

Machine Learning HW5 Report

學號：B05902525 系級：資工三 姓名：葛洵為

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)

我所使用的 proxy model 為 resnet50，方法為 Iterative target model，epsilon 為 0.2，總共 5 個 step，每個 step 移動 0.01，比起 FGSM 來說，因為每次都小 size 的改動，所以攻擊成功機率變高，但是 L-norm 卻不容易增加，在 FGSM 只能達到 0.875 的成功率，但是使用此方法可以達到 0.995 的成功率。

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)。

FGSM：proxy model 為 resnet50，epsilon 為 0.05，success rate 為 0.875，L-inf norm 為 0.5

Best：proxy model 為 resnet50，epsilon 為 0.2，step 為 5，每次 step 移動 0.01，success rate 為 0.995，L-inf norm 為 0.3

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。

同為 epsilon 為 0.5，vgg16 acc=0.775，vgg19 acc=0.725，resnet50 acc=0.900，resnet101 acc=0.800，desnet121 acc=0.815，desnet169 acc=0.820，由此可見 model 比較接近於 resnet50

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)。



dung beetle = 0.8502661
ground beetle, carabid beetle = 0.12112842
leaf beetle, chrysomelid = 0.01896855

leaf beetle, chrysomelid = 0.9417743
cockroach, roach = 0.03709337
long-horned beetle, longicorn, longicorn
beetle = 0.01244056



Vase = 0.9235093
Screw = 0.04608277
Hummingbird = 0.00571818

monastery = 0.26623145
vault = 0.14099333
bell cote, bell cot = 0.08371202



bull mastiff = 0.933506

boxer = 0.02323634

Rhodesian ridgeback = 0.02308533

boxer = 0.338088

Tibetan mastiff = 0.04731672

Shih-Tzu = 0.13470995

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。