D21-鳥屎攻擊法:針對車牌辨識系統進行對抗性樣本攻擊

學生:羅政傑　　指導教授: 阮文齡　教授

摘要:

License Plate Recognition(LPR)車牌辨識近幾年被廣泛的應用, 通常會運用到CNN來協助辨識, 但經過實驗, 許多研究指出CNN在影像辨識方面, 容易因為影像的擾動而受到攻擊, 產生出錯誤的結果, 這些擾動後的影像就稱為對抗性樣本。對抗性樣本的擾動通常讓人眼難以分辨, 卻能使機器辨識錯誤, 然而鳥屎攻擊法製造的擾動是人眼可視的, 利用鳥屎出現在車上的合理性, 降低被認出是人為攻擊的可能性, 同時也能使LPR辨識錯誤, 是一種黑盒的對抗性樣本攻擊。

方法與架構 :

1. Datasets: AOLP (Application-oriented License Plate)

2. LPR system:

YOLOV8 (抓取車牌位置) + Easyocr (辨識車牌號碼)

辨識步驟 :

(1) 使用YOLOV8從圖片抓取車牌位置, 再切除多餘的影像

(2) 放大車牌後將圖片灰階化, 再用高斯模糊, 反白影像與設置邊界, 完成影像處理

(3) 使用Easyocr辨識車牌, 回傳結果 (Text: 54433ZS)

辨識826張圖片 / Accuracy = 0.9

3. Attack Method:

(1) 使用YOLOV8抓取車牌上每個號碼的位置, 截取號碼圖片後, 透過Easyocr的判斷選出信心值最低的號碼, 作為欲攻擊的對象(同時避免選擇不易攻擊的號碼)

(2) 用原號碼圖片透過標準差控制擾動位置、密度, 生成數個添加隨機擾動的照片, 並選出可能讓LPR辨識錯誤的擾動圖片

(3) 將擾動圖片當成輸入放入Genetic Alogrithm(遺傳演算法)中, 經過計算產生出對抗性樣本

隨機擾動 產生原理:

center = (round(Normal(w/2,center_std_dev)), round(Normal(h/2,center_std_dev)))

x = round(Normal(center[0],std_dev)), in range(num_points)

y = round(Normal(center[1],std_dev), in range(num_points)

perturbed_img[y,x] = [255,255,255]

adv_img = original_img + perturbed_img

Genetic Alogrthm原理:

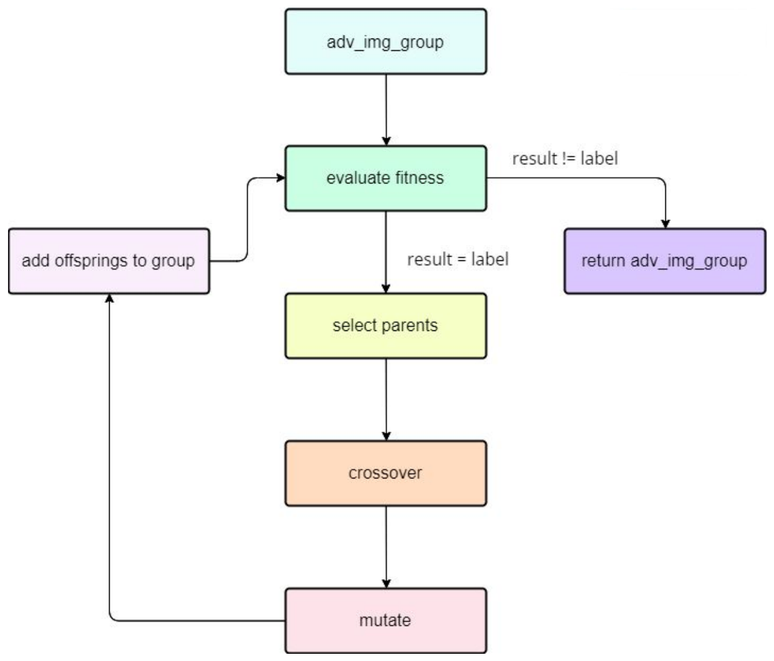
evaluate fitness: 透過Easyocr計算族群中圖片的可信度, 同時辨別是否有不等於label(正確車牌號碼)的結果, 若有則回傳該圖片, 若沒有則繼續往下執行

select parents: 從族群中選出可信度較高的幾張圖片當作parents

crossover: 讓parents互相產生offsprings, offsprings有parents較明顯的特徵

mutate: 每個offsprings都有定的機率變異, 使圖片與parents有些許差異

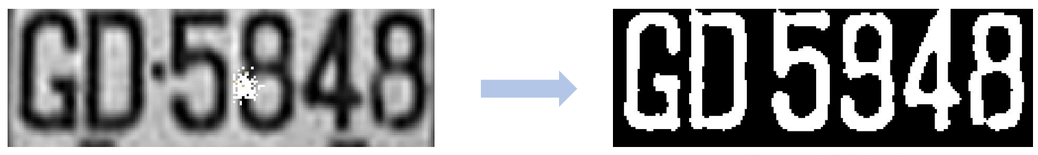
add offsprings to group: 將生成的offsprings與他們的parents合成新的族群



Genertic Alogrithm流程圖

Text: GD5848

選取容易受到攻擊的車牌號碼, 並加上擾動



Text: GD5948

將擾動後的車牌號碼圖片貼回原始車牌, 最後用LPR辨識該車牌, 可以發現從'GD5848'變成'GD5948'

其餘對抗性樣本 :

結果:

最後測試結果為嘗試攻擊250張圖片 / Accuracy = 0.73, 由於是黑盒攻擊, 所以並沒有辦法得到模型的梯度獲得模型預測的方向, 只能透過辨識的結果來製造對抗性樣本, 因此要花蠻多時間去計算跟測試圖片, 倘若測試步驟更加繁瑣, 將辨認時間拉長, 是可以有效提高攻擊的準確度。雖然鳥屎攻擊法效率跟準確度不算高, 但是我認為利用人類對於事物認知的合理性, 使人類疏忽掉細節, 也是一種有效的攻擊方式, 是資訊安全該提防的一點。