學號：B07902047　系級：資工二　　姓名：羅啟帆

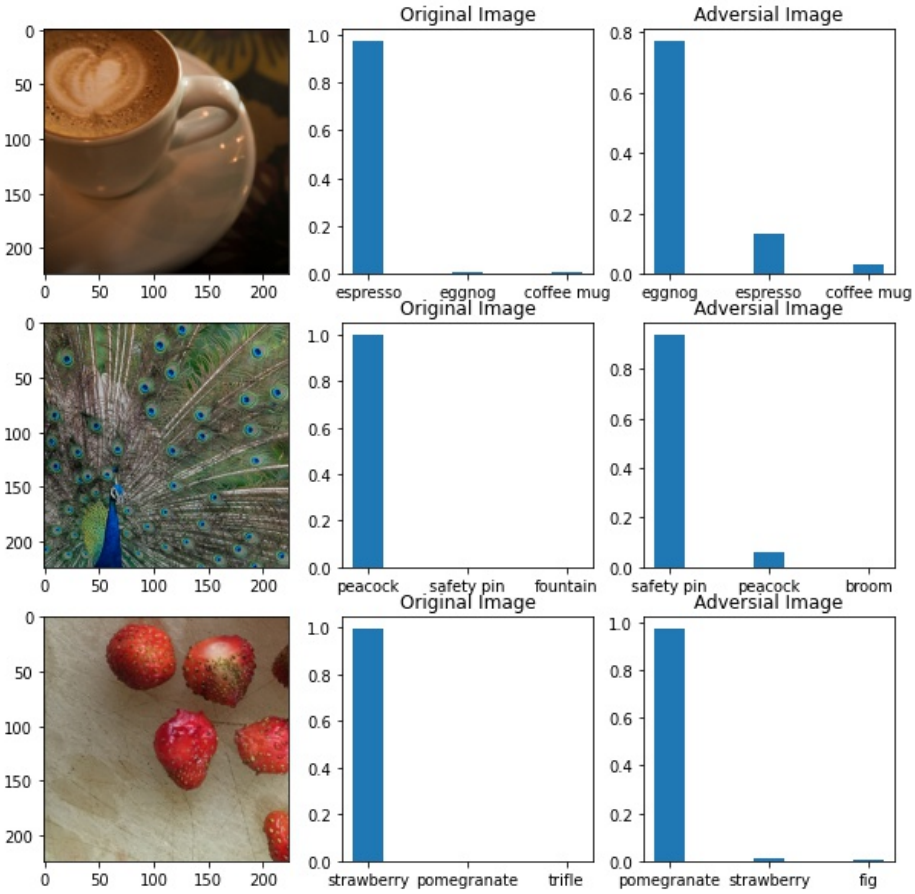
1. (2%) 試說明 hw6_best.sh 攻擊的方法，包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何？如何影響你的結果？請完整討論。(依內容完整度給分)
    - 我使用的 proxy model 是 densenet121，模型參數是 torchvision 的 pretrained 參數，而攻擊的方法是修改版的 Deepfool。大概的步驟長這樣：
        1. 將 raw picture 做預處理（縮放到 0~1，然後做 normalize）
        2. 如果 proxy model 在這張圖片上本來就答錯，直接跳過這張圖片
        3. 開始 DeepFool 演算法，最多迭代 50 次：（下面的符號街沿用論文內的符號）
            1. 將圖片跑過 model 得到預測，記錄機率前 3 高的 classes
            2. 如果目前圖片可以騙過 proxy model：
                1. 檢查轉換回 0~255 的圖片有沒有成功騙過 proxy model
                2. 有的話就回傳，沒有的話就對圖片做 $x_{i+1} = x_i + \frac{r_{i-1}}{max\, r_{i-1}} \times eps \times 0.2$的更新，最後再 clip 圖片
            3. 取出 $\hat{l} = \arg\min_{k \neq \hat{k}(x_0)} \frac{|f_k|}{\|w_k\|_1}$
            4. 對圖片做更新：$x_{i+1} = x_i + (1 + 0.001)\frac{|f_k|}{\|w_k\|_1} \times sign(w_i)$
            5. Clip 圖片使得 L-inf 不會超出一個預設的範圍（clip 的範圍我設定 1.00001/255/0.225，這樣換回 0~255 時 L-inf 會在 1 上下）
    - 與原 deepfool 不同的點在於
        1. deepfool 沒有 3-2-2 這個步驟
        2. 3-3 和 3-4 原本論文是用 2-norm
        3. 3-4 更新的時候原本論文不是用 $sign(w_i)$ 來更新。
    - 與 fgsm 的差異是主要在於
        1. fgsm 只根據原本類別計算更新方向，並且只更新一次；相較之下 deepfool 在計算更新方向時考慮了其他類別，計算的更新方向更加準確，而且更新次數多，步伐小，比 fgsm 精準許多。
        2. 總結來說，fgsm 計算速度快，但沒有 deepfool 精準；deepfool 可以在 L-inf 更小的情況下 fool model，但花費時間較久。
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。

| | Success rate | L-inf |
|---|---|---|
| Densenet121 | 100% | 0.935 |
| Densenet169 | 13.5% | 0.925 |
| Resnet50 | 10% | 1.01 |
| Resnet101 | 11.5% | 0.94 |
| Vgg16 | 8.5% | 0.87 |
| Vgg19 | 9% | 0.875 |

以上數據是用 hw6_best 得到的。由以上的數據推測 black box model 最有可能是 densenet121。

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)。

我取編號 6, 77, 121 這三張圖片來畫圖。




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。
   - Smoothing method：opencv GaussianBlur, kernel_size=(3, 3) with default sigma
   - 防禦前後的攻擊 success rate：
     1. 防禦前：100%
     2. 防禦後：69%
   - 可以發現 GaussianBlur 可以降低攻擊的效果。
   - 防禦對原本圖片 accuracy 的影響：
     1. 防禦前：92.5%
     2. 防禦後：81%
   - 可以發現此防禦方法雖可以有效降低攻擊，卻也同時降低了原本的準確率。