Machine Learning HW5 Report

學號：B0502128　　系級：資工三　　姓名：鄭百凱

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)

我在 hw5_best 當中，使用的 proxy model 是 RESNET50，方法是 iterative FGSM，並且在本機做比對，若該照片在本機端並沒有被 attack 成功，就會反覆做 FGSM 直到 iterate 成功；反之，若一次就 attack 成功，便不再繼續做 FGSM，才能降低 L-inf. norm。此外，我將 epsilon 設為 0.015，這樣子每做一次 FGSM 每個 pixel 最多變化正負 1。

相較於調整 FGSM 的參數，這樣的方法能夠對每一張圖片作客製化次數的 FGSM，可以同時升高 success rate 和降低 L-inf. Norm。若是單純的 FGSM，在調整 epsilon 的時候，只會同時升高或同時降低。此外，普通的 FGSM，也有可能因為 epsilon 太大，修正過多，而錯過 minimum，所以也不見得 epsilon 越大，success rate 就會越好。像是我自己測試時，epsilon = 0.2 比 0.3 要來的好。這也是為甚麼要用 iterative FGSM，這樣才能確保能夠接近 minimum。

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)。

	success rate	L-inf. norm
hw5_fgsm	0.920	4.0000
hw5_best	1.000	1.3350

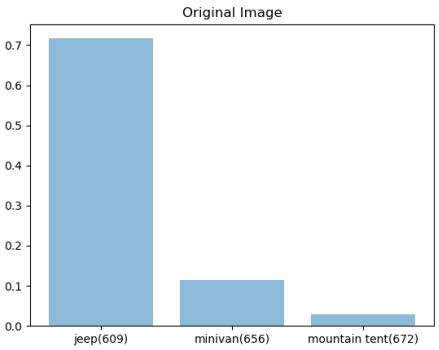
兩個 proxy model 都是 RESNET50。

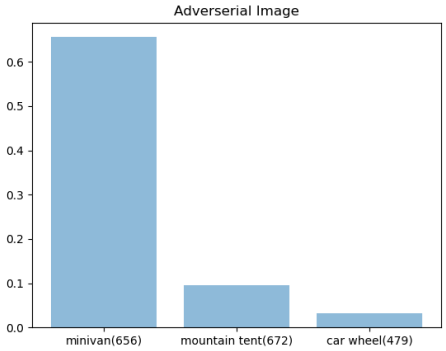
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

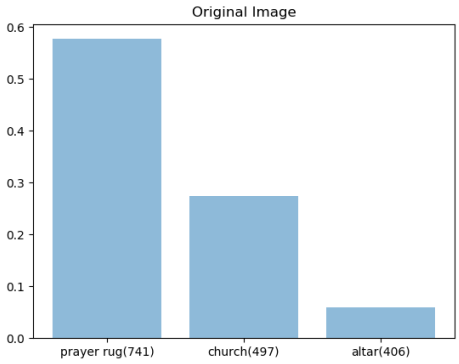
	success rate	L-inf. norm
VGG16	0.265	6.0000
VGG19	0.260	6.0000
RESNET50	0.855	6.0000
RESNET101	0.395	6.0000
DENSENET121	0.320	6.0000
DENSENET169	0.290	6.0000

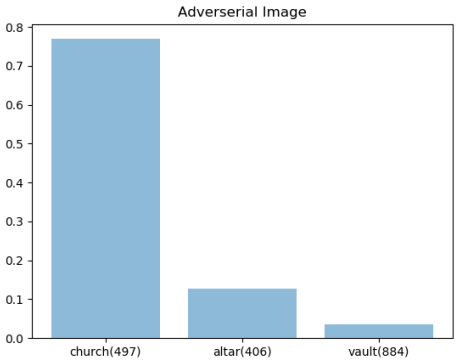
在使用 FGSM 的方法之下，固定 epsilon 的大小，改變不同的 proxy model，得到的結果如上圖。可見 RESNET50 為成功率最高的模型，因此推測 RESNET50 最後可能是背後的 black box 模型。

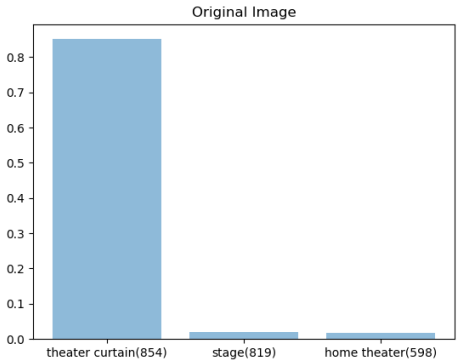
4. (1%) 請以 hw5_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

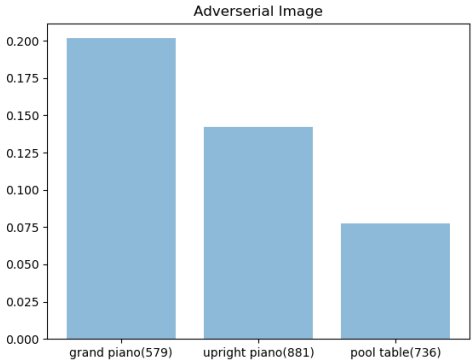












5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

我用的是高斯 filter，以下是在$\sigma = \sqrt{2}$的情況下做出來的結果，有一定程度的降低 model 誤判的比例，但並沒有到很有效。

防禦前攻擊 success rate：0.920

防禦後攻擊 success rate：0.745

此防禦對原始圖片的影響：有 87%的圖片還是能夠判斷出原本的 label，此外，圖片在物體交界處會變得比較模糊。

隨著σ 變大，model 對原始圖片的辨識能力會和攻擊成功的機率一起降低，若是要有效降低 model 誤判的比例，那麼 model 也比較容易判斷錯原始圖片。