Machine Learning HW6

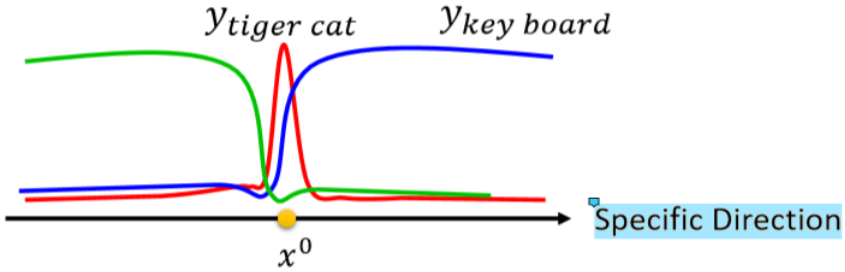
學號:B06902060　系級:資工三　姓名:鄒宗霖

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)

   在此題中我所使用的 proxy model 為 densenet121，而實作的方法如下，首先 epsilon 的值為 0.01，如果這個 epsilon 利用 FGSM (Fast Gradient Sign Method) 更新完參數的 perturbed image 可以成功使 proxy model 做出錯誤的分類，我們就把這個 perturbed image 存起來；反之，我們將 epsilon 的值增加 0.01，直到利用 FGSM 更新完參數的 perturbed image 可以成功使 proxy model 做出錯誤的分類才停止；若很不幸的在 epsilon 增加到 1.5 後還無法成功攻擊 proxy model，我們就把原本的 image 存起來，才不會增加到 L-inf. norm。在原本的 FGSM model 中，每張 perturbed image 的 epsilon 值都一樣，而此方法和 FGSM 的差異為我們給予每張 perturbed image 的 epsilon 都不一樣，而且此 epsilon 是可以成功攻擊 proxy model 的最小值，所以能最小化 L-inf. norm；因為在還沒有辦法成功攻擊 proxy model 時，我們會不斷將 epsilon 的值增加直到成功，所以能夠提高 success rate (success rate : 0.975 , L-inf. norm : 1.34)。

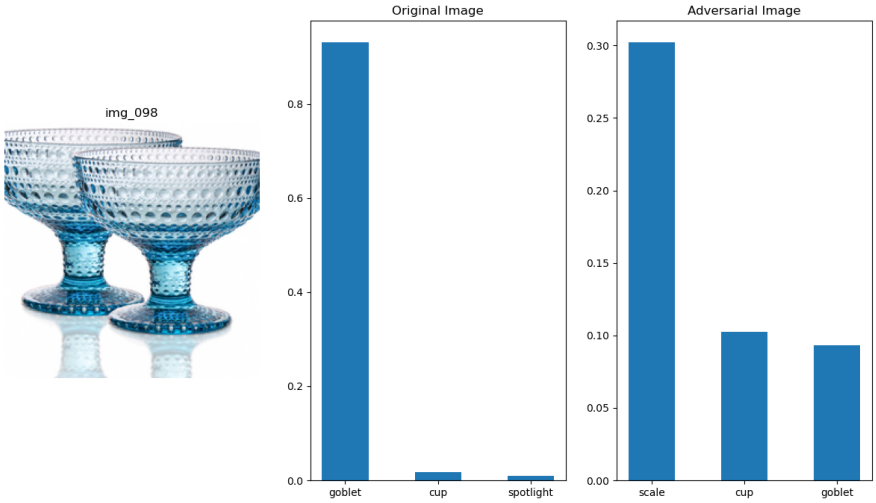
2. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

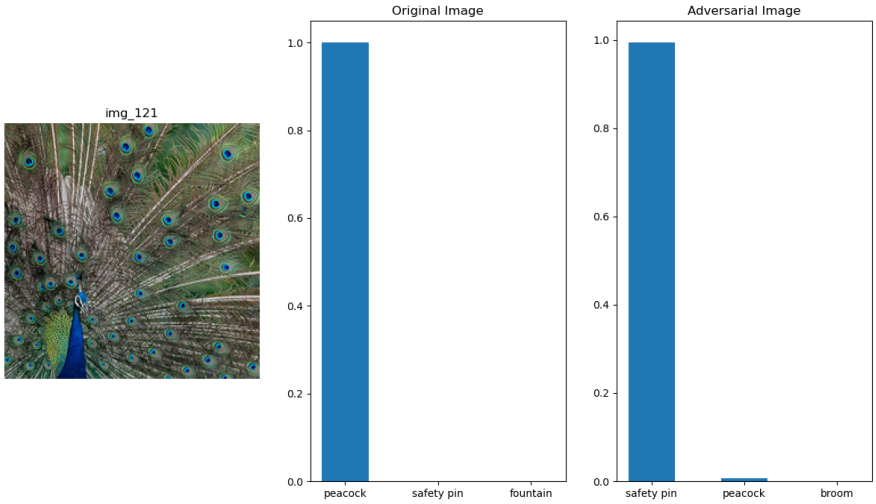
vgg16	vgg19	resnet50	resnet101	densenet121	densenet169
0.575	0.545	0.640	0.585	0.925	0.715

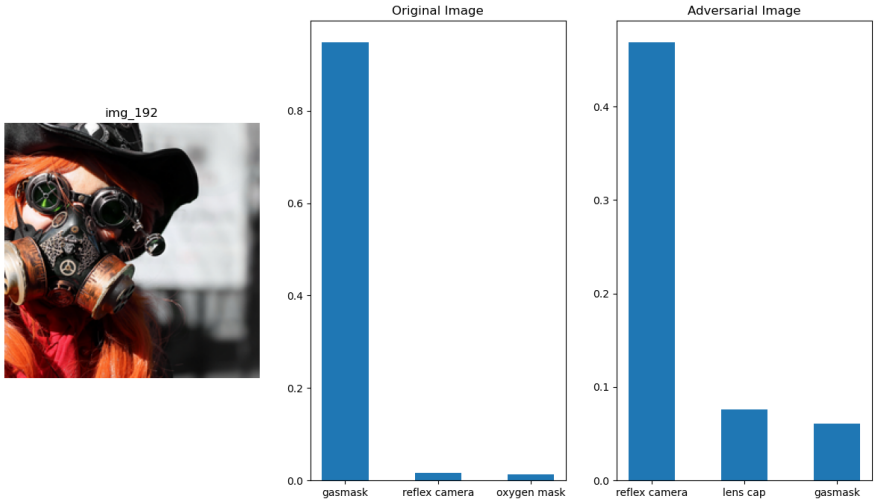


   上表為利用不同的 proxy model 在 epsilon = 0.3 時產生 perturbed image 去攻擊 black box 的成功率，可以看到 densenet121為被攻擊最徹底的 model，因此可以推測背後的 black box 最有可能為 densenet121。我們之所以能夠將圖片改變一點點就達到攻擊的成效，是因為在高維空間中特定的方向上 (Gradient 的方向) 正確答案為機率最高的範圍狹窄，如上圖老師的課程投影片所示，因為不同的 model 有著不同的結構，上述高維空間中特定的方向也會有所不同，然而 densenet121被攻擊的最徹底，代表它和 proxy model 有著同樣的結構，推測背後的 black box 就是它。

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)。







上面三張圖片為 proxy model 被攻擊前後的機率分布圖，我們可以看到高腳杯、孔雀、防毒面具在被攻擊前都有著接近 1 的機率，然而在被攻擊後的機率均掉到小於 0.1，但是都有維持在前三高的機率。

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。

	before smoothing	after smoothing
adv_img	0.975	0.325
ori_img	0.075	0.160

上表為 adversarial, original image smoothing 前後攻擊 black box 的 success rate，本題是利用 Gaussian filtering 實作 smoothing 被動防禦 (使用 cv2 中的 GaussianBlur 函式，filter 大小為 5 * 5)，我們可以看到adversarial image 在 smoothing 後攻擊 black box 的 success rate 從 0.975 降低到 0.325，有效的降低了被攻擊的比率。如第二題所提到，我們之所以能夠將圖片改變一點點就達到攻擊的成效，是因為在高維空間中特定的方向上 (Gradient 的方向) 正確答案為機率最高的範圍狹窄，然而 smoothing 破壞了 adversarial image 先前利用 FGSM 更新參數的方向，自然能達到被動防禦的效果。此外，雖然此方法能成功抵擋攻擊，但從上表我們也可以看到實作 smoothing 降低了原始圖片的辨識率，增加了模型誤判的比例，攻擊成功率從 0.075 增加到 0.160，然而增加的幅度不大，實作 smoothing 被動防禦可能利大於弊。