TẤN CÔNG HỘP ĐEN DỰA TRÊN TẬP CÁC MÔ HÌNH DỰ ĐOÁN DÀY ĐẶC

Hà Trọng Tài - 21520436
Trường Đại học Công nghệ thông tin TP Hồ Chí Minh

Phan Trường Trí - 21520117
Trường Đại học Công nghệ thông tin TP Hồ Chí Minh

Tóm tắt

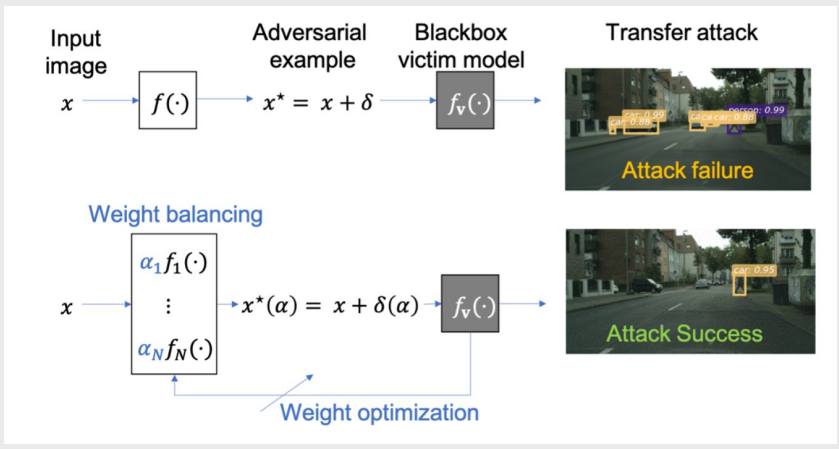
Tạo các cuộc tấn công có mục tiêu vào các mô hình dự đoán dày đặc (object detection và segmentation)

- Tạo một lớp nhiễu nhằm đánh lừa các mô hình dự đoán dày đặc
- Tấn công có mục tiêu thường khó hơn tấn công không mục tiêu
- Đánh giá khả năng tấn công của phương pháp

Động lực

- Các nghiên cứu hiện nay thường nhắm đến mục tiêu là tăng hiệu suất của mô hình nhưng sự an toàn của mô hình vẫn khá quan trọng.
- Tấn công hộp đen vào các mô hình dự đoán dày đặc vẫn là một nhiệm vụ khá thử thách

Tổng quan



Hình 1: Tổng quan về mô hình tấn công đề xuất

Input: 1 tấm ảnh
Output: ảnh sau khi thêm nhiễu

Hình 2: Input/Output của bài toán

Mô tả

1. Tìm hiểu bài toán

- Tìm hiểu các nghiên cứu về tạo các cuộc tấn công vào các mô hình dự đoán dày đặc
- Tìm hiểu các nghiên cứu liên quan đến tấn công hộp đen, đặc biệt là các phương pháp Transfer attack và query attack.

2. Xây dựng mô hình

- Tìm hiểu quá trình tấn công vào victim của phương pháp Transfer attack.
- Xây dựng mô hình Ensemble attack dựa trên ý tưởng của model Transfer attack với cải tiến thay vì sử dụng một model thì ta sử dụng một tập các models whitebox để tạo cuộc tấn công vào victim model. Tăng tính bao phủ khi tấn công vào các victim models.
- Đưa ra phương hướng chọn tập hợp models whitebox để tối ưu việc tính toán đồng thời vẫn đảm bảo tính bao phủ khi tạo cuộc tấn công vào các victim models.
- Đưa ra giải pháp cho việc kết hợp các models whitebox để tạo ra một cuộc tấn công duy nhất đảm bảo các models này đều bị tấn công thành công cùng lúc.

3. Đánh giá mô hình

- Tấn công vào các models object detection thì ta sử dụng dataset COCO bao gồm 328K hình ảnh, với hơn 200.000 hình ảnh có nhãn cho 1,5 triệu mẫu đối tượng.
- Tấn công vào các models segmentation thì ta sử dụng dataset Pascal VOC 2012 chứa tổng cộng 11.540 hình ảnh, mỗi hình ảnh bao gồm một tập hợp các đối tượng từ 20 lớp khác nhau
- Đánh giá mô hình đối với victim model là dựa trên tỷ lệ tấn công thành công của mô hình khi thực hiện tấn công vào victim model, đồng thời cũng đánh giá mức độ ổn định (chống bị tấn công) của victim model.
- So sánh với các phương pháp khác để đưa ra nhận xét về ưu điểm và khuyết điểm của mô hình

4. Đề xuất hướng đi chống lại các cuộc tấn công

- Phân tích kết quả đánh giá của mô hình, đặc biệt là các trường hợp mà victim model bị tấn công thành công nhằm rút ra đặc điểm chung của các điểm dữ liệu.
- Từ đó đưa ra kết luận về nguyên nhân mà victim model bị đánh lừa để đưa ra hướng giải quyết, hạn chết nguy cơ bị tấn công.