Machine Learning HW5 Report

學號：b04104040　系級：工海三　姓名：解正安

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。

採用 torchvision.models 中 pretrain 好的 resnet50 去進行攻擊。方法分兩階段，第一步是先用 Iterative Method 去實現 FGSM，iteration= 500，epsilon=0.001，如果該張圖片已攻擊成功就跳出 iteration。跑完後檢查圖片攻擊結果，會發現['006', '061', '074', '090', '092', '094', '098', '102', '121', '142', '156', '165', '182']這幾張特別難攻擊(攻擊失敗)，這時用 “攻擊後”的圖片做同樣的第二次 Iterative Method，iteration= 500，epsilon=0.001，第二次後會剩下['121']。針對這張圖片，直接用已攻擊過圖片，用 FGSM 手動調 epsilon，epsilon 為 0.7 即可達到 success rate=1.0，L-infinity= 1.7950 的結果。

使用 Iterative Method 比單純用 FGSM 的好處是可以省時，他事實上就是 FGSM 用很小的 step size 去慢慢找最佳解，這樣比起一次加很多的躁點，更能為每張客製化適合的躁點，也不用自己慢慢找。而難攻擊的圖片則適合一次加大量躁點，雖然也可以用 Iterative Method，但這裡我直接手動調，也能得到不錯的結果。

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	hw5_best.sh
proxy model	resnet50	resnet50
success rate	0.690	1.00
L-inf. norm	1.0000	1.7950

hw5_fgsm.sh 的 eplison 為 0.007

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。

使用 hw5_best.sh 的結果:

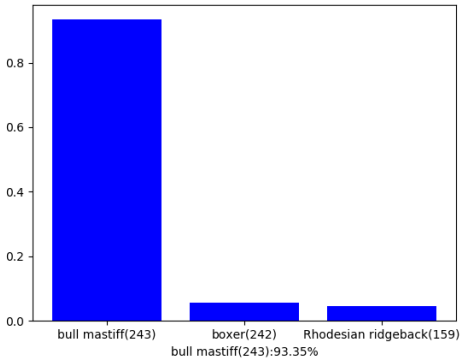
Model	Success rate
VGG-16	0.025
VGG-19	0.025
ResNet-50	1.00
ResNet-101	0.090
DenseNet-121	0.055
DenseNet-169	0.040

black box 最有可能是 ResNet50。從結果可以知道我們所做的攻擊主要在 fit model，只有攻擊對的 model 才會有較高的準確率，如果是不同 model 的攻擊結果，並不一定是能騙過其他的 model。

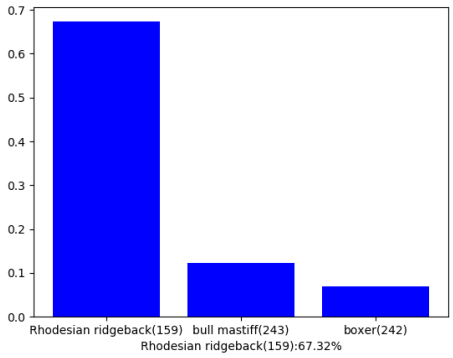
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前

三高的機率)。

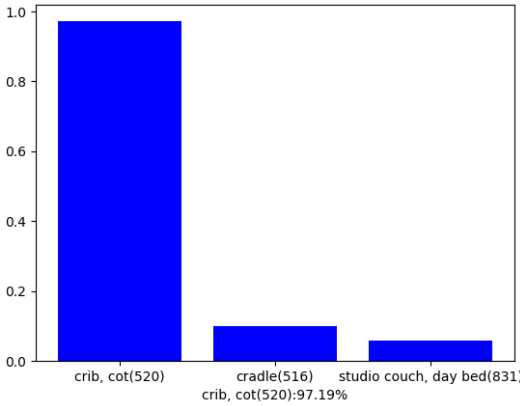
Origin images



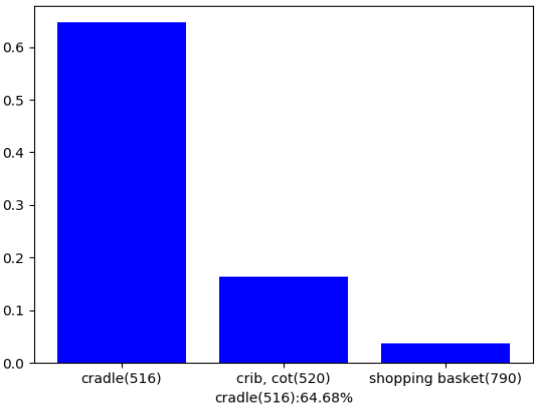
Adversarial images



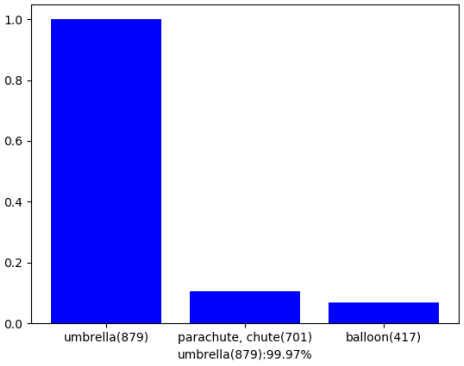
Origin images



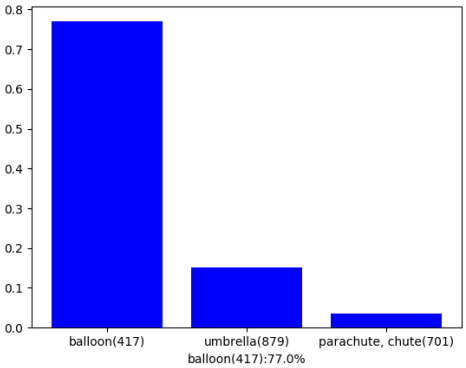
Adversarial images



Origin images



Adversarial images



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，附上你攻擊有無的 success rate，並簡要說明你的觀察。

採用 Median filter 的方式，將圖片使用 median filter 可以有效降低一些雜訊(如 Salt and pepper noise)。因此我們在攻擊時所產生的部分高斯雜訊可以因此而濾除或是降低效果，實際圖片如圖:

Attack Figure

Defense Figure

使用 ndimage.median_filter，參數為 3

Success rate:

Attack Figure	Defense Figure
1.00	0.295