學號：R07943150　系級：電子碩一　姓名：吳辰鋐

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)

Best 的 Proxy model 為 Pytorch 的 resnet50，方法為將 FGSM iterative 10 次，將每次 FGSM 產生的結果再帶回 FGSM 並將 L-infinity 與原圖比較做限制，結果差異如題 2，每次的步長設定為 0.5/(255*0.224)，L-infinity 也是 0.5/(255*0.224)，除 255 及 0.224 乃因資料 preprocessing，另外不設定 1 而設定成 0.5 乃因會自動變化成標準化後的一單位，若設定成 1 會進位成標準化後的兩單位

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 皆為 Pytorch 的 resnet50

	Success Rate	L-infinity
hw5_fgsm	0.73	1
hw5_best	1	1.02

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。

Pytorch	Success Rate	L-infinity
Resnet50	1	5
Resnet101	0.56	5
Vgg16	0.275	5
Vgg19	0.305	5
Densenet121	0.445	5
Densenet169	0.435	5

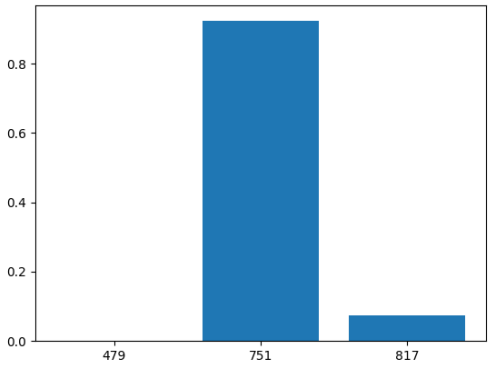
由以上數據觀察，在相同 L-infinity 限制下，很明顯的可以發現背後模型應為 Resnet50

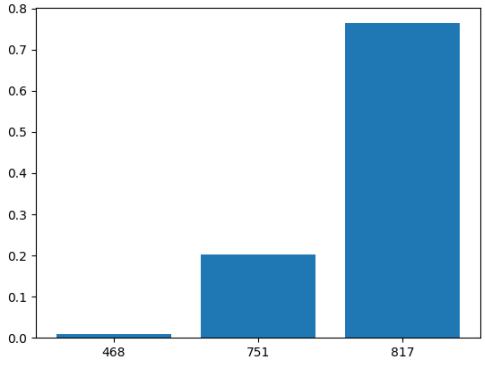
Keras	Success Rate	L-infinity
Resnet50	0.5	5.95
Resnet101	0.425	5.975
Vgg16	0.5	5.975
Vgg19	0.5	5.975
Densenet121	0.445	5.975
Densenet169	0.365	5.875

Keras resnet50 和 Pytorch resnet50 在差不多的 L-infinity 下成功率差非常多 QQ

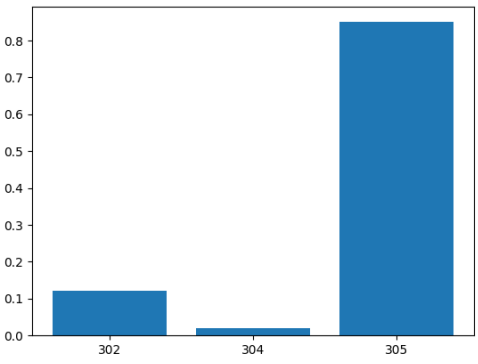
可得知即使知道模型，不同訓練方式得到的參數也對攻擊效果影響很大

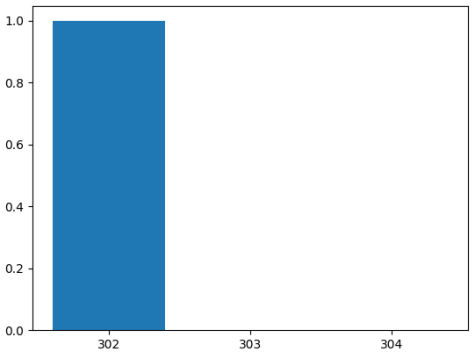
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。



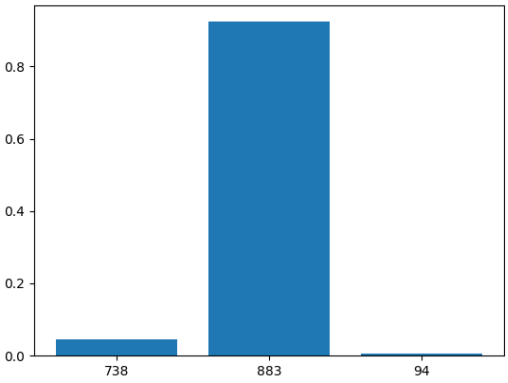


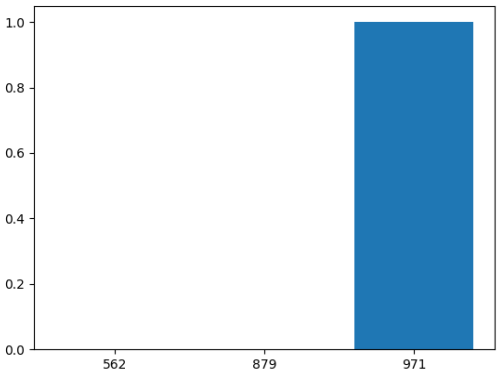
Original　　Before: race car　　After: sport car





Original　　Before: dung beetle　　After: ground beetle





Original　　Before: vase　　After: bubble

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。

	Success Rate	L-infinity
防禦前	1	1.02
防禦後	0.315	131.7100

原本攻擊效果很好的圖片，再使用 Median filter 過後即使 L-infinity 很大，success rate 依然很低，另外附上直接將原圖進行防禦的數值，Success rate = 0.1 L-infinity=131.42，可見防禦本身並不會造成黑盒子誤判，並使得攻擊僅能將 Success rate 從 0.1 拉至 0.315