Machine Learning - Homework 6

資工四 B05902023 李澤諺

April 30, 2020

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 的差異爲何？如何影響你的結果？請完整討論。(依內容完整度給分)

我使用的 proxy model 爲 DenseNet-121，並且使用了 basic iterative method (BIM) 對其進行攻擊，BIM 爲 FGSM 的改良，FGSM 使用了較大的 ϵ 並只對照片進行了一次的更新，爲了達到較高的 success rate，FGSM 必須使用較大的 ϵ，然而其會使得 L-inf norm 上升，而 BIM 則是使用較小的 ϵ 並對照片進行了多次的更新，如此一來便可以在成功攻擊一張照片的同時盡可能降低 L-inf norm。在 BIM 的實作上，我使用的 ϵ 爲 0.0135，並且對於每一張照片，我會對其不斷更新直到其無法被 proxy model 辨認便停止，故每張照片所需的更新次數不同，以在成功攻擊的條件下減少照片的更新次數，意即在 success rate 爲 100% 的條件下儘可能降低 L-inf norm。最終，我所得到的 success rate 爲 100%，而 L-inf norm 爲 1.055。

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 最有可能爲哪一個模型？請說明你的觀察和理由。

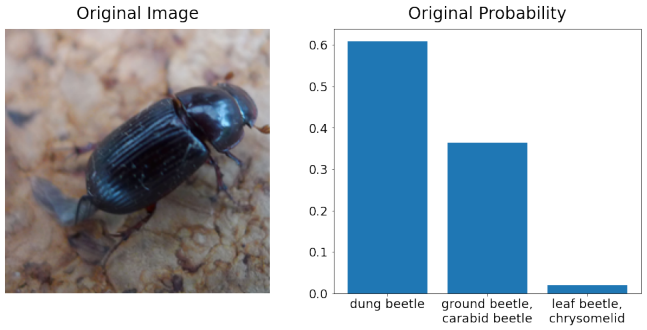
我將本次作業中所有可能的 black box model 作爲 proxy model，並使用 FGSM 對其進行攻擊，其中 ϵ 爲 0.01，所得到的 success rate 如下表所示：

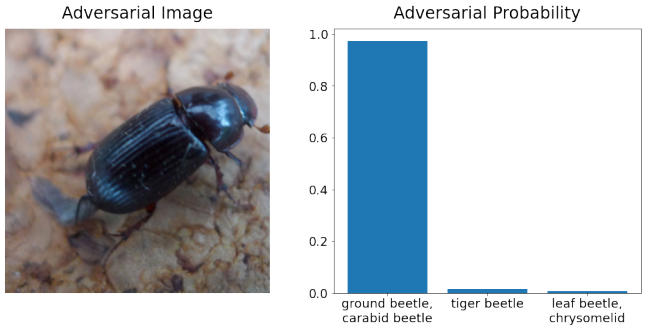
Proxy Model	Success Rate
VGG-16	0.100
VGG-19	0.125
ResNet-50	0.170
ResNet-101	0.125
DenseNet-121	0.905
DenseNet-169	0.160

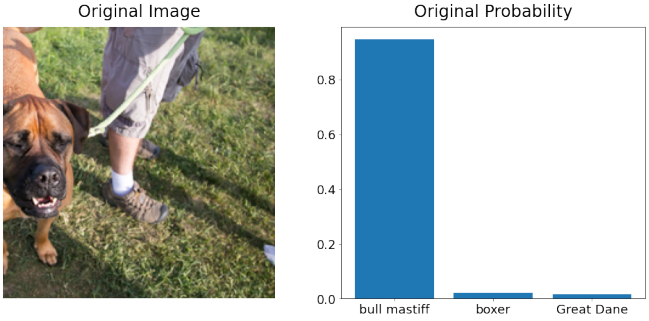
其中 DenseNet-121 的 success rate 最高，因此我猜測 black box model 爲 DenseNet-121，也因此我於第 1 題之中使用了 DenseNet-121 作爲 proxy model 對其進行攻擊。

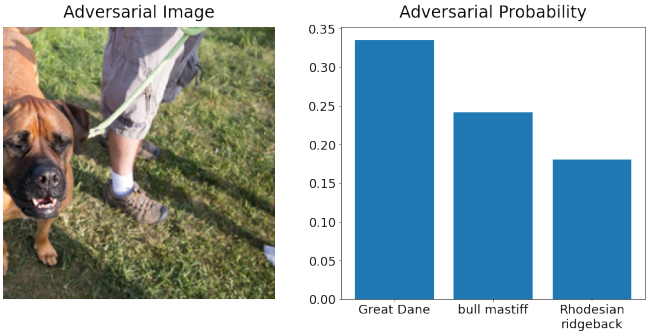
3. (1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖(分別取前三高的機率)。

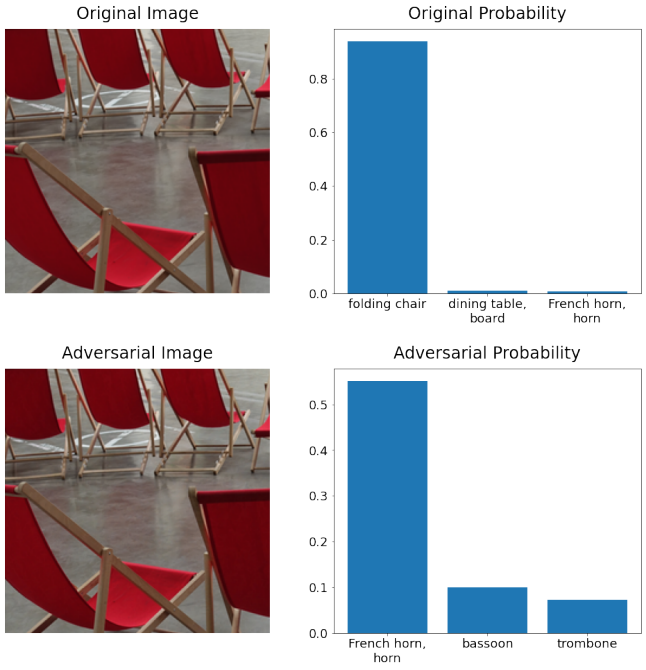
以下爲我任取三張照片對其進行攻擊，並使用 DenseNet-121 對攻擊前後的照片進行辨識所得到的機率分布圖：











4. (2%) 請將你產生出來的 adversarial image，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。

我將被 BIM 攻擊的照片使用 Gaussian filter 進行 smoothing 之後，再將其輸入 DenseNet-121 進行辨識，此時攻擊的 success rate 從 100% 下降到了 59%。Gaussian filter 會將照片中的邊界模糊化，而 noise 的尺寸更小，因此可以使用 Gaussian filter 將其去除，進而達到被動防禦，降低攻擊的 success rate，然而，將圖片使用 Gaussian filter 進行 smoothing 再輸入 model，可能會使得未被攻擊的正常照片其外觀被稍微改變，導致 model 將其辨識錯誤。