學號：B06902067　系級：資工二 姓名：許育銘

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)

proxy model：resnet50，直接取用 pytorch model 的參數。
方法：類似 FGSM，但每次只改動 1*sign(gradient)，直到結果改變或改變次數過大(>100)。
根據觀察，有許多圖只要改動 1*sign(gradient)就可以騙過 model，因此可以在 L-inf 平均為 5 以內對部分圖做比較大的改動，藉此增加成功率。也可以說是動態調整 FGSM 中 epsilon 的大小。

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	hw5_best.sh
proxy model	resnet50(pretrained)	resnet50(pretrained)
success rate	0.925	0.965
L-inf. norm	5.0000	2.6200

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。

black box 應該是 resnet50
以下是我對各種 model 做 FGSM(epsilon=5)的結果

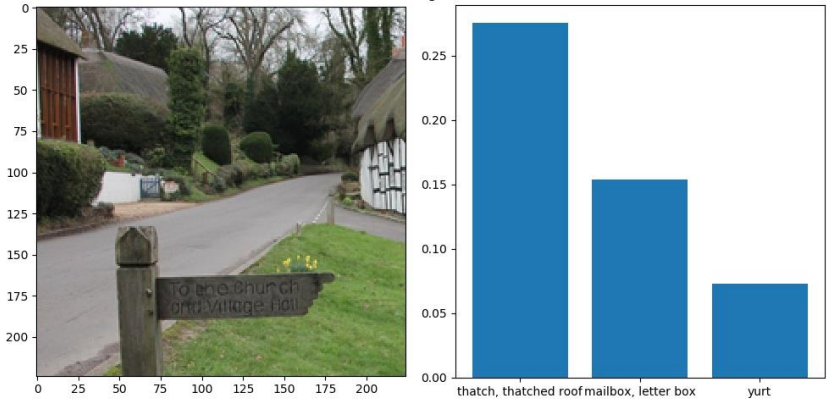
model	vgg16	vgg19	resnet50	resnet101	densenet121	densenet169
success rate	0.265	0.250	0.925	0.445	0.365	0.360

可看出對 resnet50 作攻擊成功率特別高。

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)。

上方為原圖，下方為攻擊過的圖。

Image 7



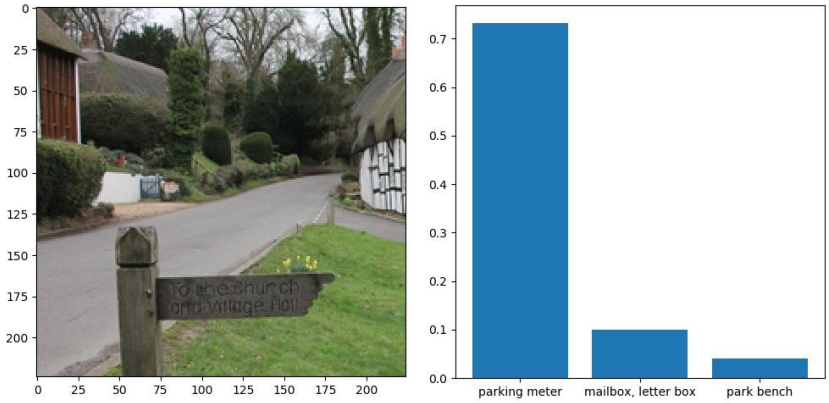
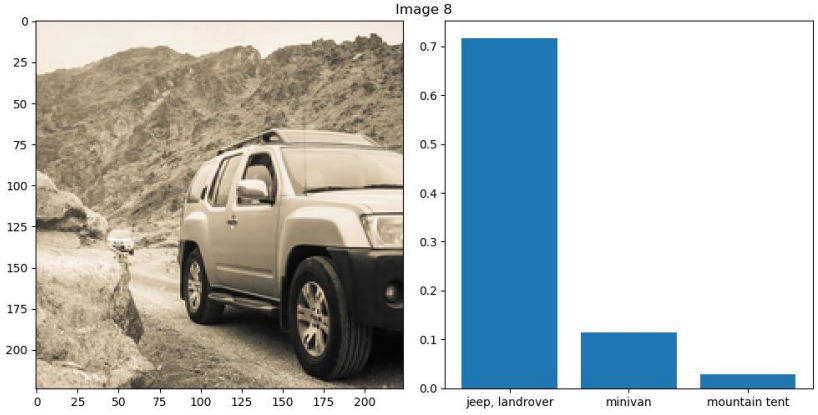
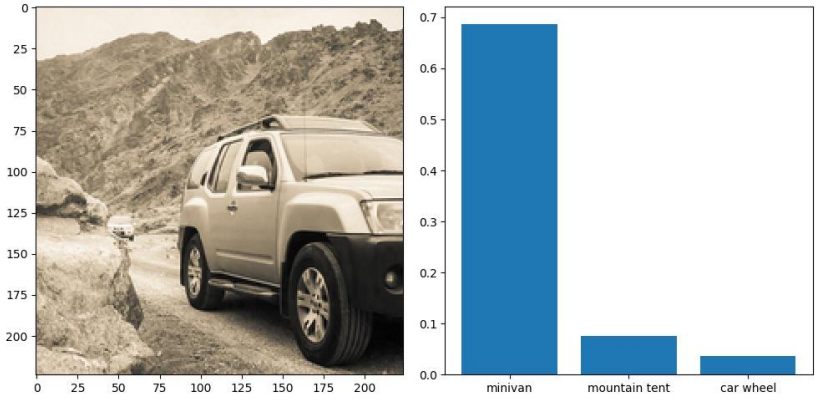
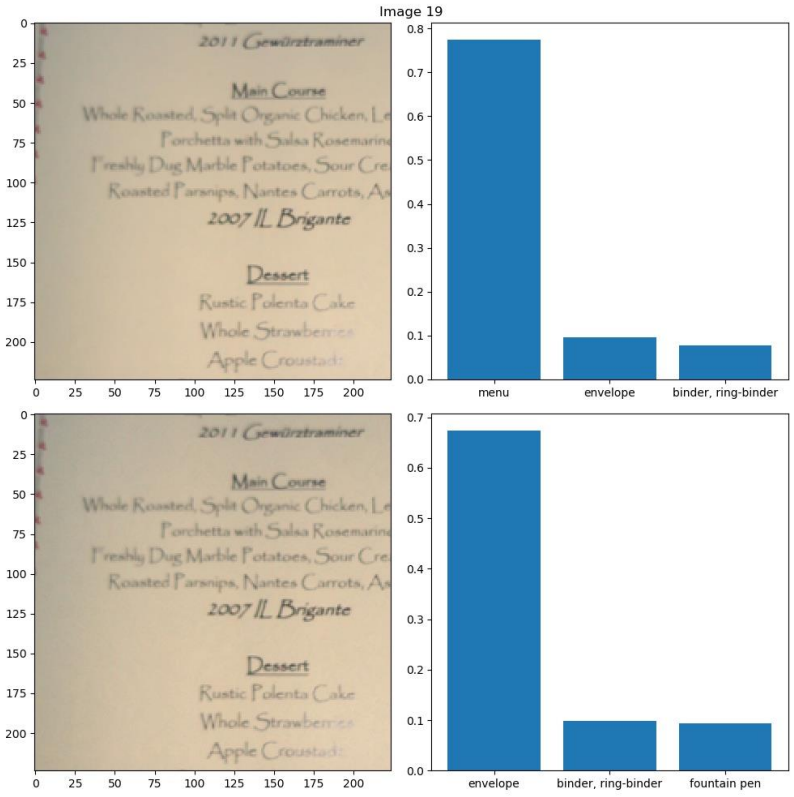


Image 8







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。

方法：blur，將每個 pixel 改成自己與周圍 8 個的平均(邊界則沒有改動)。

原圖做防禦：success rate=0.21，有 21%的圖片會誤判。

攻擊後的圖做防禦：success rate=0.41，有一半以上防住了。

此方法可以降低攻擊性，但離原圖還是有一段距離(誤判率約為 2 倍)，此外，此方法會造成原圖 21%的誤判率，算是有點缺陷。