HW5 Report

學號：B06901020 系級：電機二 姓名：張恆瑞

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)

Proxy model: Resnet50 Learning rate = 2.5 Epochs = 10 Epsilon = 3.0

我使用了 *iterative FGSM*，也就是每次只修改圖片一點點(設定小的 learning rate)，然後經過數次微調後可以成功 attack 的圖片就固定不動，並且繼續更新剩下的照片。這個方法和原本的 FGSM 不同之處，在於 FGSM 只更新一次，相當於 iterative FGSM 的特例，也就是將 epoch 設成 1 以及 learning rate 設得很大，FGSM 的缺點就是有些 pixel 其實不需要變動，但卻被修改了，以致於有些圖片會無法被成功 attack。

2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果。

方法	Proxy Model	Success Rate	L-infinity
FGSM	Resnet50	0.920	3.0000
Iterative FGSM	Resnet50	0.980	3.0000

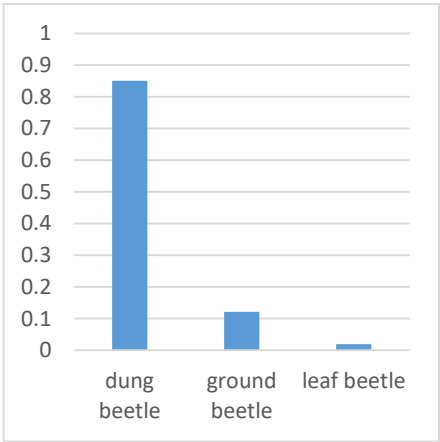
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

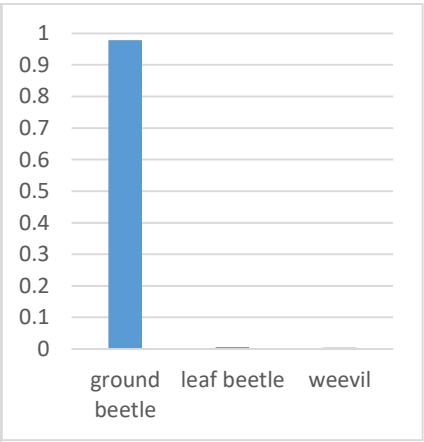
(以下不同情況在訓練的時候準確率都有 0.985 以上，並且 L-infinity 皆為 3.0000)

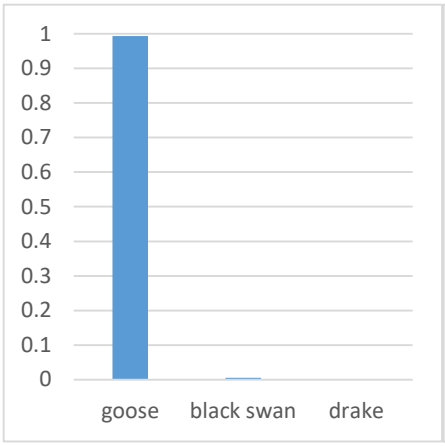
Proxy Model	Success Rate
VGG16	0.165
VGG19	0.185
Resnet50	0.980
Resnet101	0.370
Densenet121	0.270
Densenet169	0.295

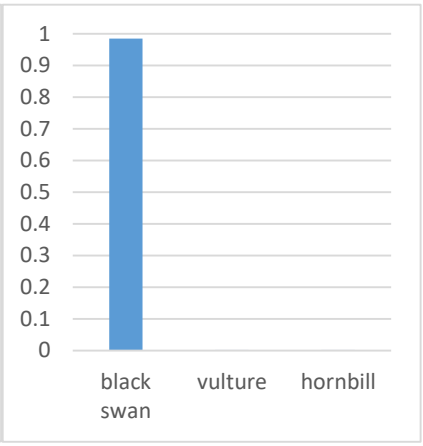
很明顯，只有 Resnet50 的 success rate 特別高，幾乎和原本在 training 時預估的 success rate (0.985)相同，因此可以推論 black box 應為 Resnet50。

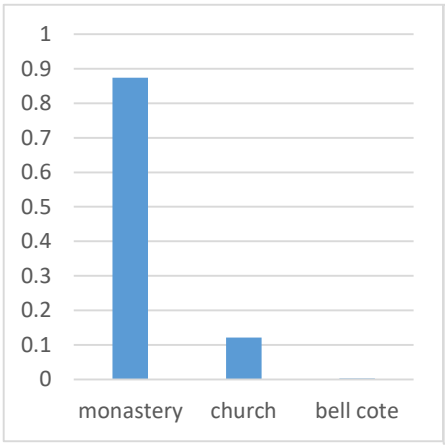
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。(左為 attack 前，右為 attack 後)

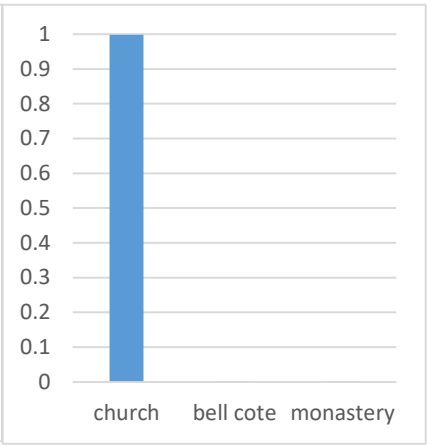












5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

我利用了 median filter 來做防禦，結果如下(success rate)：

	未經 smoothing	經 smoothing 後
原始圖片	0.000	0.335
Iterative FGSM 處理後之圖片	0.980	0.490

經過 smoothing 後，被 attack 的圖片很明顯 success rate 降低了很多，變成原本的一半，效果還挺好的。

然而，原始圖片被做 smoothing 後也降低了預測的準確率(有些被誤判了)，因此 smoothing 也有可能破壞了原本的圖片。