學號：b05901168　　系級：電機四　　姓名：陳冠豪

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)

使用的 proxy model : densenet121 with pretrain
攻擊方法: Boosting Adversarial attacks with momentum (MI-FGSM)

此種方法 pseudo code 如下

Input: A classifier f with loss function J; a real example $\boldsymbol{x}$ and ground-truth label y;
Input: The size of perturbation ϵ; iterations T and decay factor μ.
Output: An adversarial example $\boldsymbol{x}^*$ with $\|\boldsymbol{x}^* - \boldsymbol{x}\|_\infty \leq \epsilon$.

$\alpha = \epsilon/T$;
$\boldsymbol{g}_0 = 0$; $\boldsymbol{x}_0^* = \boldsymbol{x}$;
for $t = 0$ to $T - 1$ **do**
    Input $\boldsymbol{x}_t^*$ to f and obtain the gradient $\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^*, y)$;
    Update $\boldsymbol{g}_{t+1}$ by accumulating the velocity vector in the gradient direction as

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^*, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^*, y)\|_1}; \tag{6}$$

    Update $\boldsymbol{x}_{t+1}^*$ by applying the sign gradient as

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* + \alpha \cdot \mathrm{sign}(\boldsymbol{g}_{t+1}); \tag{7}$$

end for
return $\boldsymbol{x}^* = \boldsymbol{x}_T^*$.

此種方法跟 FGSM 想法類似，差別為引入 momentum，每次更新 gradient 皆會參考先前 gradient 的值，相較於 FGSM 只會取當下計算出的 gradient 來計算，而此種方法的缺點是需要重複計算 T 次, FGSM 只需要做一次就好

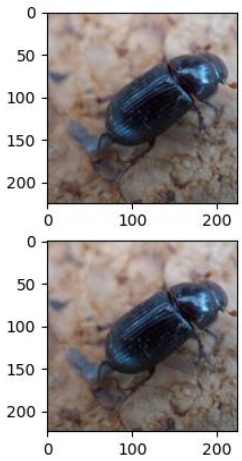
而我過 strong baseline 的參數為 μ = 0.95，另外本次作業由於是 white model，因此我把 loop 結束的條件修改成直到攻擊成功才結束,所以 ε 對每張圖片都不一樣,取決於做了多少次迴圈

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。
   A：在利用同樣的方法(FGSM)的情況下，依每張照片不同的情況，將 ε 調到對於在攻擊成功的前提下所允許的最小值，上傳到 JudgeBoi 的結果如下

model	Success Rate	L-infinity
VGG-16	0.095	1.0000
VGG-19	0.095	1.0000
ResNet-50	0.150	1.0000
ResNet-101	0.155	1.0000
DenseNet-121	0.785	1.0000
DenseNet-121	0.550	1.0000

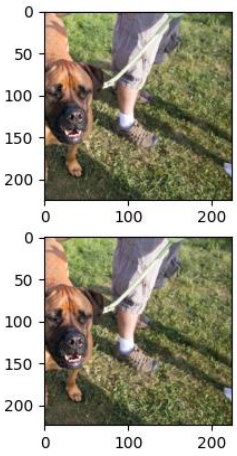
   由實驗結果可以看出 black box 可能為 DenseNet-121
3. (1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

   1)



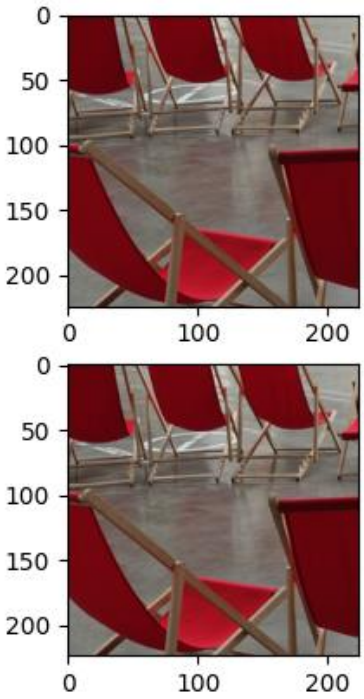
Origin label	Origin confidence
Dung beetle	0.6088
Label after attack	Confidence
Ground beetle	0.4908
Dung beetle	0.4659
Lead beetle	0.0320

2)



Origin label	Origin confidence
Bull mastiff	0.9485
Label after attack	Confidence
Great Dane	0.2681
Bull mastiff	0.2679
Rhodesian ridgeback	0.0756

3)



Origin label	Origin confidence
Folding chair	0.9391
Label after attack	Confidence
Shield	0.1374
Folding chari	0.1318
French horn	0.0867

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。

   實作 Gaussian Filter，並以 hw6_best.sh 的方法產生攻擊的圖片後，將圖片餵進 DenseNet-121 pretrain model 做辨識，下表為有無 Gaussian Filter 的比較

	Without Gaussian Filter	With Gaussian Filter
Success Rate	1.000	0.880

   可以看出經過 Gaussian Filter 能夠降低約 12% 的成功率，效果算是普通，而如果是用 FGSM 並依每張照片不同的情況，將 ε 調到對於在攻擊成功的前提下所允許的最小值，結果如下

	Without Gaussian Filter	With Gaussian Filter
Success Rate	1.000	0.750

   這個結果降低了 25% 的成功率，由這兩個結果可以看出 Gaussian Filter 是能夠抵擋一些改動較少的圖片(ε 較小)，但整體的抵抗效果還是有限.
   而使用 Gaussian Filter 會讓圖片變得較平滑,所以看起來比較糊.