

Machine Learning HW5 Report

學號: B05901170

系級: 電機三

姓名: 陳柏志

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)

在 hw5_best.sh 中我使用 resnet50 作為 proxy model。並以多個 epoch、Gradient Assend 的方式更新參數。其中使用 adagrad 做為 optimizer, $\epsilon = 1$, Epoch = 100。另外每張圖片有客製化的 L-infinity 限制，一開始每張圖片的限制都是 1，若經過 100 個 epoch 後仍無法成功攻擊，則將 L-infinity 限制增加 1，反之若成功攻擊，則停止更新。反覆至所有圖片皆成功攻擊為止。另外 loss function 為 prediction 與正確答案的 label 的 CrossEntropyLoss 減去 prediction 與機率第二大的 label 的 CrossEntropyLoss，用意是除了讓圖片與正確答案愈來愈遠外，也會往原本預測結果中第 2 高的答案前進。

此方法和 FGSM 的主要差異在更新參數的方式，這個方法是依 gradient 大小來更新參數，並非像 FGSM 一樣只看 gradient 的 sign。另外也新增了客製化的 L-infinity。結論是在此方法下可以能讓更多的圖片攻擊成功，並同時能省下不必要的 L-infinity 浪費。

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	resnet50	0.935	13.6550
hw5_best.sh	resnet50	1.000	1.0050


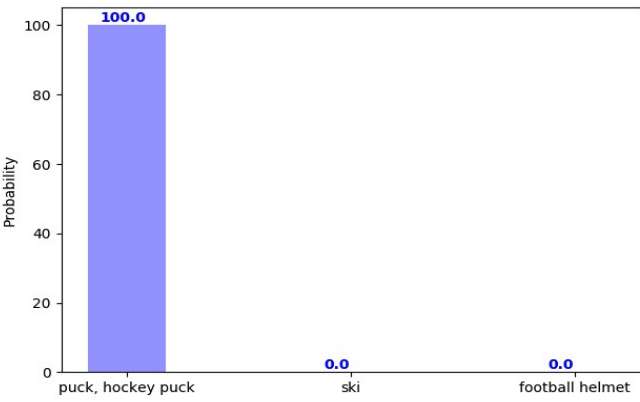
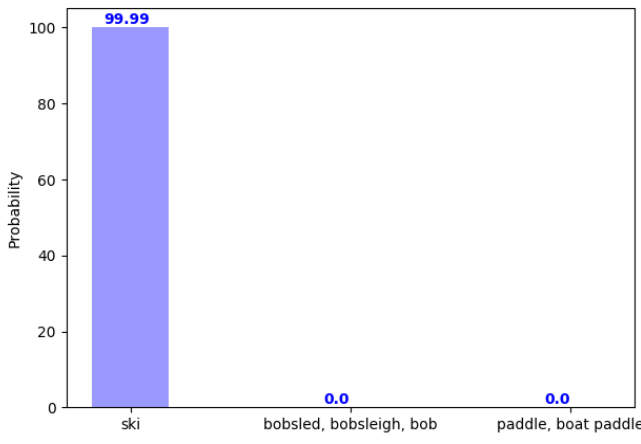
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

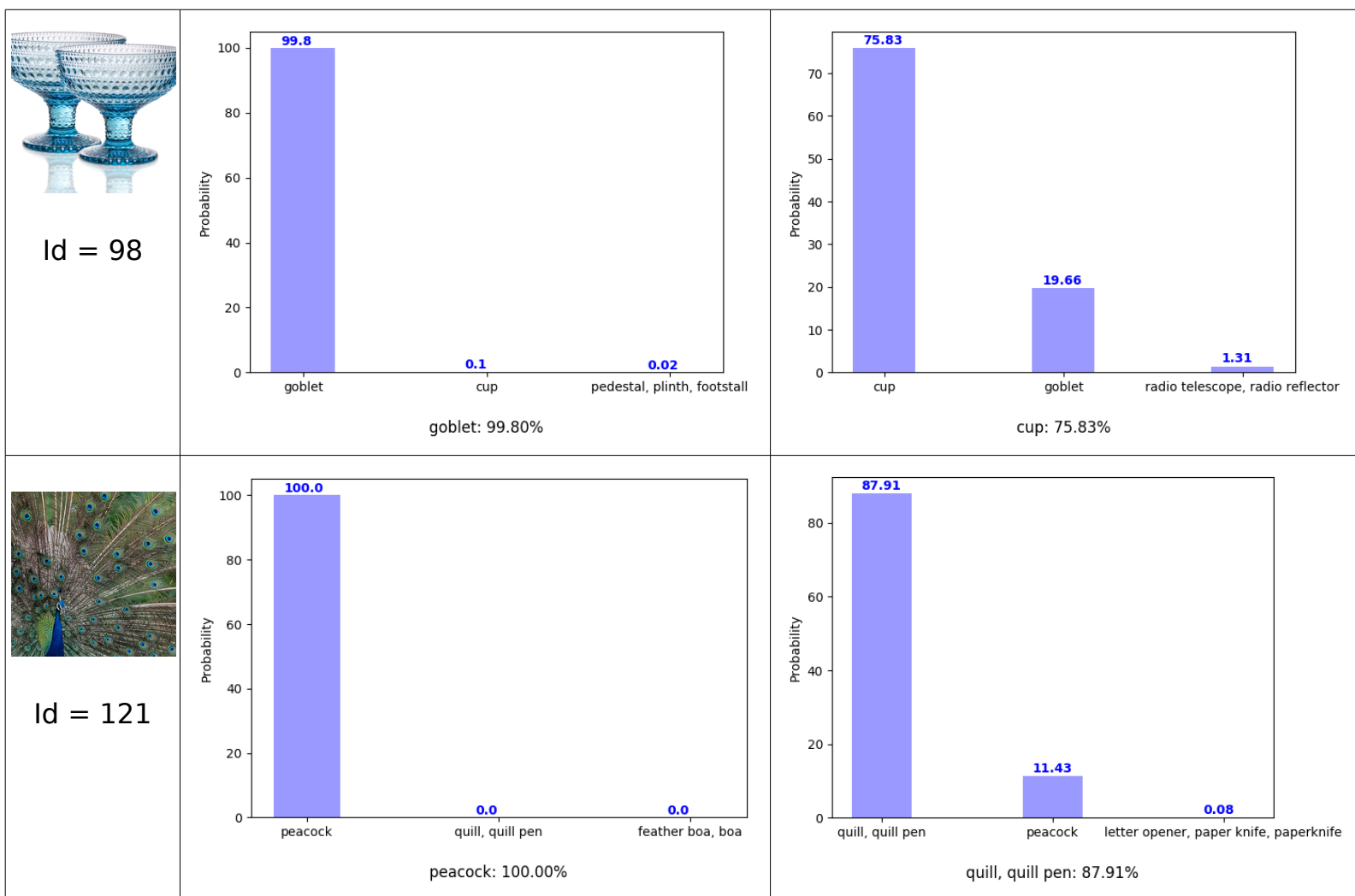
下表為實作結果：（implimented with FGSM, epsilon = 0.097）

Proxy Model	Success Rate	L-inf. norm
VGG16	0.375	10.9900
VGG19	0.355	10.8850
ResNet-50	0.935	10.7950
ResNet-101	0.510	10.8950
DenseNet-121	0.420	10.9450
DenseNet-169	0.435	10.9200

依以上實驗結果推斷，其背後的 black box 最有可能為 ResNet-50，因在 L-inf. norm 差距不大的情況下，其成功率明顯高出其他 proxy model。


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)。

Image	Original Probability Distribution	Adversarial Probability Distribution
 <p>Id = 50</p>	 <p>puck, hockey puck: 100.00%</p>	 <p>ski: 99.99%</p>



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。

使用 Pillow.ImageFilter 中的 Blur Filter 實作被動防禦。在讀進預測圖片前先過一次 Blur Filter, 再進行預測。下表為實驗結果：

	Success Rate	圖片比較 (image id = 0)
Without Passive Defense	1.0000	
With Passive Defense	0.3300	

可以看出，此 filter 會讓圖片變得較模糊，而同時能大幅降低攻擊的成功率。