

Machine Learning HW5 Report

學號：B05901005 系級：電機三

姓名：賴沂謙

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)

在best裡面，我也是一樣使用FGSM的方法，proxy model是使用pretrained的resnet50、epsilon是0.065。因為使用的應該是和black box一樣的model，所以成功率相當高。而epsilon設成這樣可以使L-inf. norm維持很小，但一樣達到高的攻擊成功率。

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 | Success rate | L-inf. norm |
|------|--------------|--------------|-------------|
| fgsm | DenseNet-121 | 0.320 | 4.000 |
| best | ResNet-50 | 0.920 | 4.000 |




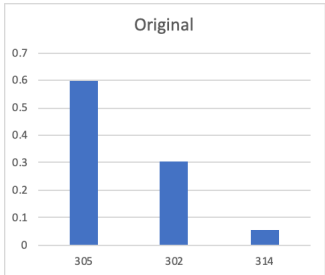
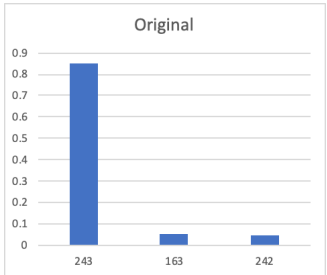
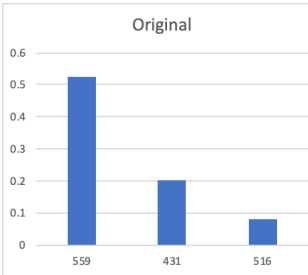



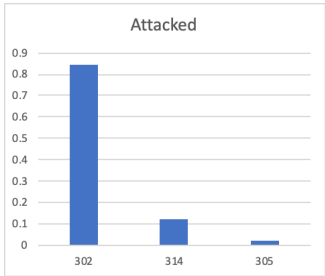
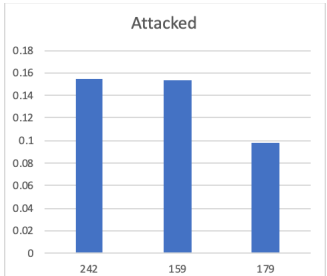
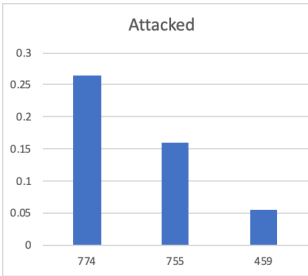
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

| Proxy model | Success rate | L-inf. norm |
|--------------|--------------|-------------|
| VGG-16 | 0.270 | 4.000 |
| VGG-19 | 0.300 | 4.000 |
| ResNet-50 | 0.920 | 4.000 |
| ResNet-101 | 0.425 | 4.000 |
| DenseNet-121 | 0.320 | 4.000 |
| DenseNet-169 | 0.325 | 4.000 |

每個proxy model所使用的epsilon都是0.065，然後從成功率結果看來，助教背後使用的back blox應該是ResNet-50，因為成功率比其他的好很多。

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)。

| 攻擊前 圖片 |  |  |  | | | | | | | | | | | | | | | | | | | | | | | | |
|-----------|--|---|--|-----|------|-----|------|-----|------|--|----------|-------------|-----|-------|-----|-------|-----|------|--|----------|-------------|-----|------|-----|------|-----|------|
| 機率類別 | <p>Original</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>305</td><td>0.6</td></tr><tr><td>302</td><td>0.3</td></tr><tr><td>314</td><td>0.05</td></tr></tbody></table> | Category | Probability | 305 | 0.6 | 302 | 0.3 | 314 | 0.05 | <p>Original</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>243</td><td>0.85</td></tr><tr><td>163</td><td>0.05</td></tr><tr><td>242</td><td>0.05</td></tr></tbody></table> | Category | Probability | 243 | 0.85 | 163 | 0.05 | 242 | 0.05 | <p>Original</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>559</td><td>0.52</td></tr><tr><td>431</td><td>0.2</td></tr><tr><td>516</td><td>0.08</td></tr></tbody></table> | Category | Probability | 559 | 0.52 | 431 | 0.2 | 516 | 0.08 |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 305 | 0.6 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 302 | 0.3 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 314 | 0.05 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 243 | 0.85 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 163 | 0.05 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 242 | 0.05 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 559 | 0.52 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 431 | 0.2 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 516 | 0.08 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 攻擊後 圖片 |  |  |  | | | | | | | | | | | | | | | | | | | | | | | | |
| 機率類別 | <p>Attacked</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>302</td><td>0.85</td></tr><tr><td>314</td><td>0.12</td></tr><tr><td>305</td><td>0.03</td></tr></tbody></table> | Category | Probability | 302 | 0.85 | 314 | 0.12 | 305 | 0.03 | <p>Attacked</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>242</td><td>0.155</td></tr><tr><td>159</td><td>0.155</td></tr><tr><td>179</td><td>0.1</td></tr></tbody></table> | Category | Probability | 242 | 0.155 | 159 | 0.155 | 179 | 0.1 | <p>Attacked</p>  <table border="1"><thead><tr><th>Category</th><th>Probability</th></tr></thead><tbody><tr><td>774</td><td>0.26</td></tr><tr><td>755</td><td>0.16</td></tr><tr><td>459</td><td>0.05</td></tr></tbody></table> | Category | Probability | 774 | 0.26 | 755 | 0.16 | 459 | 0.05 |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 302 | 0.85 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 314 | 0.12 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 305 | 0.03 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 242 | 0.155 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 159 | 0.155 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 179 | 0.1 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Category | Probability | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 774 | 0.26 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 755 | 0.16 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 459 | 0.05 | | | | | | | | | | | | | | | | | | | | | | | | | | |

- (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。

我是用PIL的ImageFilter裡面的SMOOTH_MORE對我的hw5_best產生出來的圖片做smoothing，確實可以讓攻擊成功率下降。

| Smoothing 前攻撃成功率 | Smoothing 後攻撃成功率 |
|------------------|------------------|
| 0.920 | 0.875 |

