

HW6 Report

學號：b07901039 系級：電機二 姓名：劉知穎

1. (2%) 試說明 hw6_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)

使用Basic Iterative Method。遞疊更新adversarial image，每次更新 $\alpha \times \text{sign}(\text{grad})$ ，並clamp回與原本圖片相距 ε (L-inf norm)的距離。為了增加運算速度，若攻擊成功就停止更新圖片，最多更新五次。

$$\mathbf{X}_0^{adv} = \mathbf{X}, \quad \mathbf{X}_{N+1}^{adv} = \text{Clip}_{X,\varepsilon} \left\{ \mathbf{X}_N^{adv} + \alpha \text{sign}(\nabla_X J(\mathbf{X}_N^{adv}, y_{true})) \right\}$$

Proxy model: densenet121。

參數：

- $\varepsilon = \frac{20}{255} = 0.078$
- $\alpha = 0.1$

結果：

演算法	Success rate	L-inf norm
FGSM (epsilon=0.1)	0.905	5.550
Basic Iterative Method	1.0	4.625

使用Basic Iterative Method的success rate和L-inf norm均較FGSM好，因為FGSM只更新一次圖片，而Basic Iterative Method則可以更新多次，所以success rate較佳。在這次實驗中，使用basic iterative method，更新一次就成功的圖片有166張，更新兩次成功的有19張。另外，每次更新完將圖片clamp與到原圖相聚epsilon，讓最後平均的L-inf norm較小。

2. (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。

Densenet121。

因為實作的時候評分網站已經關閉，所以沒有online的數據。但這份報告會以densenet121為proxy model討論。

3. (1%) 請以 hw6_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)。



4. (2%) 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 5*5 的 Gaussian Filter 做被動防禦。實作方法：cv2.GaussianBlur(img, (5,5), 0)

	Success rate
Adversarial 防禦前	1.0
Adversarial 防禦後	0.73

	Test accuracy
Original	0.925
Original + smoothing	0.835

