Machine Learning HW5 Report

學號：B04501073 系級：土木四 姓名：李利元

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)

   我使用的 proxy model 為 resnet50 訓練而成，攻擊方法不同於範例的地方是，我沒有全部的參數都是往 sign_()方向移動同樣 epsilon 的單位，我會依照 grad 的大小來決定 epsilon 的值(如果 grad 的值大於某一 threshold 值就直接給定值)，如此一來可以讓 loss 下降更多，此外我的 loss function 除了讓預測結果與真實 label 產生偏差，我增加一個 loss 讓預測值可以越靠近次高機率的 label，這個方法實作上也可以增加我攻擊成功的機率，可以在 proxy model 成功率達到 1，但最後在真實 model 上成功率仍然只有 0.6 我猜想可能是因為在 model 訓練過程沒有好好的資料前處理，讓 model 有點 overfitting，若有做一定程度的資料處理應該可以再提升成功率。

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)。

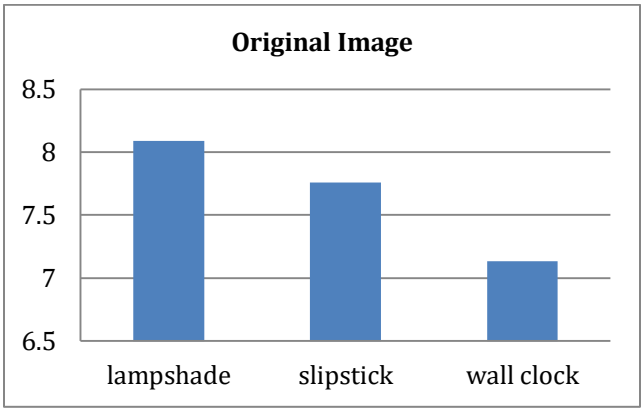
   皆使用 resnet50 訓練後的 model，hw5_fgsm 成功率為 0.535、L-inf 為 2.0， hw5_best 成功率為 0.61、L-inf 為 3.0。

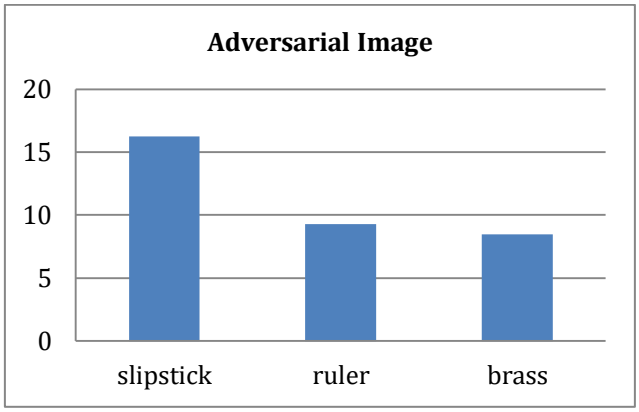
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

   在作業給定的 6 個 pretrained model 中直接進行攻擊，所得到的結果為 resnet50 的成功率可達到 0.45，其餘皆小於 0.2，以此結果判斷 black box 模型為 resnet50 為架構。

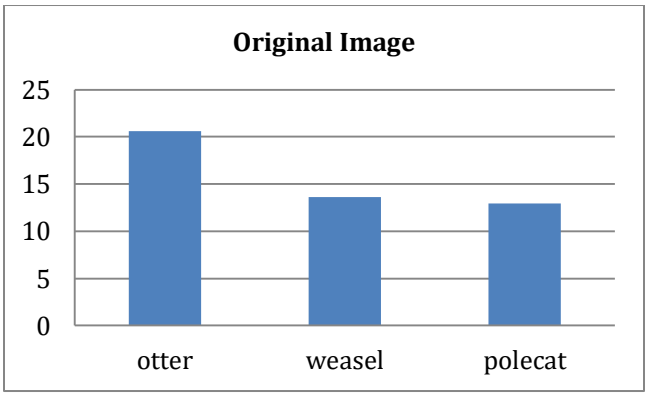
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

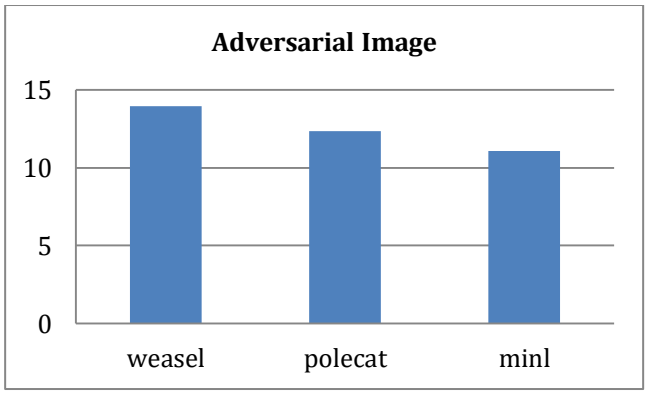
(1) 012.png



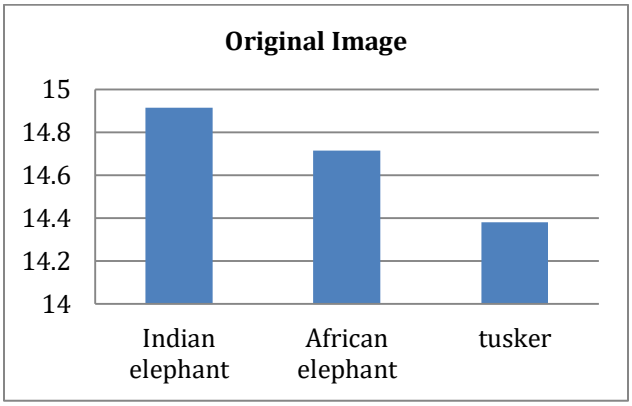


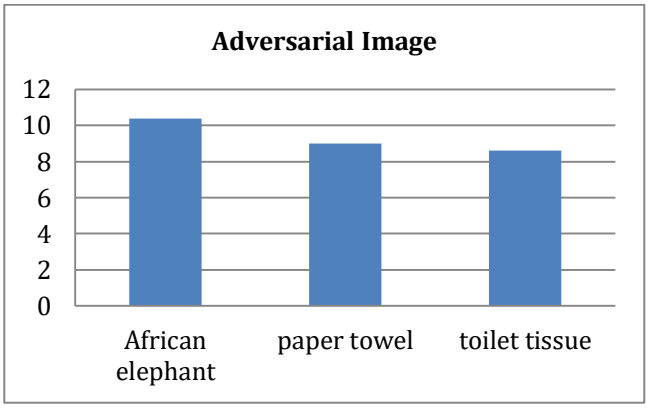
(2) 047.png





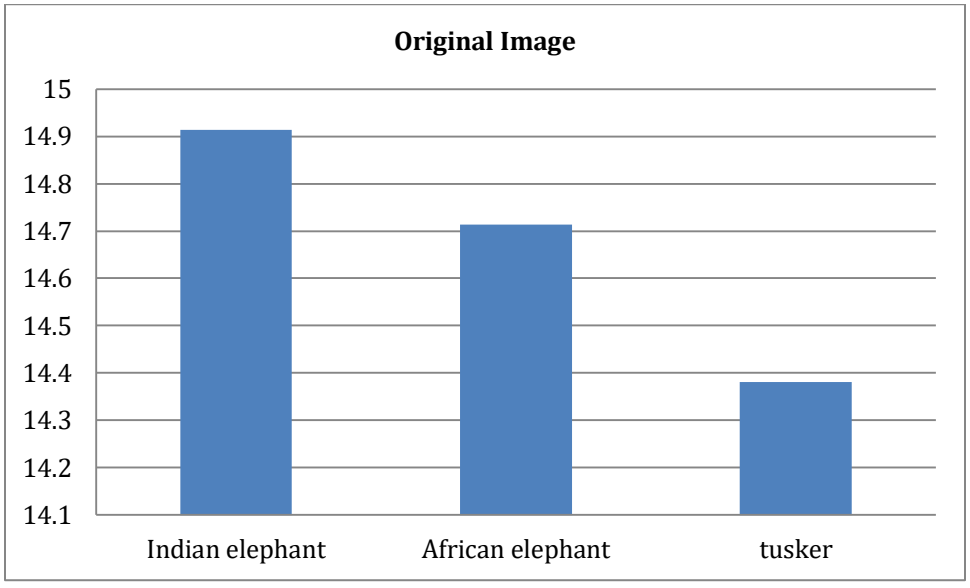
(3) 191.png

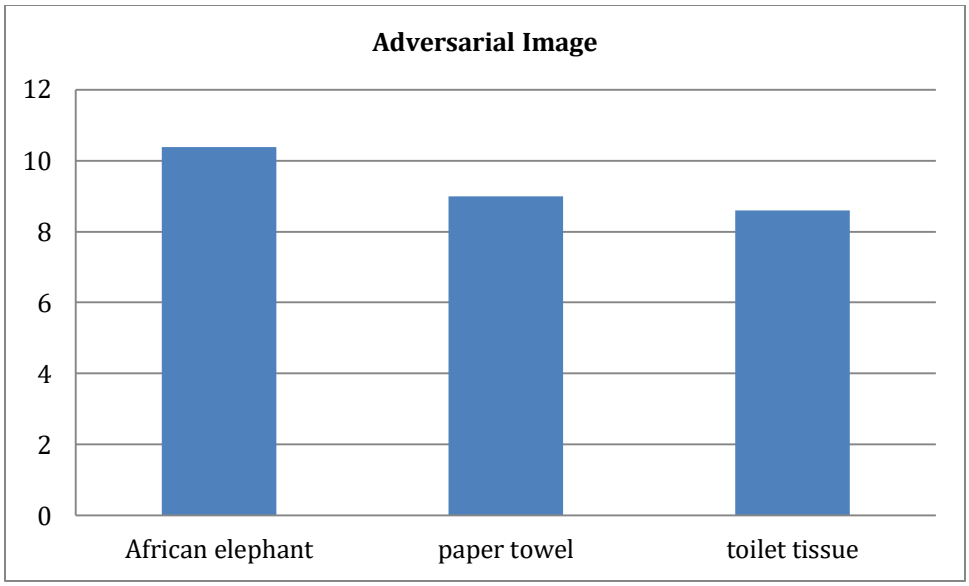


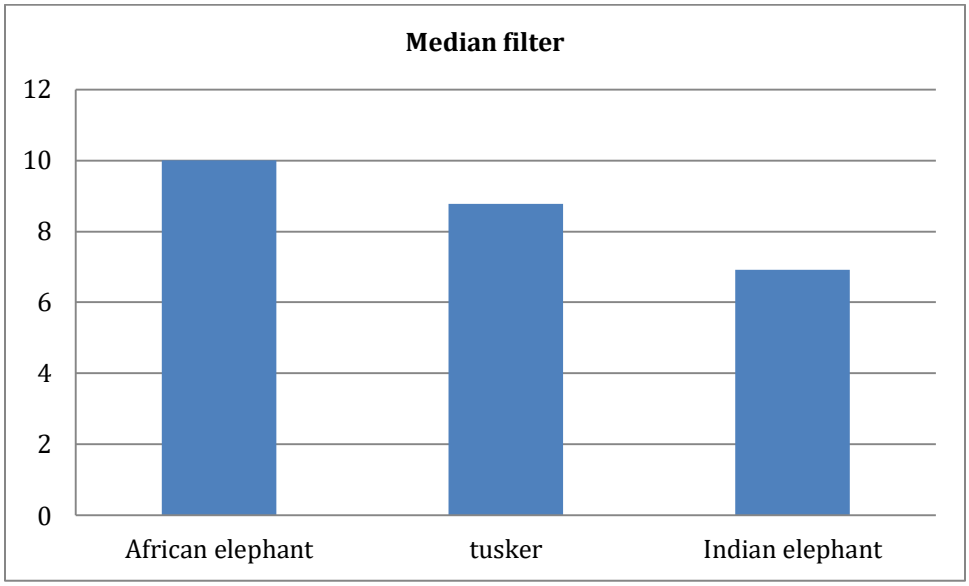


5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你攻擊有無的 success rate，並簡要說明你的觀察。

   我使用的方法為 median filter，實驗對象為 191.png 結果如下：







此圖片之 label 為印度象，結果雖然還是無法預測真正的 label，但是至少此 label 回到前 3 名的機率，不像毫無防備情況下攻擊，第二高及第三高的 label 都是跟印度象相差非常遙遠的物品。