Machine Learning HW5 Report

學號：B05901068 系級：電機三 姓名：蕭如芸

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

   方法：使用 Iterative FGSM，每一步做的就是 FGSM，但重複多個 epoch。

   參數：epsilon = 0.005, epoch = 10

   結果：Success rate = 0.995, L-inf. norm = 3.0000

   因為會跑多個 epoch，所以可以用比較小的 epsilon，同時能達到較低的 L-inf. norm。而每個 epoch 會重新計算 loss 和 gradient，可以朝最佳的方向 update image，在重複多個 epoch 的情況下，即使用很小的 epsilon 也可以達到相當高的攻擊成功率。hw5_best.sh 的方法比起 FGSM，可以達到更高的 success rate 及更低的 L-inf. norm。

2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。

	hw5_fgsm.sh	hw5_best.sh
Proxy model	ResNet50	ResNet50
Parameters	epsilon = 0.3	epsilon = 0.005, epoch = 10
Success rate	0.885	0.995
L-inf. norm	18.0000	3.0000

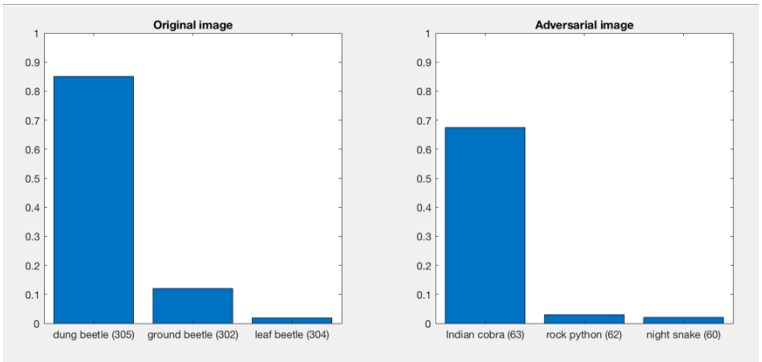
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

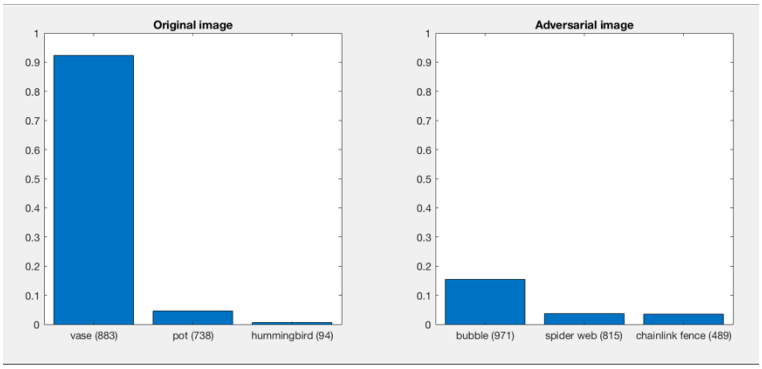
   使用 FGSM，epsilon = 0.3，不同 proxy model 的實作結果如下：

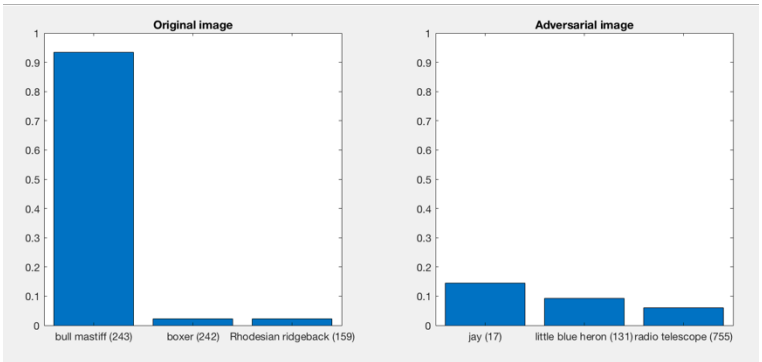
Proxy model	VGG16	VGG19	ResNet50	ResNet101	DenseNet121	DenseNet169
Success rate	0.605	0.595	0.885	0.700	0.670	0.695
L-inf. norm	18.0000	18.0000	18.0000	18.0000	18.0000	18.0000

   不同 proxy model 的 L-inf. norm 皆相同，success rate 最高的是 ResNet50，因此推測背後的 black box 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)。

<table>
<tr><td colspan="2">000.png</td></tr>
<tr><td>  </td><td> 

Dung beetle 85.03% Indian cobra 67.54%</td></tr>
<tr><td colspan="2">001.png</td></tr>
<tr><td>  </td><td> 

Vase 92.35% Bubble 15.45%</td></tr>
<tr><td colspan="2">002.png</td></tr>
<tr><td>  </td><td> 

Bull mastiff 93.35% Jay 14.48%</td></tr>
</table>

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。

使用 Gaussian filtering (sigma = 1.0)的方法將 hw5_best.sh 產生的 adversarial image 平滑化。以 012.png 為例，可以看出通過 Gaussian filter 後，影像變得比較模糊。加入原始影像的雜訊被平滑化會使模型比較不容易誤判。

Original image	通過 Gaussian filter
	

原本的 adversarial image 的攻擊成功率為 0.995。

經過 Gaussian filtering 後的攻擊成功率為 0.785。

將 adversarial image 平滑化確實會降低模型誤判的比例。