Machine Learning HW5 Report

學號：b067005001 系級：資管二 姓名：楊力行

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)

   resnet50 方法是，隨機選個不是正確預測的物品，將它做為目標的 target attack，用多次疊代的 FGSM 進行 20 次的疊代，每次的步長為 1，並且限制雜訊不能超過 5pixel，原本的 FGSM 是一次就沿著梯度移動所限制的誤差的長度，而這個方法則是多次沿著梯度移動，但是移動多次，每次的步長較短，並且總共的誤差不能超過所限制的，可以在較小的誤差內影響到判斷的準確。

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 resnet101 0.620 16.0

   hw5_best.sh vgg16 0.455 5.0

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。

   pytorch 的 resnet50 我自己使用 KERAS，我的疊代 FGSM 在各模型中的表現都差不多，不太好，

   但截止後改用 pytorch 發現在 resnet50 的效果遠超其他模型，因此我認為是 resnet50，並且是 pytorch 所 train 出的

   resnet50 0.955/5.5300

   resnet101 0.120/5.5350

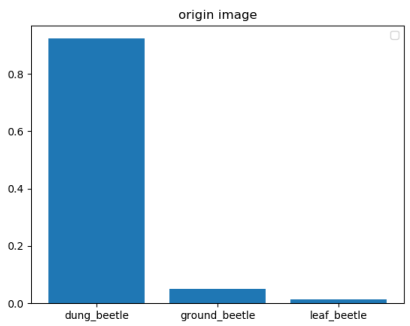
   Densenet121 0.085/5.5650

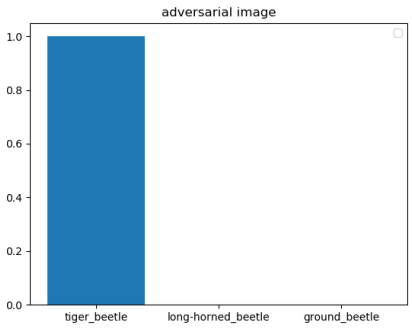
   Densenet169 0.085/5.5600

VGG16	0.05/5.5200
VGG19	0.06/5.5150

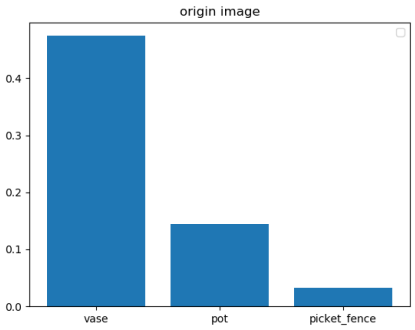
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

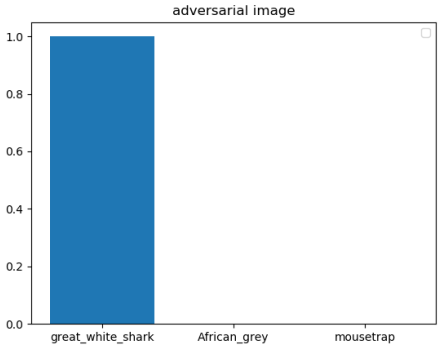
000



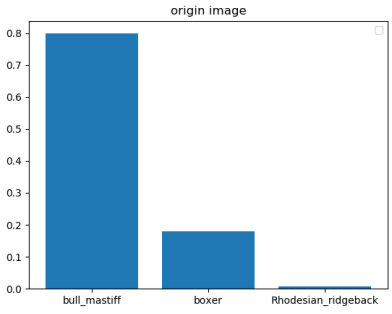


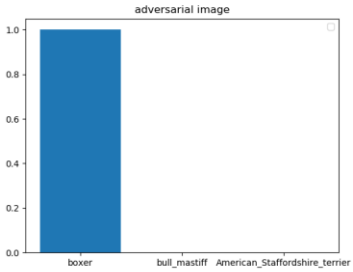
001





002





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。

使用 median filter filter 為 3，把周圍的中間值，當作是中間那格的值。

防禦前 0.455　　防禦後 0.460，沒能有效的降低，還提高了，我想是因為我的初始成功率太低，因為，進行防禦後，沒攻擊的原本圖片也多了 0.1 的成功率，我想可能是我的成功地攻擊被防禦的數量比原本就沒成功的攻擊被變成功的還少。

圖片變得模糊，且較不菱角分明。