

Machine Learning HW5 Report

學號：b04507025 系級：電機四 姓名：韓秉勳

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)

事實上只要能選對 proxy model 用 FGSM 攻擊已能達到最好效果。我先把投影片上列出的 model 都試過以找出最佳解，發現 resnet50 最接近此作業的模型。其中要注意在 normalize 時要用 pytorch 規定的 normalize 方式喂入 model 才不會出錯。在訓練方面，我利用原本的標準答案 gradient 取 sign，並利用此 gradient 反向改變 input。如此一來，原本的圖片就會朝原標準答案的反方向走。我的一張圖片會走 70 個 step，每個 step 都會擾動 $0.01 * \text{sign}(\text{grad})$ 的量。另外為了減少 L-infinity，每個擾動都會縮在 $5e-4$ 的範圍內，且擾動完我也會將輸出 clip 在原本圖片像素大小內以避免超過標準。最後照片也要 denormalize 回來以維持同樣的照片輸出。原本過 simple 時並沒有 fit 與走那麼多 step，這也導致了差別。

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)。

由於我的方法都是 fgsm，故我直接取 iteration 較少次的做為比較。可看出當我做比較少 iterate 時準確率會降低，但 L-inf 也會降低，這可以代表擾動的實際影響。

	hw5_fgsm.sh	hw5_best.sh
proxy model	resnet50	resnet50
success rate	0.905	0.995
steps	15	70
L-inf norm	1.0000	3.0000


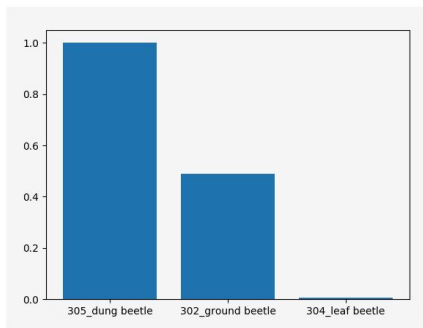
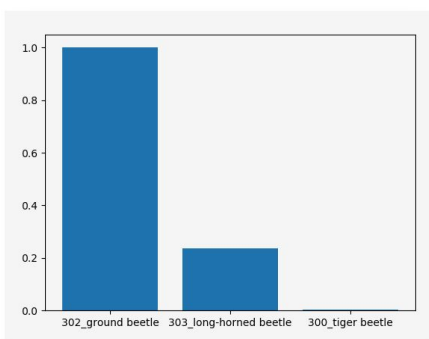

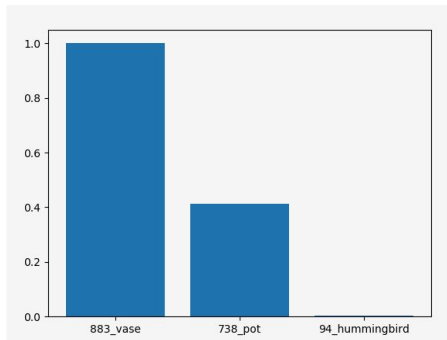
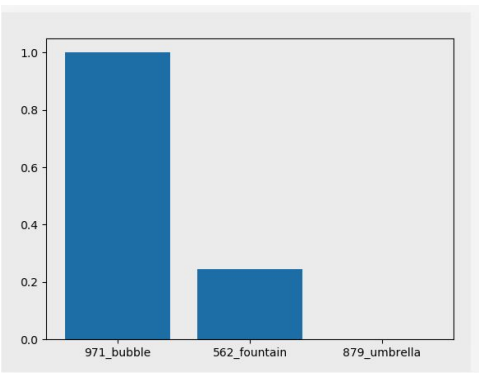
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

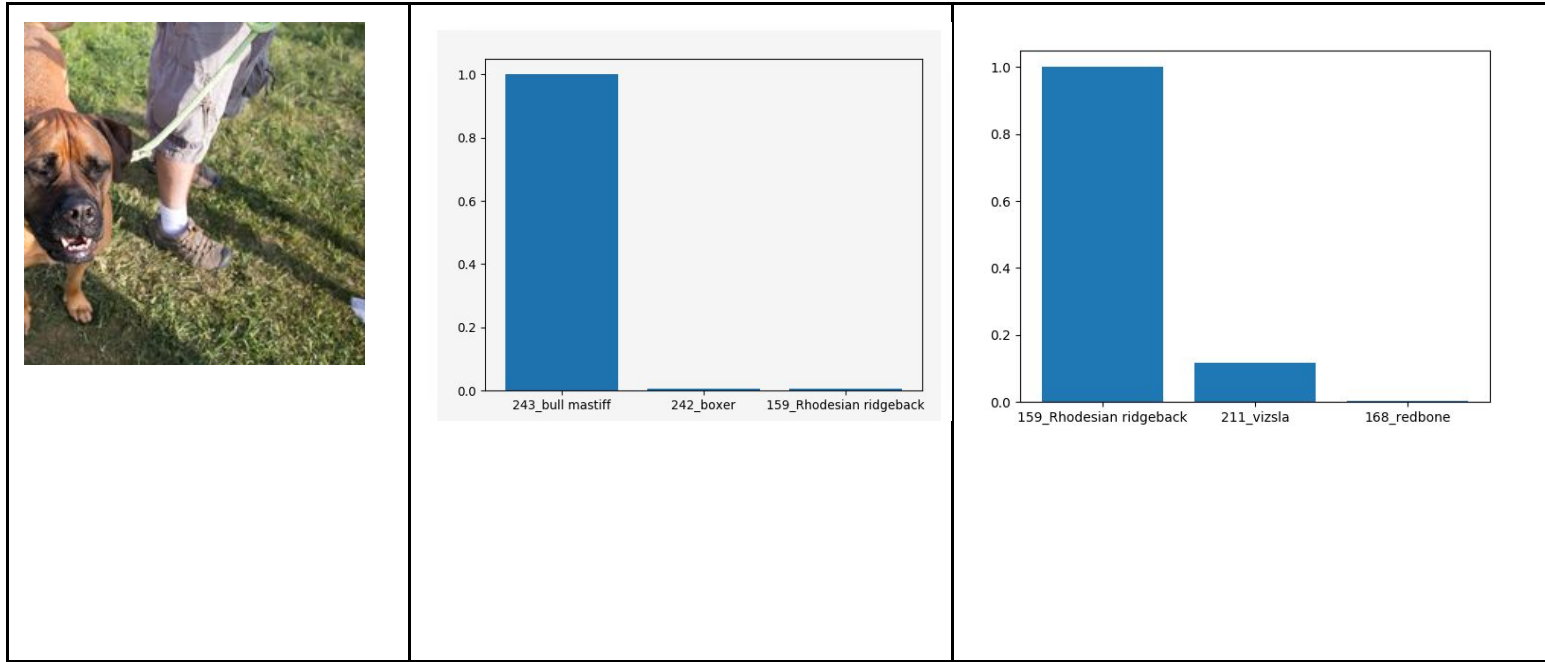
最有可能的是 resnet50，因為當我用同樣的方法對所有六種 model 做攻擊，只有 resnet50 達到最好的效果。其實我也有把所有可能 model 對原本圖片做 predict，得出來的結果也只有 resnet50 是完全吻合的，如此一來就更能確定該 model 是這次的目

標。

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)。

可以看到這三張圖的attack都有改變model預測結果，但除了第二張圖vase讀成bottle外，其他都是同類東西（如第一章圖是讀成另一種蟲，最後一張也是另一狗），應該是為了fit L-infinity而不能擾動太多，導致辨識結果沒有太多不同導致。

	before attack	after attack																
	 <table><tr><th>Category</th><th>Probability</th></tr><tr><td>305_dung beetle</td><td>1.0</td></tr><tr><td>302_ground beetle</td><td>0.5</td></tr><tr><td>304_leaf beetle</td><td>0.0</td></tr></table>	Category	Probability	305_dung beetle	1.0	302_ground beetle	0.5	304_leaf beetle	0.0	 <table><tr><th>Category</th><th>Probability</th></tr><tr><td>302_ground beetle</td><td>1.0</td></tr><tr><td>303_long-horned beetle</td><td>0.25</td></tr><tr><td>300_tiger beetle</td><td>0.0</td></tr></table>	Category	Probability	302_ground beetle	1.0	303_long-horned beetle	0.25	300_tiger beetle	0.0
Category	Probability																	
305_dung beetle	1.0																	
302_ground beetle	0.5																	
304_leaf beetle	0.0																	
Category	Probability																	
302_ground beetle	1.0																	
303_long-horned beetle	0.25																	
300_tiger beetle	0.0																	
	 <table><tr><th>Category</th><th>Probability</th></tr><tr><td>883_vase</td><td>1.0</td></tr><tr><td>738_pot</td><td>0.4</td></tr><tr><td>94_hummingbird</td><td>0.0</td></tr></table>	Category	Probability	883_vase	1.0	738_pot	0.4	94_hummingbird	0.0	 <table><tr><th>Category</th><th>Probability</th></tr><tr><td>971_bubble</td><td>1.0</td></tr><tr><td>562_fountain</td><td>0.25</td></tr><tr><td>879_umbrella</td><td>0.0</td></tr></table>	Category	Probability	971_bubble	1.0	562_fountain	0.25	879_umbrella	0.0
Category	Probability																	
883_vase	1.0																	
738_pot	0.4																	
94_hummingbird	0.0																	
Category	Probability																	
971_bubble	1.0																	
562_fountain	0.25																	
879_umbrella	0.0																	

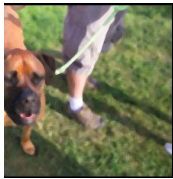


5. (1%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。

我實做了兩種filter, guassian filter 比較將edge模糊化，而median filter 會將圖片色塊化。然而可能因本來Model辨識效果很強，加上filter本身基本上就很容易辨識錯誤(L-inf 皆在100以上)，自然在攻擊後success rate也很高，故嚴格來講並沒有達到防禦的目標，即便我試了不同filter參數差異也不大，可能只有sigma=3(guassian std=3)時有稍微達到防禦功效，但原本影像也是會有0.16辨識錯誤。median filter 當kernal size=3時圖片也還算是判斷正確，但在攻擊後仍然會被影響。

guassian n std	origin	gaussian filter added	gaussian attacked	success rate before filtering	success rate with filtered image	success rate after attack
sigma = 1				0.000	0.16	1.000

sigma=3				0.000	0.16	0.995
sigma=5				0.000	0.865	0.995

median filter size	origin	median filter added	median attacked	success rate before filtering	success rate with filtered image	success rate after attack
size = 3				0.000	0.100	1.000
size = 5				0.000	0.345	1.000