Machine Learning HW5 Report

學號：B05901025 系級：電機三　姓名：王鈺能

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)

   我的 hw5_best.sh 攻擊方法就是 FGSM(可通過 strong baseline)，proxy model 為 resnet50，方法是算出每個 pixel 對 output 的 gradient 取正負 1 再乘以 epsilon 遠離原本的預測結果，參數只有 epsilon=0.0364。因為方法即是 FGSM，因此結果留於第 2、3 題討論。

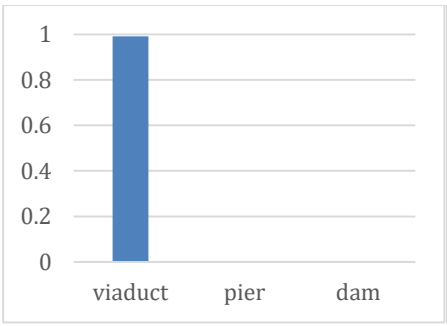
2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。

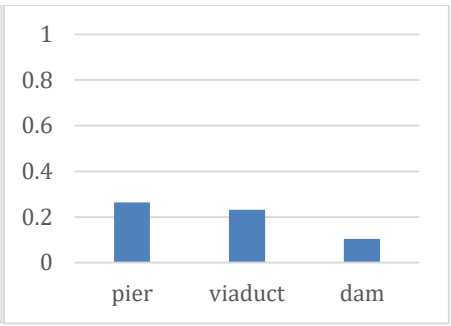
	Proxy model	Success rate	L-inf. norm
best(FGSM)	Resnet50	0.915	3.00

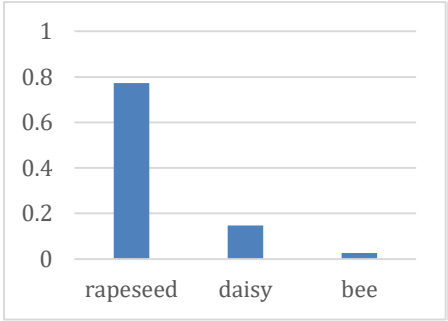
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

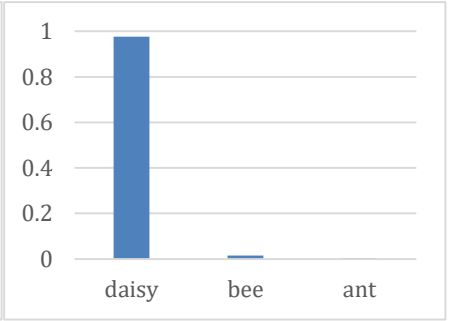
背後的 proxy model 應為 resnet50，因為只有 resnet50 在直接丟沒有 raw image 進去時是全部正確辨識的，其他都會和給的 label 有所出入；而且同樣做法下，其他 model 表現皆較差(success rate: vgg16=0.635, denseNet121=0.400；此外，還必須是 pytorch pretrain 的 resnet50，用 tensorflow 的 pretrain 參數和 pytorch 不同，攻擊效果差很多(success rate=0.495)。

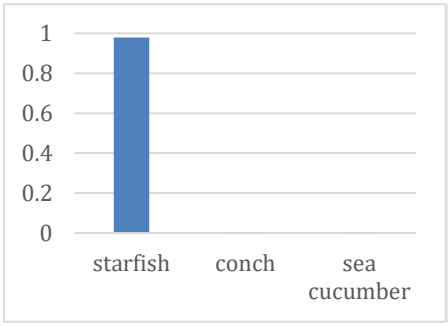
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

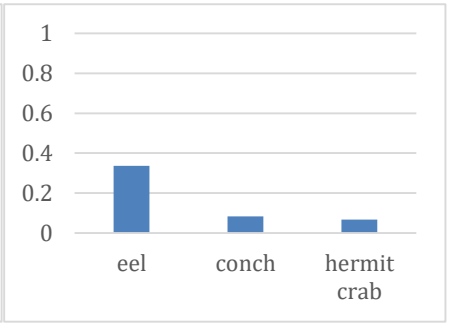









5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

	Proxy model	Success rate	L-inf. norm
best(after filtering)	Resnet50	0.595	137.1050

我使用 scipy 內建的 median filter(scipy.ndimage.median_filter)，使用後明顯的 success rate 降低很多，但相對的圖片變得模糊了，L-inf norm 變得很大。