FGSM 攻擊後 proxy model 仍能辨識成功的圖片再用同樣的 FGSM 演算法攻擊一次。結果是能在幾乎不使 Linfinity 變大的情況下讓辨識成功率大幅下降，在同樣的限制下，一次 FGSM 可使辨識率下降至 23.5%；而第二次即可讓辨識成功率下降至 2.5%，同時 Linfinity 均為 2。若要求僅使用一次 FGSM 達成 2.5%辨識成功率，結果 Linfinity 約為 30。

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)。

   均使用 Res50 作為 proxy model。

   Hw5_FGSM 的結果為 success rate:0.865 Linfinity:9.00

   Hw5_BEST 的結果為 success rate:0.975 Linfinity:2.00

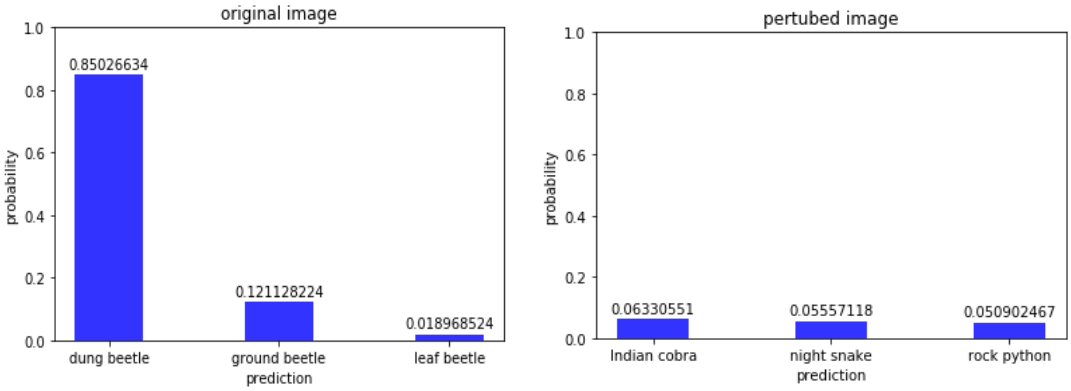
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

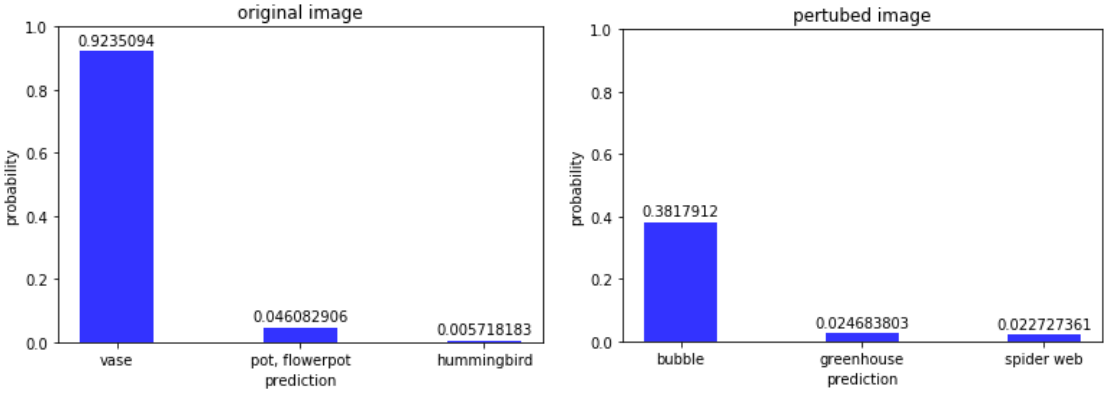
   將 6 種模型都嘗試了以後，我認為 proxy model 即為 pretrain 過後的 res50。其他五種 model 經過同樣的演算法攻擊過後對於助教的黑盒子 Linfinity 約為 20 的狀況下 success rate 大約都落在 0.3~0.5 間。

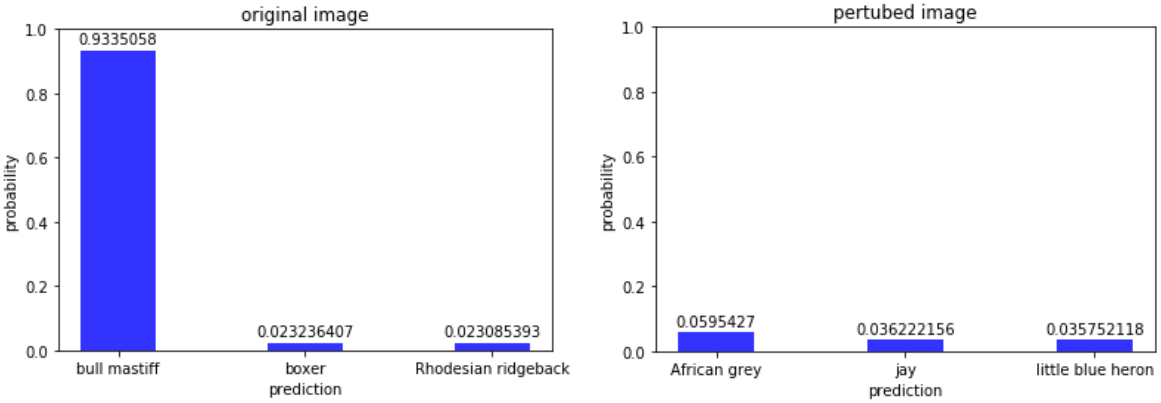
   另外我也有試著自己用那 200 張圖片自己 train res50 network，確保 training set accuracy 達 100%後作為 proxy model 攻擊。對此 proxy model 攻擊成功率達 96%，Linfinity 為 38。然而上傳後 success rate 僅剩 55%。

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)。


original image
probability
1.0
0.8
0.6
0.4
0.2
0.0
0.85026634
0.121128224
0.018968524
dung beetle
ground beetle
leaf beetle
prediction
pertubed image
probability
1.0
0.8
0.6
0.4
0.2
0.0
0.06330551
0.05557118
0.050902467
Indian cobra
night snake
rock python
prediction





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 處理圖片，可將辨識正確率從 2%提升至 12%。不多，但還是有提升。原理投影片內連結有，外觀其實就是讓他看起來變霧，對人類而言應該還看得出是什麼。