Machine Learning HW5 Report

學號： B06902052系級：資工二 姓名：張集貴

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: ResNet-50，使用的方法是 https://arxiv.org/pdf/1611.01236.pdf 中的 target class method 的 iterative 版本。

   令 X 為原始輸入資料，y_{true} 為資料的正確分類，y_{target} 為目標 class（我實做時都選第 0 個 class），$J(X, y)$ 為 cost function，第 i 次更新後的結果為 X_i，則 $X_0 = X,\ X_{i+1} = X_i - \alpha \cdot sign(\nabla_X J(X_i, y_{target}))$。

   首先對 X 進行 standardization，進行 170 次更新，$\alpha = 0.000415$。

   此方法和 FGSM 的差異是，FGSM 是讓 y_{true} 的機率盡量降低，而此方法是讓 y_{target} 的機率盡量提高。

   看第 4. 題的圖可發現，我攻擊後 model 預測的結果都是一樣的 class（y_{target}），可見此方法確實有效。

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)。

	proxy model	success rate	L-inf. norm
hw5_fgsm.sh	ResNet-50	0.925	5.0000
hw5_best.sh	ResNet-50	1.000	4.9150

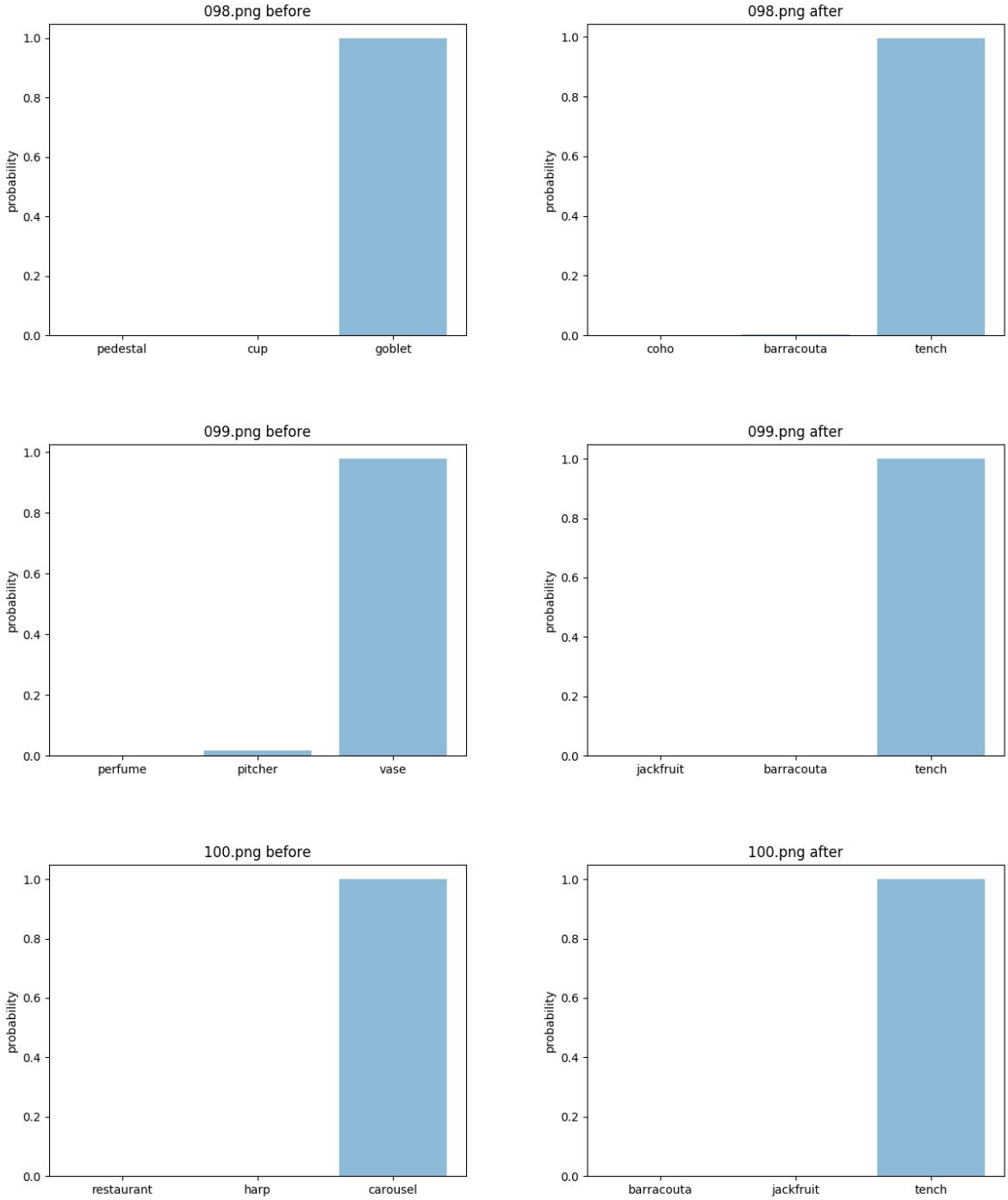
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box

最有可能為哪一個模型？請說明你的觀察和理由。

proxy model	success rate	L-inf. norm
VGG-16	0.035	4.9300
VGG-19	0.040	4.9200
ResNet-50	1.000	4.9150
ResNet-101	0.075	4.9250
DenseNet-121	0.060	4.9400
DenseNet-169	0.045	4.9200

ResNet-50 的 success rate 遠高於其他五個模型，所以背後的 black box 最有可能為 ResNet-50。

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)。



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。

對 hw5_best.sh 生成的圖片使用 Gaussian filtering。

	success rate	L-inf. norm
hw5_best.sh	1.000	4.9150
hw5_best.sh -> Gaussian filtering	0.195	108.9050
原始圖片->Gaussian filtering	0.155	108.4900

觀察可發現 Gaussian filtering 有效降低模型的誤判。

肉眼可看出此防禦產生的圖片明顯變模糊了，L-inf. norm 也因此變很高。

對原始圖片做 Gaussian filtering 的話則會產生一部分的誤判。