學號：b06901153 系級：電機三 姓名：林瑩昇

1. (2%) 試說明 hw6_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model用的是densenet121，方法則是iterative FGSM，以下是比較：

(1) FGSM：

A. 因為只計算一次gradient，所以eps必須要調得夠大，才能確保一定的攻擊成功率
B. 如果我們把成功attack當作一個training的目標，用一般的FGSM相當於只做一個iteration，因此lr(eps)就要很大，而且可能攻擊的方向還不對
C. 如果要達到100%成功攻擊，eps要到接近1

(2) iterative-FGSM：

A. 我每做一次attack就會做一次inference，如果沒有成功則用心的圖片再計算一次gradient
B. 與(1)B相同，這相當於做很多iteration，lr(eps)也就可以調小
C. 因為每次attack都會計算gradient，所以每次都會調整攻擊的方向

(3) 我最後用的是eps=0.01, iteration=50，便可以達到100%attack

2. (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。

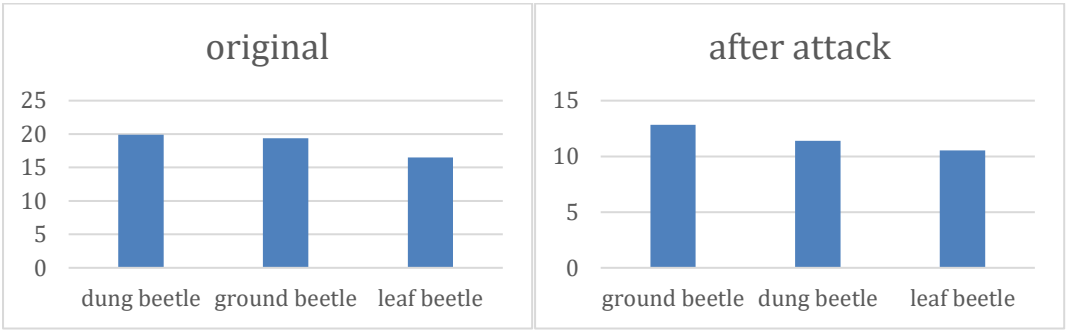
(1) proxy model ：densenet121

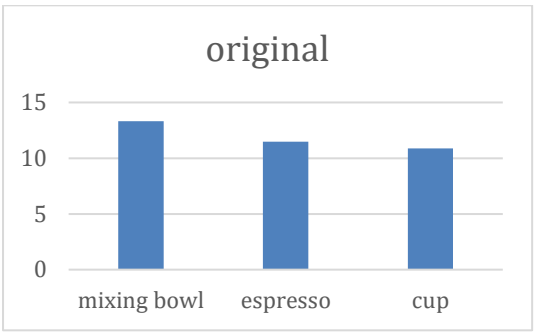
(2) 用助教提供的sample code固定eps後，用所有可能的model下去做攻擊，丟到judgeBoi上實測，可以發現除了densenet121以外其他網絡的攻擊成功率都在0.3左右，有的甚至到0.1。

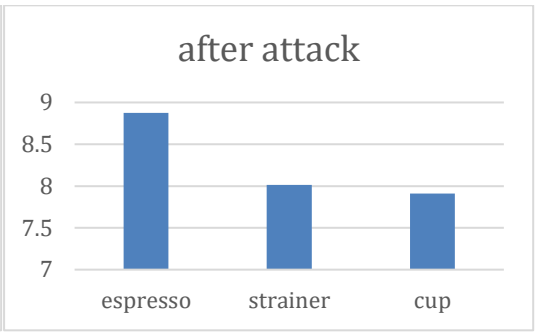
(3) 這裡有一個地方要注意就是eps一開始不能開太大，像我一開始開太大，導致攻擊過後的圖片不管是哪一種model都很難辨識，導致我誤以為black box是其他model。

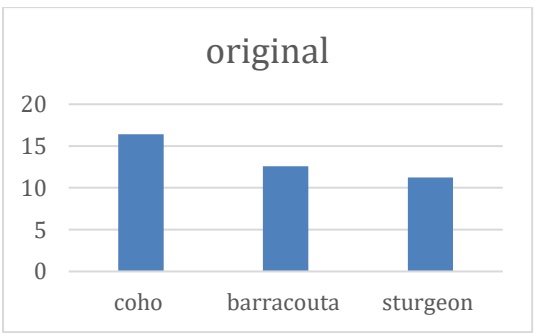
3. (1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

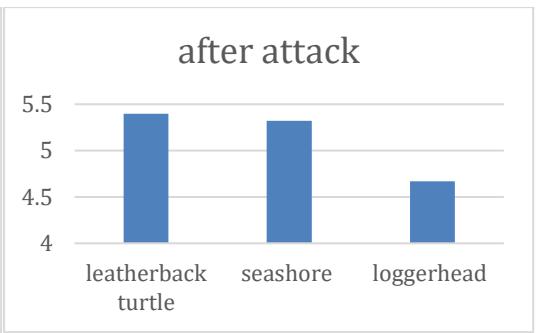
取這三張圖(000.png, 075.png, 150.png)











4. (2%) 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。

(1) 我用的是gaussian filter，實作方法是繼承nn.Module並更改裡頭kernel的weight，forward便是一層的conv2d。其中我試過的kernel size有5*5、7*7、9*9，其中以9*9效果最好

(2) 在未使用filter前，success rate為0.975，而使用了filter後，success rate降到了0.315

(3) 下圖為000.png經過個步驟的圖片，可以看出gaussian filter有模糊邊界的效果，雖然會試圖片稍微不清楚，但卻可以稍微降低fgsm加在圖片上的noise

A. original

B. attacked (hw6_best.sh)

C. filtered