Machine Learning HW5 Report

學號：B05705053 系級：資管 姓名：蔡涵如

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best在leaderboard上面是用keras vgg16實作，做完preprocees image，先降低top1的confidence，再升高將完之後第一名的confidence，在做postprocess把還原回去attack後的照片，但在leaderboard success rate大概只有0.45左右 L-inf 5，然後後來deadline之後我有使用pytorch在實做一次攻擊resnet50，使用L-inf 5的FGSM就達到成功率0.92，我的推測是因為preprocess的時候tensorflow backend會除127.5再-1，而pytorch則是除255，導致後面用的std跟mean雖然是一樣的，但是preprocess的方法不一樣，或是model weight不一樣就導致結果差異非常大，所以猜測到proxy的package也是一件很重要的事情。

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

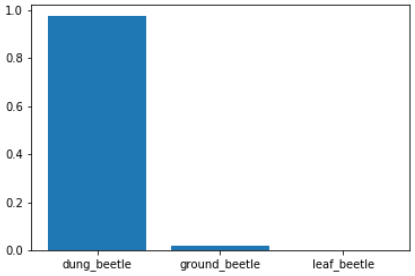
	Proxy	Success Rate	L-inf
hw5_fgsm.sh	VGG16	0.205	5.0000
hw5_best.sh	VGG16	0.435	4.9600

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。

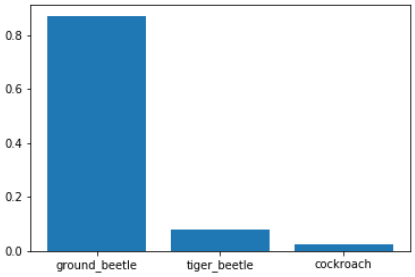
我推測是由resnet50的模型，因為在leaderboard上面做的同一個方法是resnet50是最好的，甚至用fgsm的L-inf=5就可以攻擊到0.92

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)。

000.png

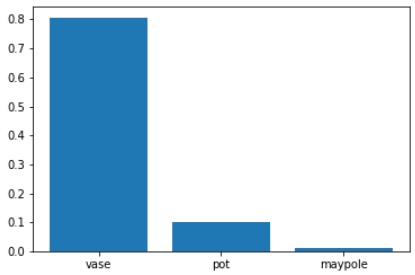


97%

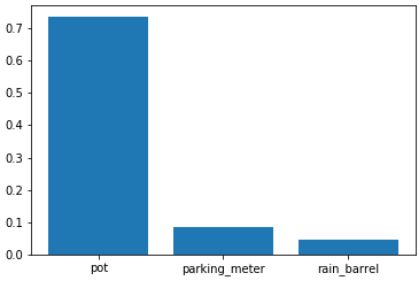


87%

001.png

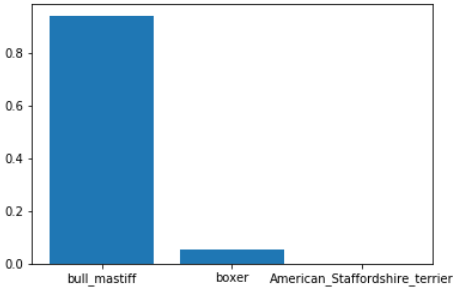


80%

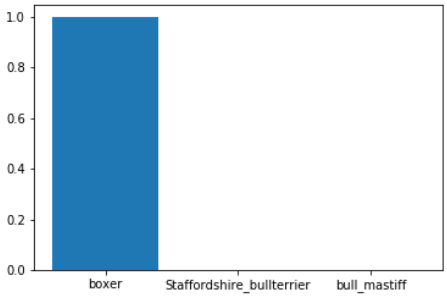


73%

002.png



94%



99.8%

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。

原本我的model再local端攻擊，會100%的完全攻擊成功，我將RGB每一維做gaussian filter image_location_generator由下面兩張圖可以看到左邊的曲線是做過defense可以將曲線變的平滑一些，而在local成功率大概降到大概51%左右。