

Machine Learning HW5 Report

學號：b05901063 系級：電機三

姓名：黃世丞

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也是使用FGSM，但是和hw5_fgsm使用不同的proxy model，參數也經過仔細調整。hw5_fgsm是用vgg19，hw5_best是使用resnet50，epsilon是0.06，雖然直覺來說epsilon越大會增加攻擊成功率和L-infinity，但是實際實驗發現epsilon從0.3開始往下調，可以同時增加攻擊成功率和降低L-infinity，直到差不多epsilon=0.06為止，再降低就會讓成功率下降。

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_fgsm.sh	vgg19	0.58	18.0000
hw5_besh.sh	resnet50	0.92	4.0000

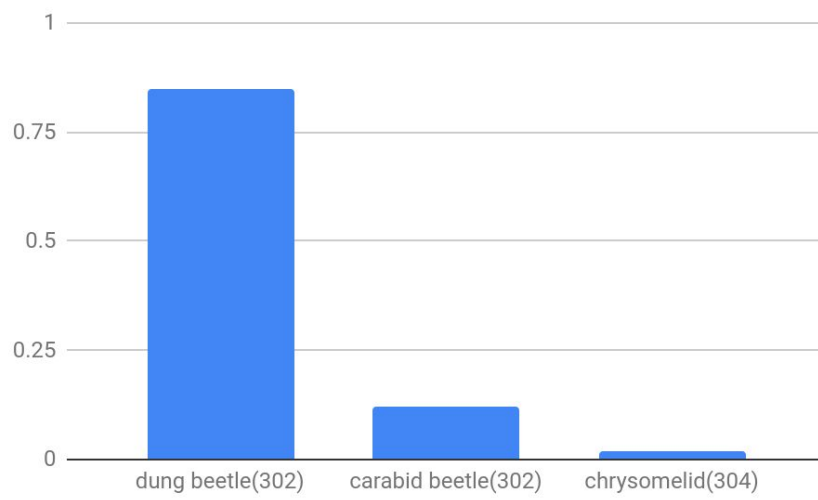
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

resnet50。無論參數怎麼調，resnet50做出來的結果會有非常明顯的差異，success rate大幅領先其他model。除了resnet50以外的model很難過strong baseline (但simple還可以過)。

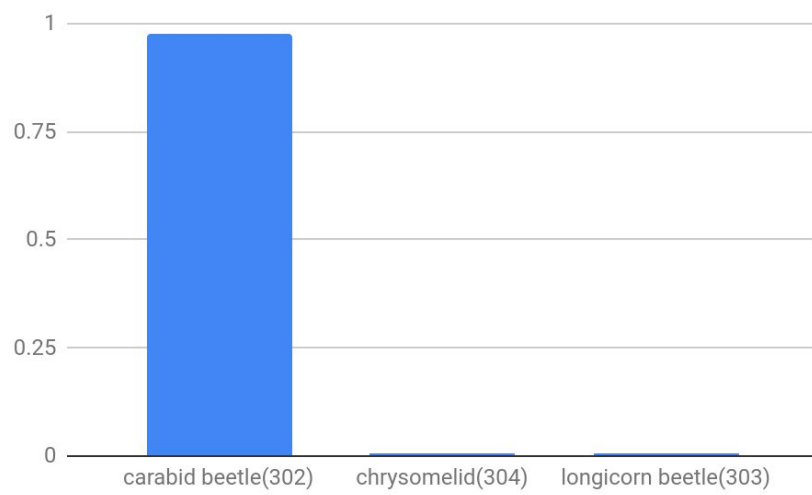
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。



Original

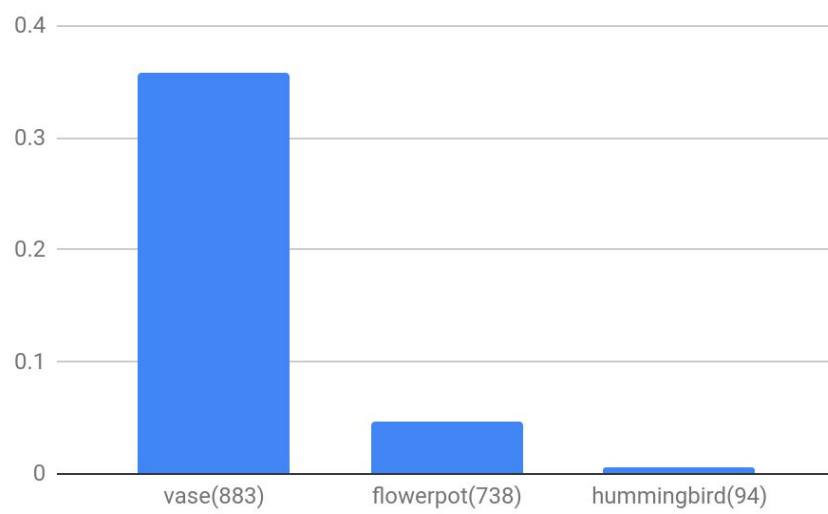


Adversarial

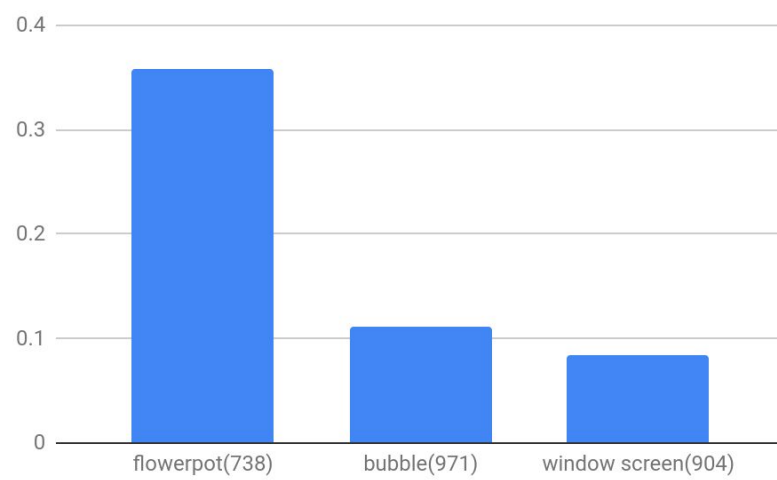




Original

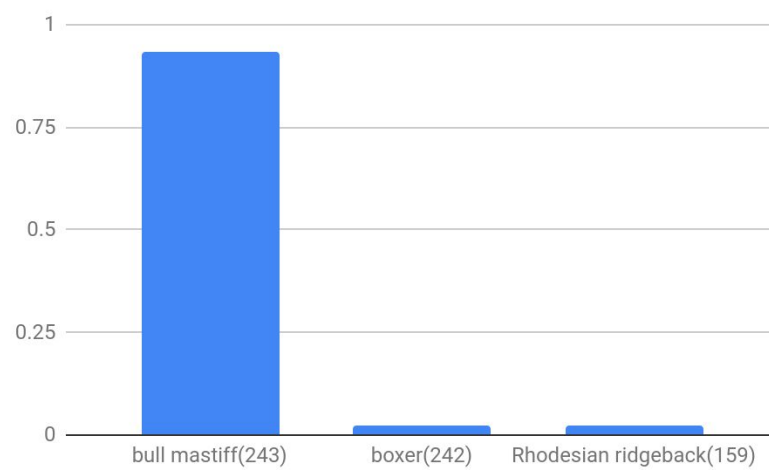


Adversarial

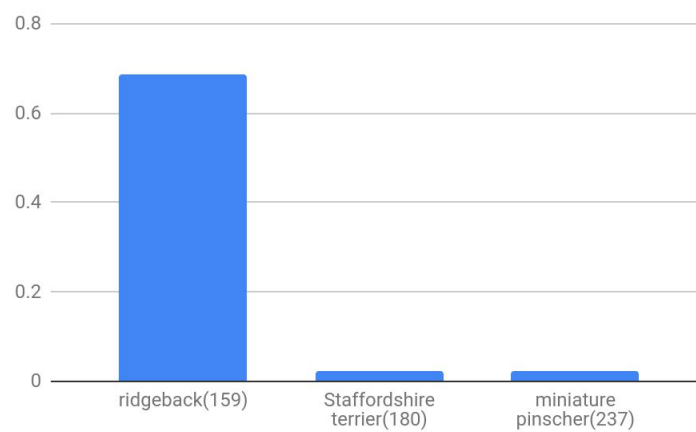




Original



Adversarial




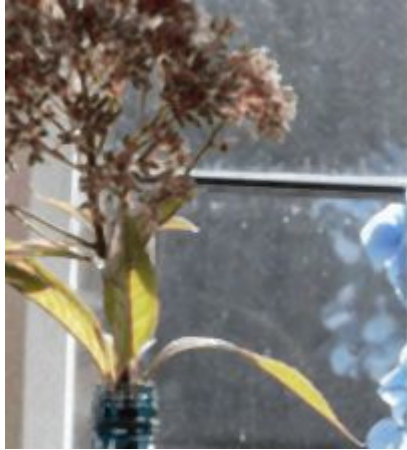


5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

我使用scipy.ndimage.median_filter，他會取設定好的大小的格子內的中位數當作中心點的值，我把kernel_size設為3，也就是每3x3x3的格子的中心點的值會換成原本的值的的中位數。防禦前後的success rate數值如下：

防禦前：0.92 防禦後：0.675

圖片比較如下：

防禦前	防禦後
	
	



可以發現防禦後的圖和原圖比起來較為模糊，而且在彩度比較高(單一RGB值較大)的地方會有明顯的色偏，推測這是因為有一個通道的RGB值特別高，但是取中位數的情況下會大幅降低原本的值，才會讓圖片變得較為黯淡。而success rate大幅降低了0.245，圖片也沒有太誇張的失真，是還不錯的防禦手段。