Phát hiện giả mạo tín hiệu GPS trong xe tự hành sử dụng mô hình học máy bán giám sát CNN

Tác giả[1] : Ngô Thái Hưng

[1] Trường ĐH Công nghệ Thông tin – ĐHQG HCM

Mục tiêu

Mục tiêu của nghiên cứu này là:

- Phát triển mô hình học máy bán giám sát CNN hiệu quả để phát hiện giả mạo tín hiệu GPS trong phương tiện tự hành.
- Nâng cao khả năng tổng quát của mô hình bằng cách sử dụng dữ liệu mô phỏng.
- Đề xuất giải pháp tích hợp mô hình vào hệ thống lái tự động.

Lý do chọn đề tài

- Phương tiện tự hành có kết nối (CAVs) tiềm năng mang lại nhiều lợi ích. Tuy nhiên, sự phụ thuộc lớn vào Hệ thống Định vị Toàn cầu (GPS) để điều hướng và kiểm soát khiến các CAVs dễ bị tổn thương trước các cuộc tấn công giả mạo GPS (GPS spoofing). Tấn công này có thể khiến CAVs xác định sai vị trí, dẫn đến tai nạn hoặc mất kiểm soát. Do đó, việc bảo vệ CAVs khỏi tấn công giả mạo GPS là rất quan trọng.

Overview

Chuẩn bị dữ liệu

Label Dataset:
A DATASET for GPS Spoofing Detection on Autonomous Vehicles

Unlabeled Dataset:
Extract from CARLA Simulation

Huấn luyện ban đầu

Huấn luyện mô hình CNN có khả năng phân loại tín hiệu GPS với dữ liệu có nhãn (Label Dataset)

Gán nhãn giả

Sử dụng mô hình CNN đã được huấn luyện để dự đoán nhãn cho tập dữ liệu không có nhãn (unlabeled dataset)
Kiểm tra lại và gán nhãn thủ công.

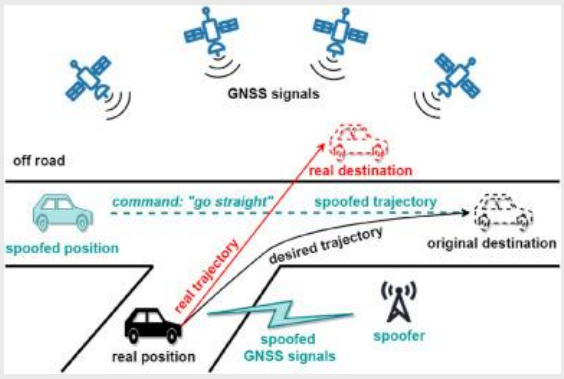
Kết hợp và huấn luyện lại

Tăng cường mô hình bằng cách huấn luyện lại với nhiều dữ liệu hơn.

Mô tả

1. Giới thiệu

Sự phụ thuộc vào GPS khiến AVs dễ bị tấn công giả mạo GPS (GPS spoofing), dẫn đến hậu quả nghiêm trọng như mất kiểm soát hoặc tai nạn. Do đó, phát hiện và ngăn chặn các cuộc tấn công này là rất cần thiết.



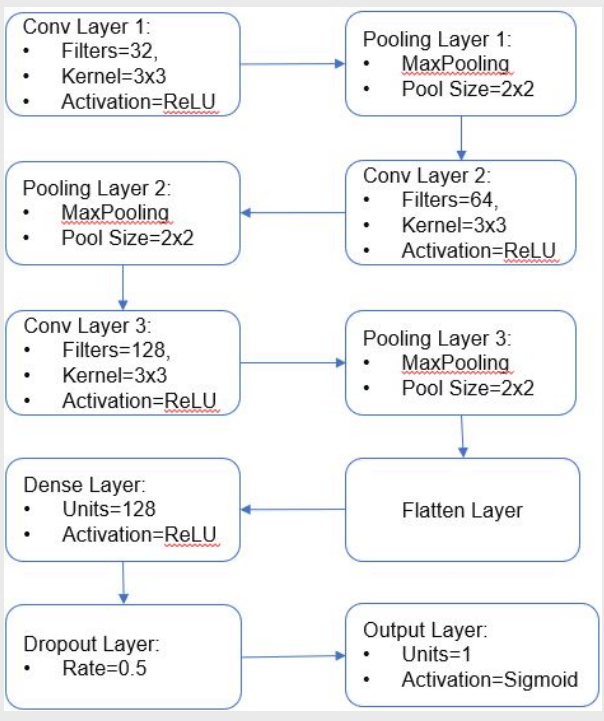
Hình 1. Cuộc tấn công giả mạo GPS nhằm vào xe tự hành

2. Chuẩn bị bộ dữ liệu

Label Dataset:
A DATASET for GPS Spoofing Detection on Autonomous Vehicles. Bộ dữ liệu kết quả chứa tổng cộng 158,170 mẫu, bao gồm 55% tín hiệu hợp lệ và 45% mẫu tín hiệu giả mạo.

Unlabeled Dataset:
Các mẫu tín hiệu GPS mô phỏng từ công cụ mô phỏng **CARLA Simulation**.

Sử dụng tập dữ liệu GPS có nhãn (label dataset) để huấn luyện mô hình CNN. Đây là bước huấn luyện ban đầu nhằm tạo ra một mô hình có thể học được các đặc trưng cơ bản từ dữ liệu có nhãn.

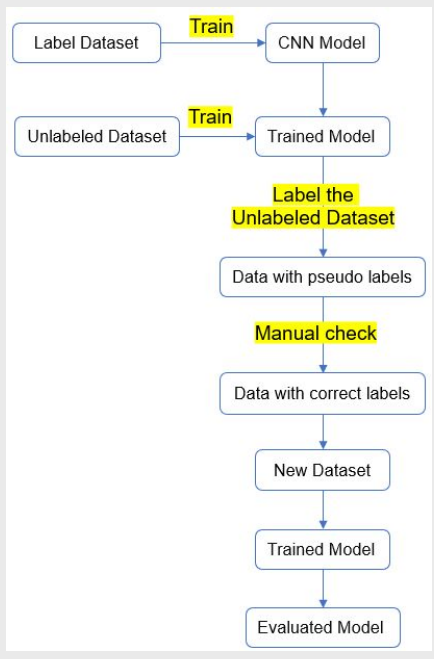


Hình 2. Mô hình máy học CNN

Sử dụng mô hình CNN đã được huấn luyện ở bước 1 để dự đoán nhãn cho tập dữ liệu không có nhãn (unlabeled dataset), tạo ra các nhãn giả (pseudo-labels). Sau đó, kiểm tra và chỉnh sửa thủ công các nhãn giả để tạo ra bộ dữ liệu lớn hơn.

3. Huấn luyện, kiểm thử mô hình

Kết hợp tập dữ liệu ban đầu có nhãn và tập dữ liệu đã được gán nhãn chính xác (sau khi kiểm tra thủ công). Sử dụng toàn bộ tập dữ liệu này để huấn luyện lại mô hình CNN, tạo ra mô hình cuối cùng với hiệu suất tốt hơn.



Hình 3. Mô hình hệ thống tổng quát

Việc sử dụng mô hình đã huấn luyện để gán nhãn giả và sau đó huấn luyện lại mô hình với dữ liệu mở rộng giúp cải thiện khả năng tổng quát hóa và hiệu suất của mô hình, đặc biệt khi dữ liệu có nhãn ban đầu hạn chế.

NII

Ngô Thái Hưng – Trường Đại học Công nghệ Thông tin – ĐHQG HCM
TEL : 0854612754 Email : hungnt.18@grad.uit.edu.vn