學號：b07901014　系級：電機二　　姓名：陳希格

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)

proxy model : densenet121

實作方式: FGSM, 但是 iteration 五次，每次的 epsilons = 0.01628 (第一次 = 0.017，不一樣只是為了方便判斷是不是第一次而已)

iteraion times	Test Accuracy
1	0.23
2	0.04
3	0.15
4	0.0
5	0.0

可以明顯看到 iteration 此的幫助很大，結果大幅改善。且因為只針對沒有正確的圖片進行 FGSM 攻擊 ，所以總體的 L-Inifinity = 1.4700 並不高。

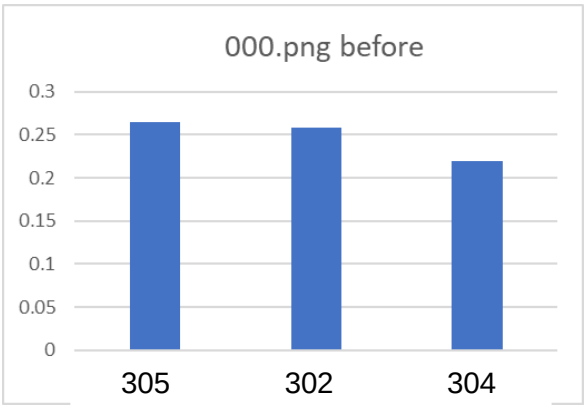
2. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

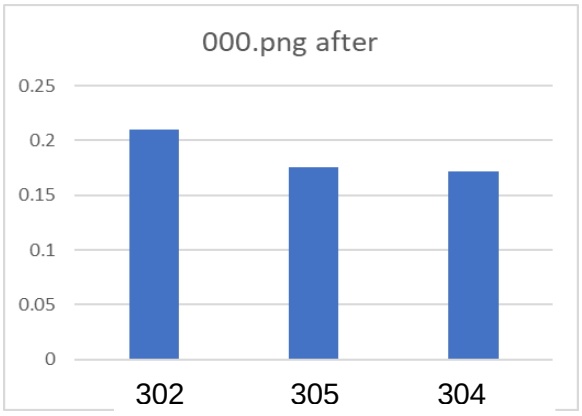
由於最後一天才開始的原因，只能上傳 5 次，且兩次因為沒有 clamp 而炸掉。所以只有以下幾種的比較。(先踢除了 vgg 因為 colab 的範例是 vgg，所以 blackbox 應該不是 vgg。)

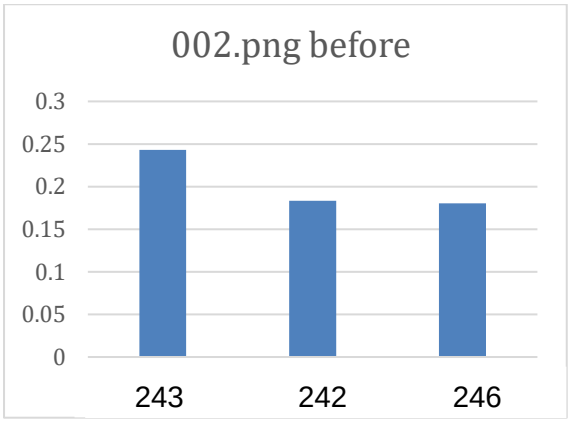
proxy model	iteraion times	epsilons	success rate on my computer	success rate online
densenet121	5	0.16(per iter)	1.00	1.00
resnet101	5	0.16(per iter)	1.00	0.10
Resnet50	1	0.475	0.895	0.345

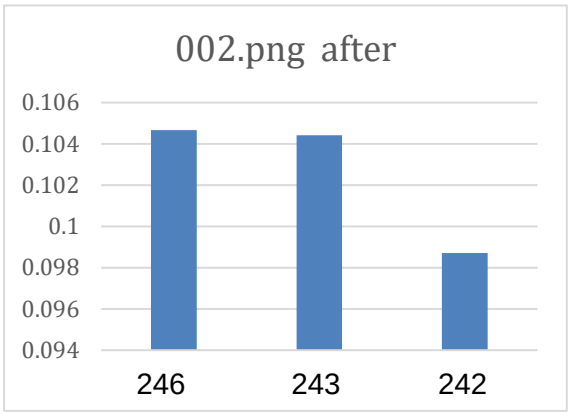
從現有結果來看，blackbox 應該是 densenet121，排除其他的原因是因為他們的 success rate on my computer 是好的，但 success rate online 不佳，代表 model 不一樣。

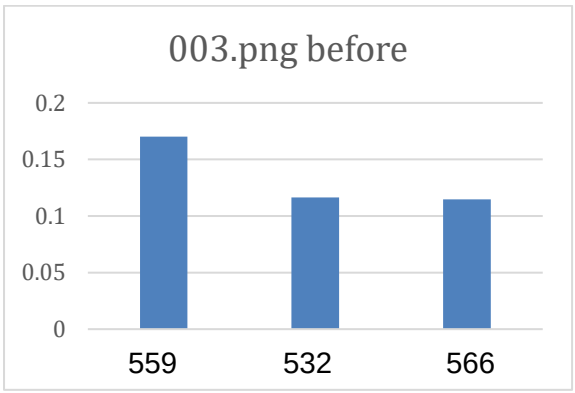
3. (1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。

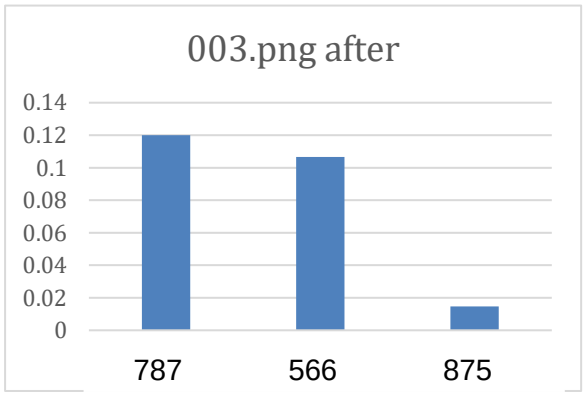












從前面兩組例子中可以發現其實前三名沒有變，只是順序變了而已。

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 方式 : GaussianBlur
success rate = 1- Accuracy

Accuracy	none	GaussianBlur
origin img	0.925	0.925
adversarial img	0.0	0.0

從結果可以發現高斯模糊完全沒有用，且對原始圖片的正確率也沒有影響。雖然這應該 smoothing 方式選擇不佳有關，但也從此可知要防禦其實是比較困難的。