學號：R08946006　系級：資料科學學程碩一　姓名：周逸平

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)**

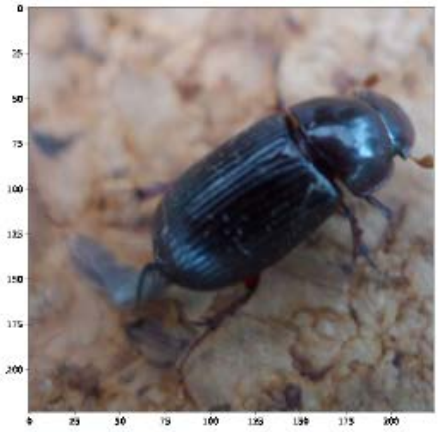
攻擊方法：basic iterative method (iterative – FGSM)
Proxy model：densenet121
Epsilon : 0.03
Alpha : 0.005
Iteration :
Judge boi acc : 1 , L-inf : 8.4

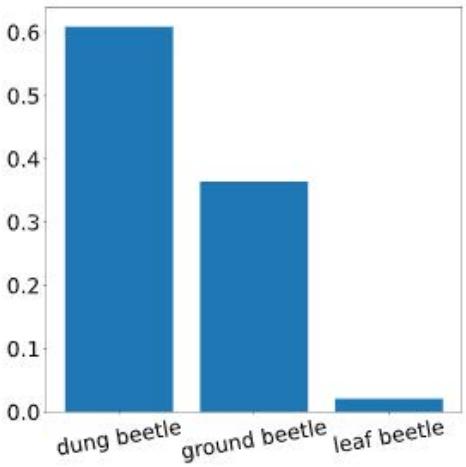
我嘗試了 alpha 自 0.005 至 0.001 以及 iteration: 5,10,15 這幾個參數來做 fine-tuning，最後決定了以上面的參數作為結果。
比起 one-shot 的 FGSM，I-FGSM 設定了一個 learning rate (alpha) 來進行多次的 FGSM 攻擊，讓最後攻擊的方向得以多次校正，以求得到夠高的攻擊成功率，同時較低的 L-inf。

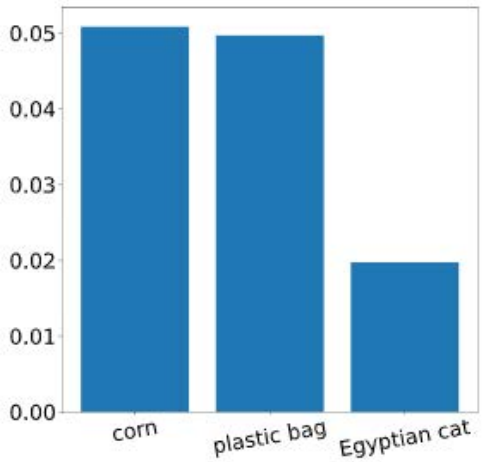
2. **(1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。**

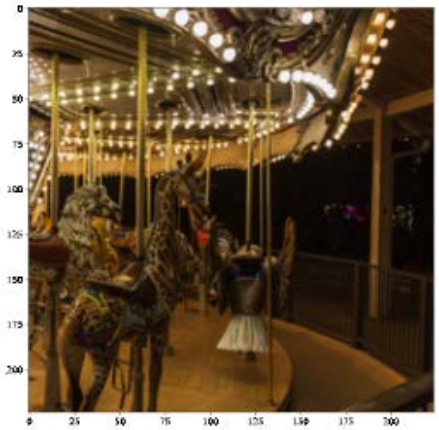
使用了 FGSM，並以 Epsilon = 0.3 針對所有 black box 可能 model 都進行了攻擊後，發現只有 densenet121 的攻擊成功率與本地的成功率相同，故我猜測 black box 是 densenet121。
選擇 0.3 是避免過低的 epsilon 使得不管怎麼扔大家都無法成功攻擊，同時過高的 epsilon 則是避免大家都完全攻擊，導致找不出誰是正確 model。

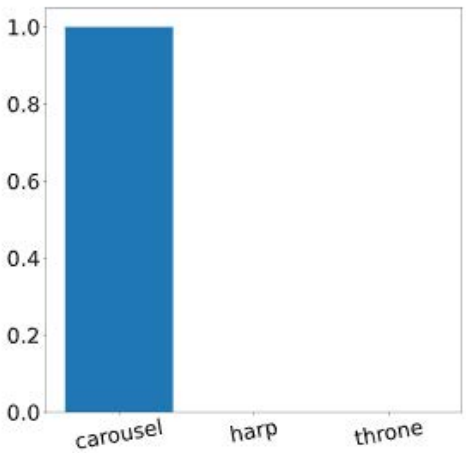
3. **(1%)** 請以 **hw6_best.sh** 的方法，**visualize** 任意三張圖片攻擊前後的機率圖**(**分別取前三高的機率**)**。

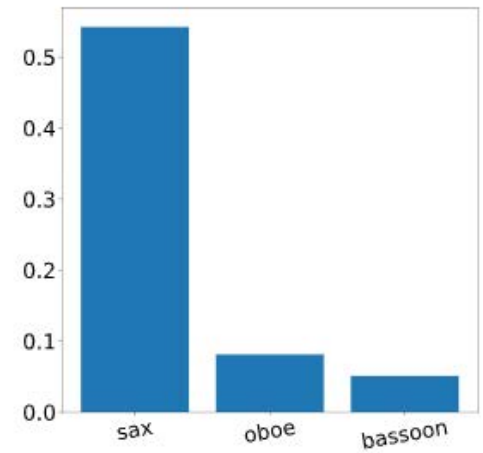


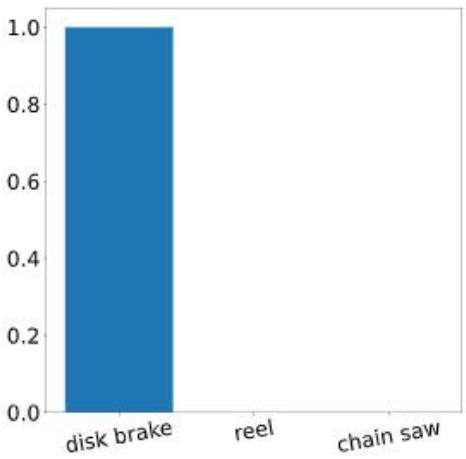


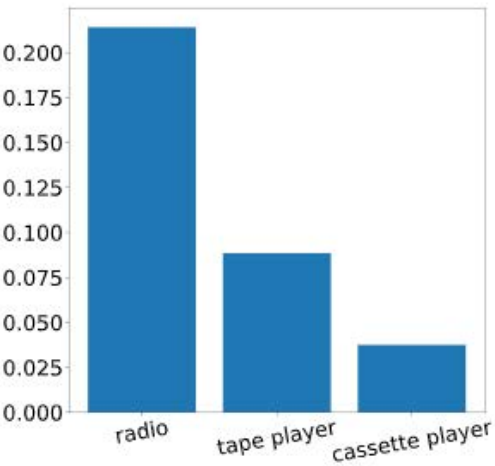






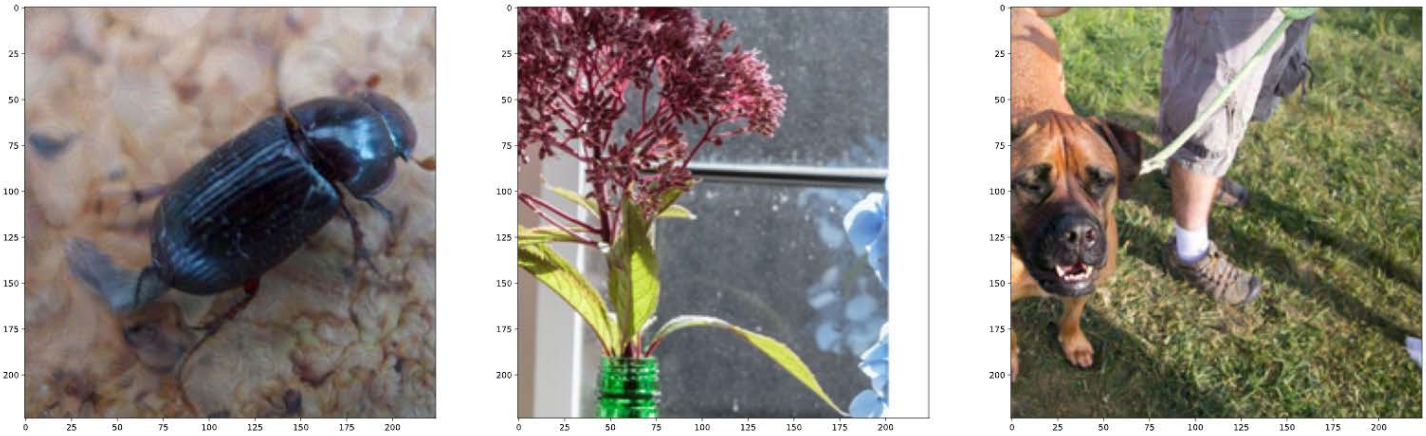




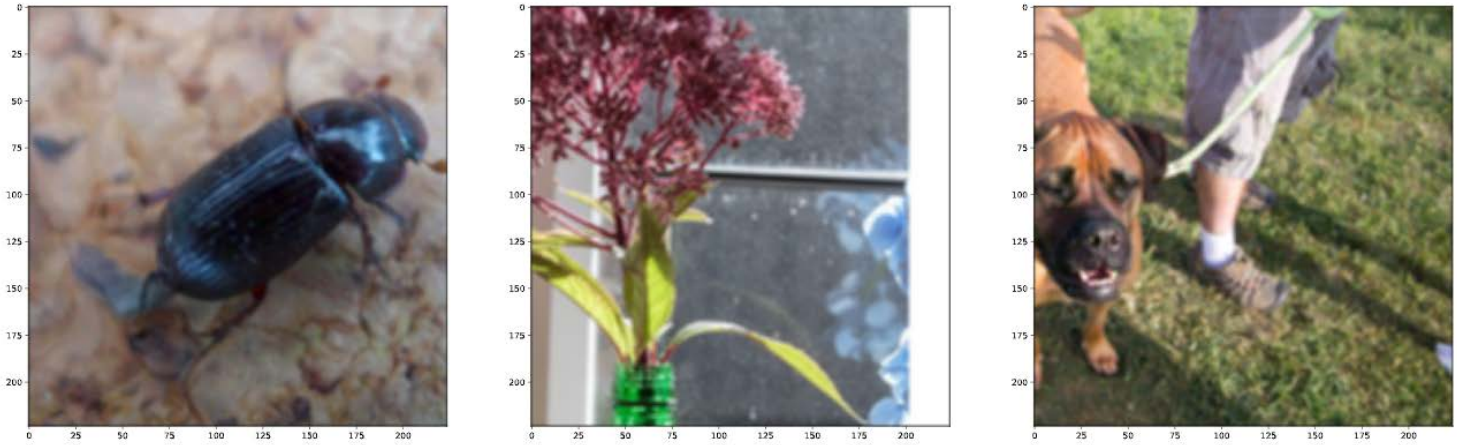


4. **(2%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。**

使用 best model 產生的 adv，其實可以以肉眼看到部分雜訊



使用 gaussian filter, sigma = 5 平滑化之後，可以發現圖片變得稍微糊糊的



而比較前後的攻擊成功率：

平滑前攻擊

```
Epsilon: 0.005 Test Accuracy = 0 / 200 = 0.0
```

平滑後攻擊，sigma = **5**

```
Epsilon: 0.005 Test Accuracy = 12 / 200 = 0.06
```

平滑後攻擊，sigma = **3**

```
Epsilon: 0.005 Test Accuracy = 10 / 200 = 0.05
```

可以發現在 sigma 越高時，由於圖片越糊，所以緩和攻擊的效果越顯著，但也有可能過高的 sigma 導致圖片糊到模型本身辨識就出問題，所以 sigma 也是個可以 tune 的參數之一。