Machine Learning HW5 Report

學號 : b06902028　系級 : 資工二　姓名 : 林柏劭

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)

   答:

   在 hw5_best.sh 中，我使用 iterative FGSM, proxy model=resnet50, epilson=0.01, epochs=100，並設了一個 limit=1 避免 L-infinity 過大，且有對圖片做 preprocess，使用 torch 的方法(除以 255，減去 ImageNet 平均[0.485, 0.456, 0.406]，除以標準差[0.229, 0.224, 0.225])。

   此方法比起一般的 FGSM，分批次增加 noise，因此能在同樣的 L-infinity 下有更大的機率攻擊成功。

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)。

   答:

	Proxy model	Success rate	L-inf. norm
hw5_fgsm.sh	resnet50	0.910	5.6550
hw5_best.sh	resnet50	0.995	4.6000

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。

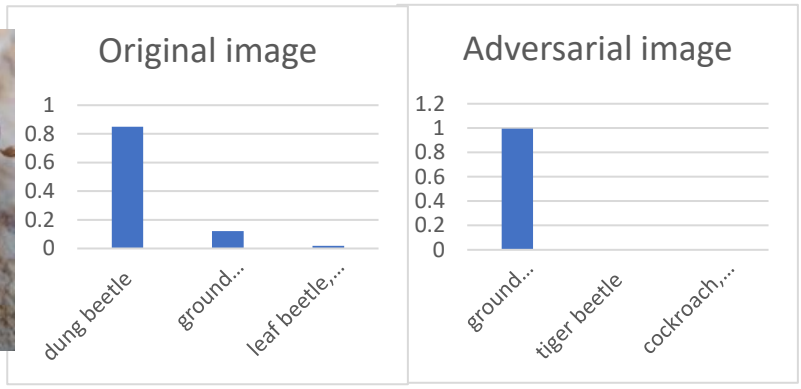
   答:

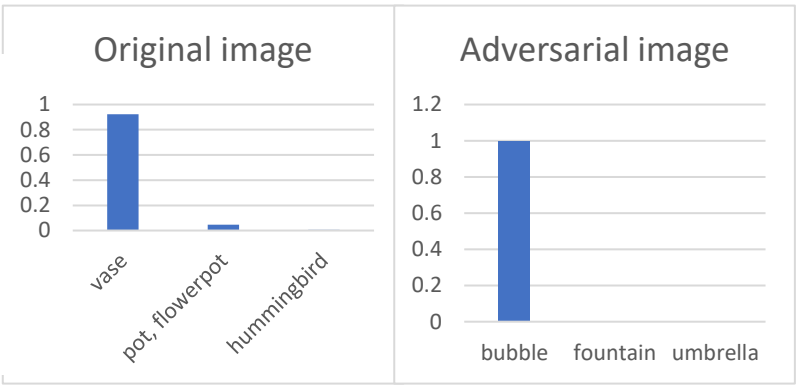
	Success rate	L-inf. norm
vgg16	0.150	5.6450
vgg19	0.140	5.6500
resnet50	0.910	5.6500
resnet101	0.335	5.6900
densenet121	0.245	5.6100
densenet169	0.250	5.6500

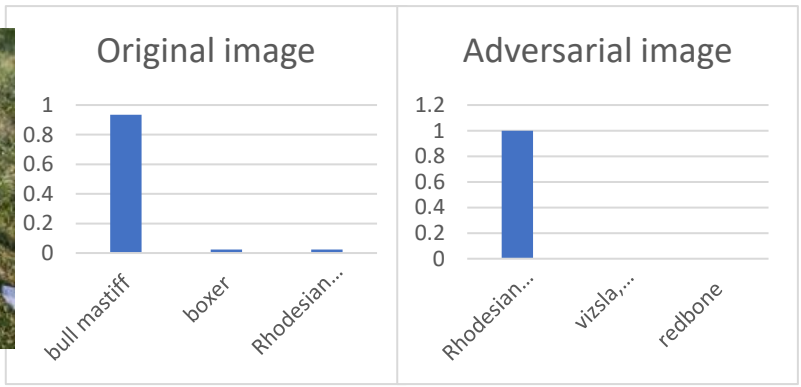
   在其他條件都一樣的情況下，使用 resnet50 作為 proxy model 的結果遠比其他的好，因此我認為背後的 black box 最有可能是 resnrt50。

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)。

答:







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。

答:

	Success rate	L-inf. norm
hw5_best.sh	0.995	4.6000
after defense	0.280	12.6450

我使用 median filter 的方法，對於防禦頗為成功。壞處是圖片本身顏色會失真，如下: