說明：請各位使用此 template 撰寫 report，如果想要用其他排版模式也請註明題號以及題目內容（請勿擅自更改題號），最後上傳至 GitHub 前，請務必轉成 PDF 檔，並且命名為 report.pdf，否則將不予計分。

中英文皆可，但助教強烈建議使用中文。

----------------------------------------閱讀完以上文字請刪除---------------------------------------

學號： 系級： 姓名：

1. (2%) 試說明 hw6\_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)

for several e with same proxy model

check acc and diff

原法：

e0.1=classification acc: 0.085 17 200 L-inf 5.55

e0.15=classification acc: 0.085 17 200 L-inf 8.325

e0.1625=classification acc: 0.085 17 200 L-inf 8.325

e0.175=classification acc: 0.09 18 200 L-inf 9.25

e0.2=classification acc: 0.095 19 200 L-inf 11.1

e0.3=classification acc: 0.065 13 200 L-inf 16.65

e0.35=classification acc: 0.075 15 200 L-inf 18.5

e0.3625=classification acc: 0.07 14 200 L-inf 19.425

e0.375=classification acc: 0.07 14 200 L-inf 20.35

e0.3875=classification acc: 0.065 13 200 L-inf 21.275

e0.4=classification acc: 0.065 13 200 L-inf 21.275

畫長條圖

black box的選擇則詳見problem2

那這樣同學可以說明你怎麼調整 FGSM 達到 strong baseline 的，途中有什麼想法，做了什麼嘗試，結果如何之類的。

1. (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。

For all model same e=0.4

check acc and diff

vgg16 , local 8/200, 19.895

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2020-04-25 23:31:51 | 0.565 | 19.8950 |

vgg19 , local 9/200, 20.01

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 9 | 2020-04-26 02:57:49 | 0.550 | 20.0100 |

resnet50 , local 18/200, 23

|  |  |  |  |
| --- | --- | --- | --- |
| 5 | 2020-04-26 02:49:59 | 0.720 | 23.0000 |

resnet101 , local 26/200, 21.39

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 8 | 2020-04-26 02:56:22 | 0.620 | 21.3900 |

densenet121, local 13/200, 21.275

|  |  |  |  |
| --- | --- | --- | --- |
| 6 | 2020-04-26 02:51:32 | 0.935 | 21.2750 |

densenet169, local 26/200. 21.045

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 7 | 2020-04-26 02:54:48 | 0.740 | 21.0450 |

1. (1%) 請以 hw6\_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)。

Plot acc before max

1. (2%) 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。

Smoothing to defense