

## 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)

- proxy model : Densenet-121
- 方法 : IFGSM
- 參數 : epsilon = 0.01, T = 20

IFGSM相對於FGSM來說，分了更多步驟去添加噪音，更能擬合模型中參數的模樣，讓新產生的圖片更能逃過模型的預測，尤其是在white box training的時候。雖然說這次的題目是black box，但如果proxy model選中了black box，題目便可能由black box轉為white box，IFGSM的效果也就比FGSM好上更多。

|               | SUCESS RATE | L-INFINITY |
|---------------|-------------|------------|
| FGSM(eps=0.2) | 0.91        | 10.18      |
| IFGSM         | 0.975       | 1.34       |

這樣的結果也驗證了，black box模型應該就是Densenet-121(更多比較可見第二題)。

## 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。

以下的proxy model測試我統一用FGSM比較，並將epsilon固定在0.2：

|              | SUCESS RATE | L-INFINITY |
|--------------|-------------|------------|
| VGG-16       | 0.91        | 10.24      |
| VGG-19       | 0.375       | 10.08      |
| ResNet-50    | 0.5         | 10.41      |
| ResNet-100   | 0.435       | 9.25       |
| Densenet-121 | 0.91        | 10.18      |
| Densenet-169 | 0.53        | 9.58       |

由上面結果可以發現，利用Densenet-121能得到最佳的success rate，而且結果大幅領先其他模型，因此我猜測Densenet-121可能是blackbox。

為了強調第一題的驗證，我將ResNet-50再拉出來進行IFGSM，會發現反而越側月不準，而Densenet-121能夠在IFGSM上越作越準，因此極有可能是black box model。

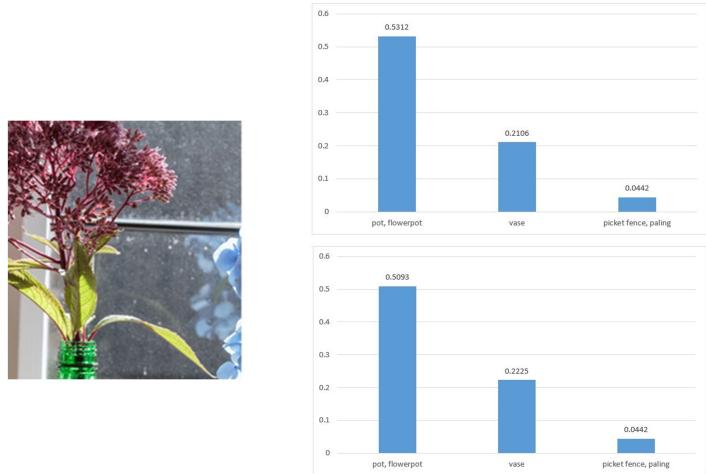
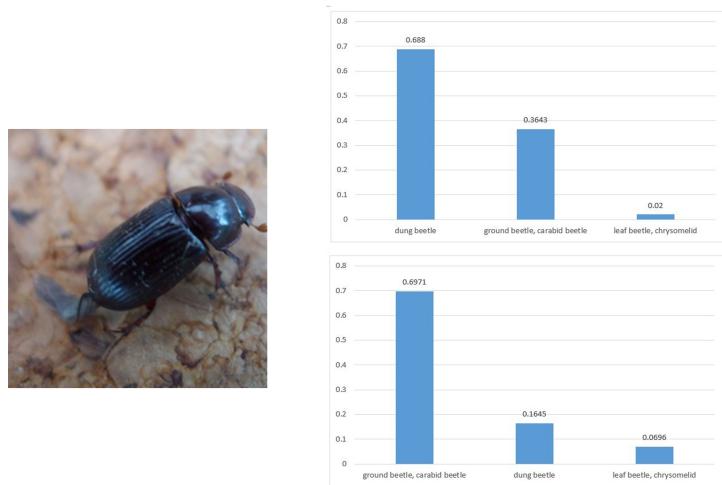
|  | SUCESS RATE | L-INFINITY |
|--|-------------|------------|
|--|-------------|------------|

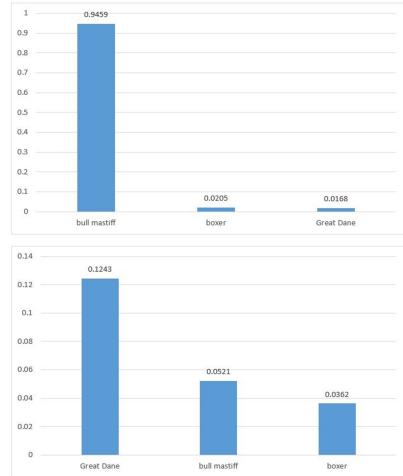
|                       | SUCESS RATE | L-INFINITY |
|-----------------------|-------------|------------|
| FGSM w/ ResNet-50     | 0.5         | 10.41      |
| IFGSM w/ Densenet-121 | 0.095       | 1.41       |

### 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)。

圖中，最左邊為原圖，右上為模型被攻擊前的預測機率，右下則為被攻擊後的機率。





#### 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。

我選擇在圖片輸入之前，套上gaussian blur來完成這項實作，其中的sigma參數調為1。成果如下：

|              | SUCES RATE | L-INFINITY |
|--------------|------------|------------|
| IFGSM        | 0.975      | 1.34       |
| IFGSM with g | 0.95       | 107.94     |

由此可見，有了gaussian filtering，我們能透有效的實行被動防禦，success rate確實降低了。然而，這樣的防禦會對原先的圖片造成變更，最明顯的就是清晰度降低(左圖為原圖、右圖為加上gaussian filtering的圖)：

