學號：R08922167 系級：資工碩一 姓名：曾民君

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)

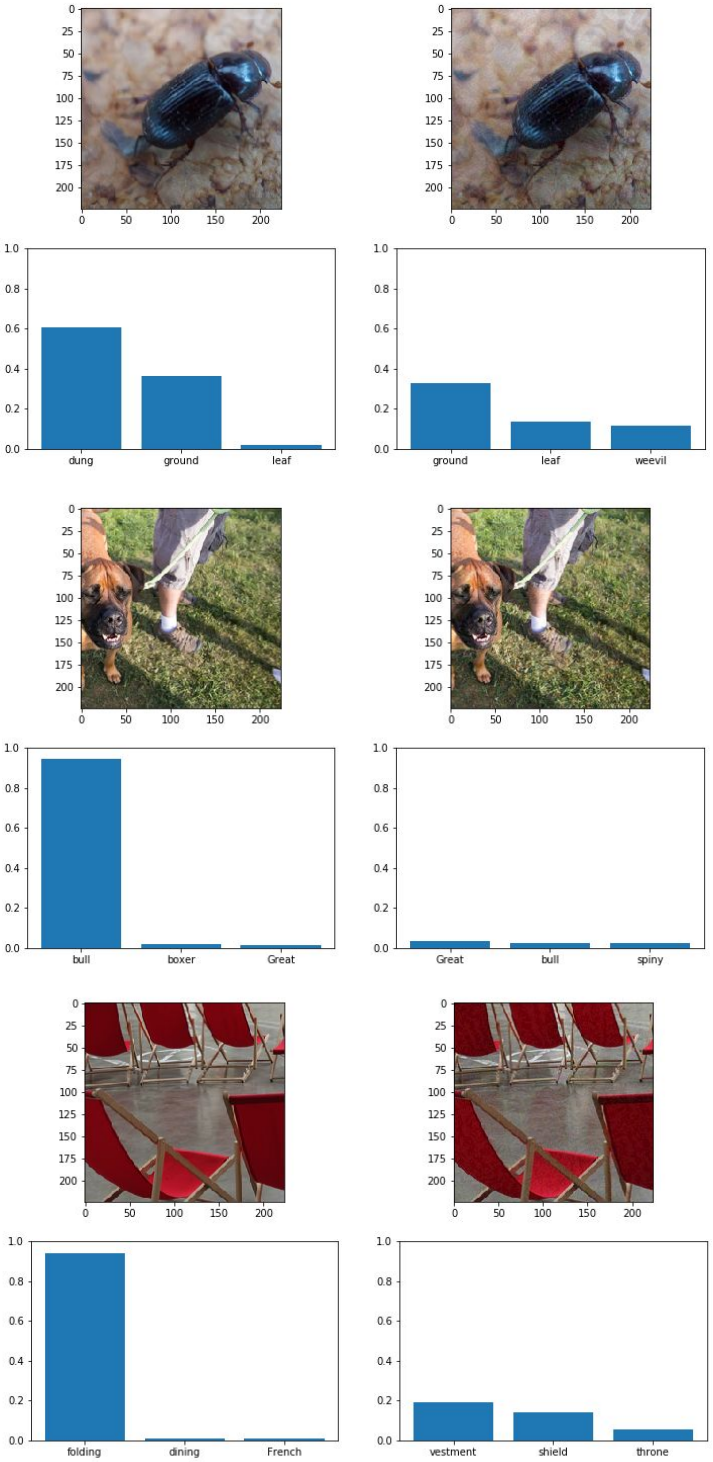
   Ans: 經過多次上傳攻擊不同 model 的結果，得出最有可能的 black box model 是 resnet-121，但首次的 FGSM 攻擊只有 91.5% 成功攻擊，所以改進方法為，將單次 FGSM 的失敗攻擊的結果進行第二次的 FGSM 攻擊，至於 epsilon 部份設為0.095 攻擊結果為 100%，avg infinite norm 為 6。 想法就是攻擊一次不夠就在攻擊一次，但問題會是攻擊兩次的影像可能會在 infinite norm 表現很遭。

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。

   Ans: 最有可能的 black box 為 resnet-121，原因是將所有有可能的proxy model 使用 FGSM 成功攻擊率只有 resnet-121 超過一半，所以判定為 resnet-121。

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)。

Ans:



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。

Ans: 這邊實作的被動防禦是使用 median filter，kernel 大小為 5*5。

比較對於只丟正常影像的結果，如果做防禦的話準確率會下降約 17%，但對於只丟備攻擊的影像，有做防禦的話準確率會上升約 32%，若只單看數字的話有做防禦平均的準確率會叫高。

Model: densenet-121	Success	Wrong	Accuracy
Ori-img + No-defense	185	15	0.925
Adv-img + No-defense	0	200	0
Ori-img + Defense	151	49	0.755
Adv-img + Defense	64	136	0.32