學號:B06705057 系級:資工二 姓名:黃資翔

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)

hw5_best 是用 FGSM 並迭代 5 次的結果， proxy model 是 pytorch 的 resnet50，每次迭代完會檢查有沒有超過 0.01，有的話會把他拉回來。我發現只要找到正確的模型，只要稍微調一點參數，就很容易攻擊，但我下面想討論當 proxy model 不正確時發生的情況。選定一個替代模型時，即使是 FGSM 也可以輕鬆癱瘓該模型，但其他的模型幾乎沒影響，也就是說純粹使用 FGSM 效果很差 (畢竟現實中要完全猜對模型很難)。因此我同時使用多個模型去迭代，會發現每個訓練的模型都可以輕鬆的誤導，而且若攻擊其他非訓練的模型也有部分的效果，當然也有些模型仍會無效。

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)。

	Proxy model	Success rate	L-inf norm
hw5_best.sh	Resnet50 (pytroch)	0.955	5.5450
hw5_fgsm.sh	VGG16 (keras) 選效果最好的	0.825	23.0
		0.23	6.0
hw5_fgsm	Resnet50 (keras)	0.130	4.650
hw5_multimodel	Resnet50, Resnet101, Densenet121, Densenet169 (keras)	0.500	4.98

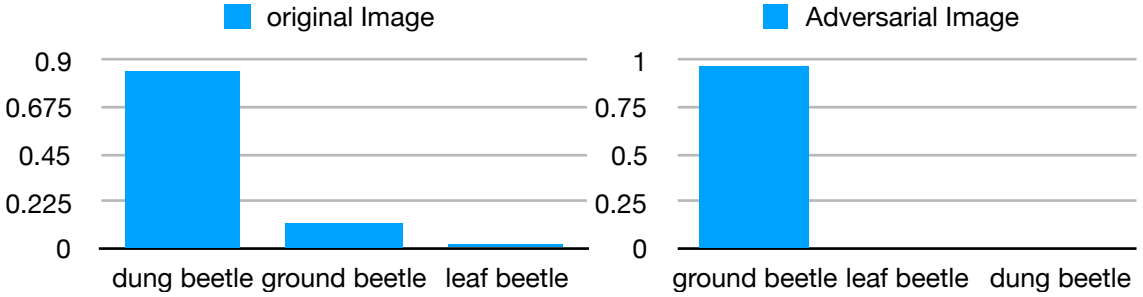
配合第一題， hw5_multimodel 沒有使用 pytorch resnet50 進行攻擊，仍有部分的效果，且可以看出 keras 與 pytorch 的 resnet50 是不同模型，但 hw5_multimodel 仍有一半的成功率。

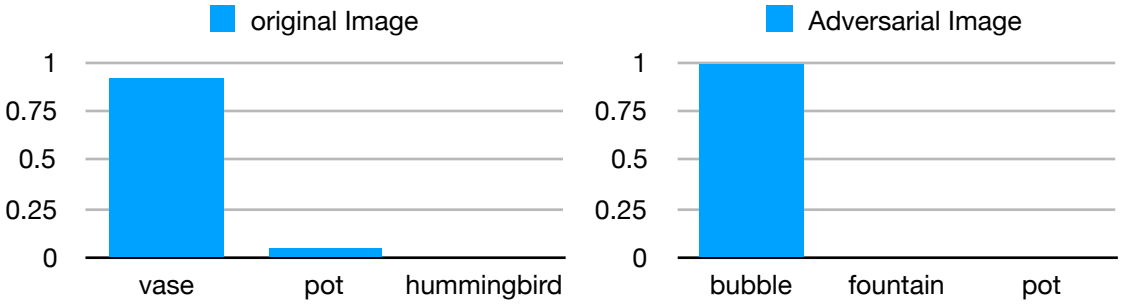
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

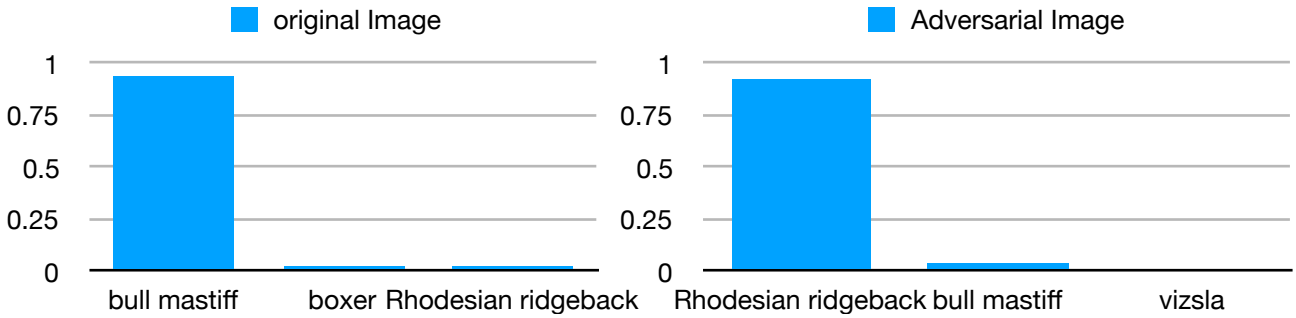
Proxy Model	Success rate	L-inf norm
VGG16	0.05	5.5400
VGG19	0.055	5.5300
Resnet50	0.955	5.5450
Resnet101	0.130	5.5300
Densenet121	0.085	5.5650
Densenet169	0.085	5.5600

很明顯背後的 black box 是 Resnet50，以上皆與 hw5_best.sh 一樣迭代 5 次，我發現當 proxy model 是錯誤的時候，迭代越多次，成功率會越低。

4. 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)。







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。

	Is Defense?	Success rate
Original Images	**Before defense**	0.0
	After defense	0.115
Adversarial Images	**Before defense**	0.955
	After defense	0.470

我對圖片進行水平與垂直的模糊化，模糊化矩陣如下。即 diagnol 是 1/2，diagnol 正負1是 1/4。將原始圖片左右兩邊乘此矩陣，左矩陣代表對垂直方向模糊，右矩陣代表對水平方向模糊。若要模糊數次可以在矩陣上加次方，但上表為各模糊一次的結果。原始圖片中約有一成會受影響，即原本正確預測會變錯誤預測，但對抗性圖片成功率會大幅下降五成。若要使原始圖片不要錯誤預測，可以在訓練模型時先讓資料做模糊化。

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 & \dots & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \dots & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$