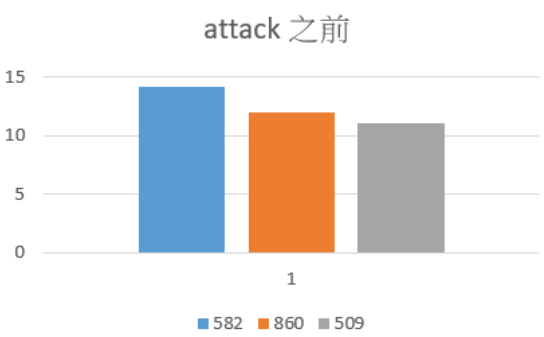
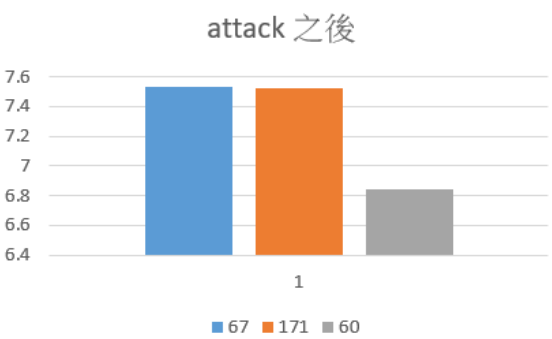
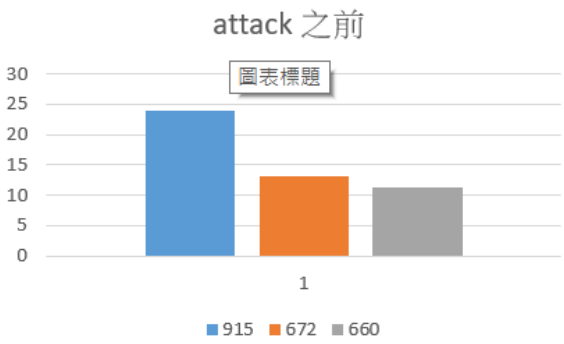
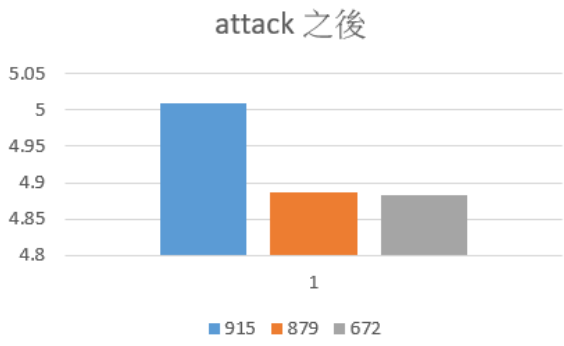
學號：b07901112　　系級：電機二　　姓名：劉聿珉

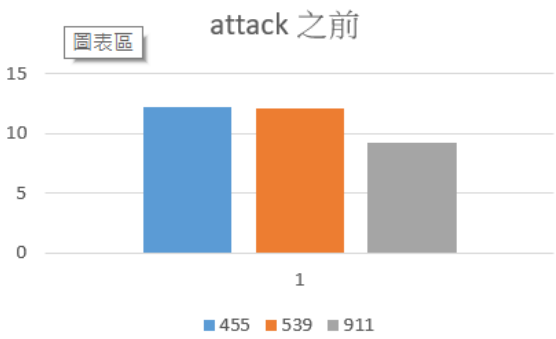
1. (2%) 試說明 hw6_best.sh 攻擊的方法，包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何？如何影響你的結果？請完整討論。(依內容完整度給分)
   我在這次的作業的 best 就是用 FGSM 的方式做出來的，proxy model 為 DenseNet-121，epsilons = 0.1，其他參數都跟助教給的 code 一模一樣，沒有改過。(這是我能想出最完整的描述方法了)
   執行 FGSM 的邏輯以及方法：
   FGSM 的想法其實非常簡單，在執行之前會有一個 original image x0
   我們會希望在 x0 上加上一個擾動，使得 model 的 loss 增加最多，並且要在一個人眼無法察覺的限制內。而顯而易見的，擾動發生在梯度方向會是最有效的，而 FGSM 中的一個合理的假設就在於，他假設現有的影象資料的儲存方式都是離散的，因此對一個在同一方向的 ε 擾動是不會被人眼發現(當然 ε 太大時也會被發現，所以 ε 要盡可能小)，因此，FGSM 會將 x0 在的梯度方向上平移 ε 得到擾動 η ，其中需要經過 sign function。
   FGSM 的概念圖像畫就像是在 x0 周圍畫上一個半徑為 ε 的圓或是邊長為 2* ε 方形，再由 gradient 的方向決定要移動的方向，然後將 x0 走到邊上，這就是 FGSM 的執行邏輯以及時作方法。
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。
   這次叫我們預測的模型有六種，他們的名字以及成功率如下：
   VGG-16：0.310
   VGG-19：0.310
   ResNet-50：0.400
   ResNet-101：0.360
   DenseNet-121：0.920
   DenseNet-169：0.440
   以上的數據 L-Inifinity 皆為 6.000
   由以上結果可以知道，black box 所使用的 model 是 DenseNet-121，因為他的正確率明顯比其他模型好太多了。
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)。

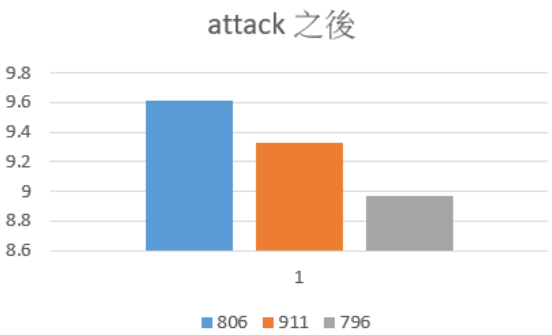

attack 之前
15
10
5
0
1
582
860
509


attack 之後
7.6
7.4
7.2
7
6.8
6.6
6.4
1
67
171
60


attack 之前
圖表標題
30
25
20
15
10
5
0
1
915
672
660


attack 之後
5.05
5
4.95
4.9
4.85
4.8
1
915
879
672





上面三張圖分別是 009,010,011 號圖，從圖表中可以看到其實 attack 是成功的，009,011 號圖的第一高機率類別都有變不一樣，就算是 010 他判定的機率也大幅度降低，而且跟第二以及第三相差不多。

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。

   這題我是用 PIL image filter 對圖像做高斯模糊，模糊之後發現對於 attack 之前的圖像，高斯模糊後機器識別正確率是降低的，從原本識別正確率 0.925 變成 0.730，而對於 attack 之後的圖片做高斯模糊會讓正確率從 0.08 上升到 0.355，雖然識別正確率都不高但是明顯看出有效果。這樣的結果是符合預期的，因為高斯分布會將圖片變的模糊，因此會讓 attack 的效果便得不明顯，但相對的他也會使機器叫難辨識原本的圖片，而用肉眼看出圖片，attack 之前及之後的照片其實都看不太出有差異，都是糊糊的一塊。