Machine Learning HW5 Report

學號：B06902030　系級：資工二　姓名：邱譯

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)

   方法：iterative FGSM

   proxy model：resnet50

   參數：epsilon = 0.01，iteration = 4

   正確率：0.995，L-Inifinity：3.0000

   此方法執行FGSM多次，當epsilon = 0.01、iteration = 4時與epsilon = 0.04、iteration = 1時相比，做多次的iteration每次可以走較小一步，且每次都往gradient方向走，因此可以走到更接近loss較高的地方，因此攻擊成功率可以較好。

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 model: resnet50
   - success rate: 0.915
   - L-inf. norm: 3.0000

   hw5_best.sh:

   - proxy model: resnet50
   - success rate: 0.995
   - L-inf. norm: 3.0000

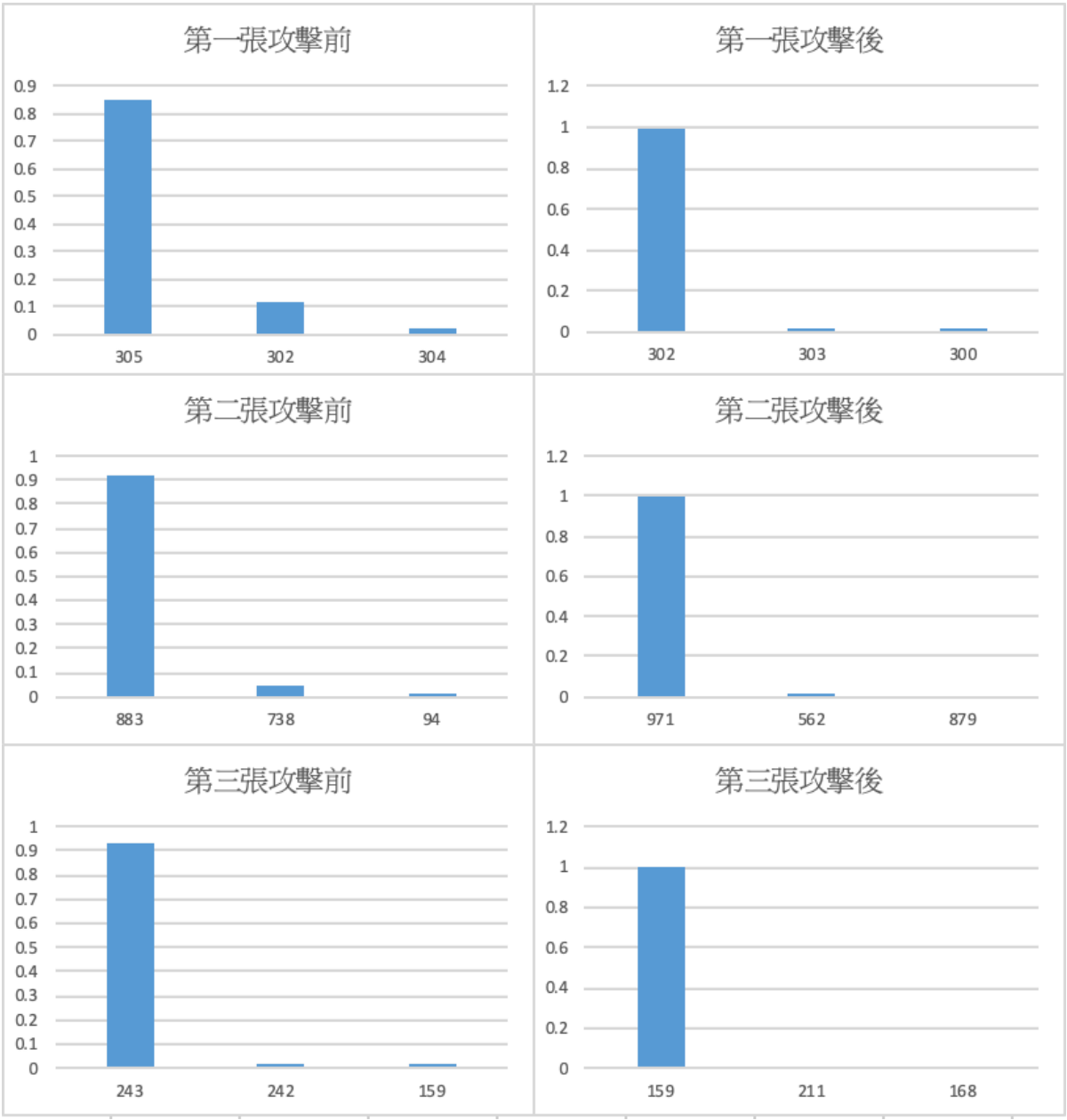
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

   攻擊成功率

   - VGG-16: 0.155
   - VGG-19: 0.140
   - ResNet-50: 0.995
   - ResNet-101: 0.345
   - DenseNet-121: 0.205
   - DenseNet-169: 0.245

   觀察可知ResNet-50的攻擊成功率最高，因此推測black box最有可能為ResNet-50

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。

   方法：medfilt

   防禦前成功率：0.995

   防禦後成功率：0.315

   經過medfilt防禦後，攻擊成功率大幅下降，表示此種防禦方式有效，但攻擊成功率仍有0.315，仍然蠻高的，因此如果要進行好的防禦仍需要再使用其他方法或綜合使用。

   此防禦會使原始圖片顏色失真，但整體大致上不變，因此仍可判斷圖片內容物為何。