자율주행 차량 보안을 위한 생성형 AI 기반 악성 트래픽 데이터 생성 방법 연구

김소영, 김기천*

건국대학교

ypd07026@gmail.com, *kckim@konkuk.ac.kr

A Study on Generating Malicious Traffic Data Using Generative AI for Autonomous Vehicle Security

Kim So Young, Kim Kee Cheon*

Konkuk Univ.

요 약

자율주행 차량은 최첨단 센서와 인공지능 기술을 활용하여 빠르게 발전하고 있지만 그만큼 보안 위협도 증가하고 있다. 기존 연구들은 다양한 적대적 공격에 대응하기 위한 시스템과 알고리즘을 개발했으나, 현실적인 악성 트래픽 데이터의 부족으로 한계를 보이고 있기에 본 연구는 생성형 인공지능, 특히 CTGAN을 활용하여 스푸핑, 데이터 조작, 통신 방해, 서비스 거부 공격 등 네 가지 주요 보안 위협 유형에 따른 악성 트래픽 데이터를 생성함으로써 데이터 불균형 문제를 해결하고자 한다. 이를 통해 AI 기반 자율주행 트래픽 분류 모델의 학습을 강화하고, 자율주행 시스템의 보안성을 보다 정밀하게 분석할 수 있는 기반을 마련하는 것을 목표로 한다.

Ⅰ. 서 론

자율주행은 최첨단 센서 및 인공지능 기술의 발달과 함께 빠르게 발전하여 활발하게 연구되고 있으며 교통사고 감소와 교통 체증 해소 등 사회 전반에 걸쳐 긍정적인 영향을 줄 수 있는 차세대 기술로 주목받고 있다. 그러나 자율 주행 기술이 발전함에 따라 자율주행 보안 문제도 중요한 과제로 떠오르고 있다.[1] 예를 들어, 악의적인 공격자가 자율주행 차량의 센서 데이터나 통신을 조작하여 주행을 방해하는 행위는 자율주행 시스템의 보안에 심각한 문제를 불러일으킬 수 있다. 따라서 이러한 보안 취약점을 미리 발견하고 효과적으로 대응책을 마련하는 것이 안전한 자율주행을 위해 필수적이다.

기존의 연구들은 자율주행에 대한 적대적 공격에 대응하기 위해 다양한 시스템과 알고리즘을 개발하였으나 이러한 연구들은 실제 환경에서 발생할 수 있는 악성 트래픽 데이터의 부족으로 다양한 공격 시나리오를 반영하는 데 한계가 있다.[2],[3],[4] 이를 해결하기 위해 악성 트래픽 데이터를 생성하고 이를 바탕으로 AI 기반 자율주행 트래픽 분류 모델을 학습시켜 보안 시스템을 구축하는 것이 매우 중요하다.

따라서 본 연구는 정상 트래픽과 악성 트래픽을 효과적으로 분류할 수 있는 보안 모델을 학습시키기 위해 생성형 인공지능을 활용하여 악성 트래픽을 생성하는 방안을 제안한다. 이를 통해 정상 트래픽 데이터와 악성 트래픽 데이터 간의 불균형을 문제를 개선하고, 자율주행의 보안 취약점을 보다 자세하게 분석할 수 있다.

Ⅱ. 본론

본 논문에서는 자율주행의 보안을 향상시키기 위해 생성형 인공지능을 활용한 악성 트래픽 데이터 생성 방법을 제안한다. 이를 위해 다양한 자율주행 공격 유형 정의와 데이터 전처리, CTGAN(Conditional Table GAN) 모델의 학습방법을 차례대로 설명한다.

2-1. 자율주행 보안 위협 유형

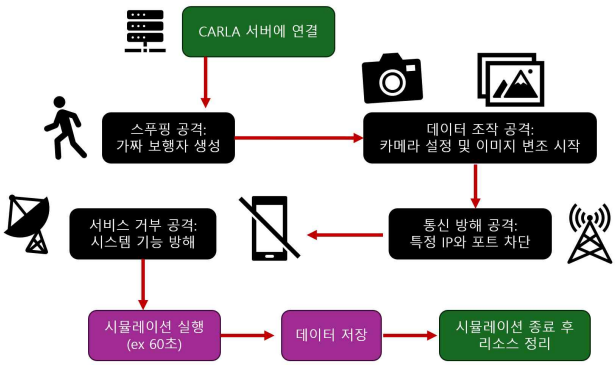
자율주행에는 다양한 보안 위협 유형이 있으며 공격 방법이 다양하고 공격 범위도 무궁무진하다. 이러한 다양한 보안 위협 중 본 논문에서는 악성 트래픽 유형을 스푸핑 공격, 데이터 조작 공격, 통신 방해 공격, 서비스 거부 공격 총 4가지로 분류한다. 스푸핑 공격은 센서 데이터를 위조하여 차량이 잘못된 정보를 인식하도록 유도하는 공격으로 위조 대상은 가짜 보행자, 잘못된 신호등 정보, 위조된 다른 차량의 위치 정보 등이 포함될 수 있다. 데이터 조작 공격은 센서 데이터나 통신 데이터를 변경하여 차량의 주행 결정을 왜곡시키는 공격으로 카메라 이미지의 일부분을 조작하여 장애물을 숨기거나, 라이다 데이터의 포인트 클라우드를 변경하여 차량의 위치를 잘못 인식하게 하는 공격이다. 통신 방해 공격은 차량 간 또는 차량과 인프라 간의 통신을 방해하여 데이터 교환을 방해하고 잘못된 정보를 주입하는 공격으로 패킷 손실 및 지연, 위조된 메시지 전송 등이 포함된다. 마지막으로 서비스 거부 공격은 시스템의 정상적인 기능을 방해하여 자율주행 시스템의 작동을 중단시키는 공격으로 과도한 데이터 전송을 통해 네트워크 대역폭을 소모하거나 CPU/메모리 자원을 고갈시켜 시스템의 응답성을 저하시키는 공격이다. 이러한 네가지 유형의 공격을 악성 트래픽 데이터 생성에 적용하기 위한 주요 파라미터는 <표 1>과 같다.

공격 유형	주요 파라미터
스푸핑 공격	가짜 보행자 수 , 신호등 정보 위조 정도, 위조된 차량 위치 범위
데이터 조작 공격	이미지 변주 범위, 라이다 데이터 변주 범위
통신 방해 공격	차단할 IP 범위, 차단할 포트, 패킷 손실률, 지연 시간
서비스 거부 공격	데이터 전송 속도, CPU/메모리 사용량, 공격 지속 시간

<표 1> 자율주행 공격 유형 별 주요 파라미터

2-2. 악성 트래픽 데이터 생성

악성 트래픽 데이터를 생성형 AI를 이용하여 증강시키기 위해서는 해당 악성 트래픽 분류 기준에 따라서 최소한의 기본 데이터셋이 필요하다. 이를 위해 오픈 소스 자율주행 시뮬레이터로, 다양한 주행 시나리오를 시뮬레이션할 수 있는 환경을 제공하는 CARLA를 이용하여 <표 1>에서 정의한 파라미터 조건에 맞춰 [그림 1]처럼 데이터를 생성하여야 한다.



[그림 1] 악성 트래픽 데이터 생성 흐름도

2-3. CTGAN 모델을 사용한 통합 데이터 생성

생성된 악성 트래픽 데이터는 한 클래스 당 100개씩, 총 400개를 사용하고 정상 트래픽 데이터는 10000개를 사용해 분류 모델에 학습시킨다. 이때 악성 트래픽 데이터와 정상 트래픽 데이터의 개수 차이로 인해 데이터 불균형 문제가 발생하여 학습이 원활하게 이루어지지 못한다. 따라서 이를 해결하기 위해 CTGAN을 사용하여 추가로 악성 트래픽 데이터를 생성한다. CTGAN은 테이블 데이터셋을 생성해 내는데 좋은 성능을 보여준다. CTGAN은 일반적으로 이미지 생성에 사용되는 GAN 기법을 활용하여 실제 데이터와 유사한 가짜 테이블 데이터를 생성한다.[5] 이때 데이터의 범위와 적절성을 높이기 위해 다양한 파라미터와 원핫 인코딩을 사용하여 <표 2>처럼 조건 벡터를 설정한다.

공격 유형 / 파라미터	스푸핑	데이터 조작	통신 방해	서비스 거부
원핫 인코딩	[0, 1, 0, 0, 0]	[0, 0, 1, 0, 0]	[0, 0, 0, 1, 0]	[0, 0, 0, 0, 1]
가짜 보행자 수	5	0	0	0
위치 범위 x	10	0	0	0
위치 범위 Y	10	0	0	0
신호등의 정보 위조 정도	0.7	0	0	0
위조 차량 위치 조정 범위	5	0	0	0
조작 센서 데이터의 비율	0	0.2	0	0
변조 범위	0	150	0	0
조작 유형 코드	0	1	0	0
차단할 IP 범위 코드	0	0	1	0
차단할 포트	0	0	2000	0
패킷 손실률	0	0	0.3	0
지연 시간	0	0	100	0
데이터 전송 속도	0	0	0	1000
CPU/메모리 소모량	0	0	0	0.8
공격 지속 시간	0	0	0	60

<표 2> 조건 벡터 예시

이렇게 CTGAN을 이용하여 생성된 데이터는 추후 AI 기반 자율주행 트래픽 분류 모델을 학습시키는 데 사용할 수 있을 것이다.

Ⅲ. 결론

본 논문에서는 자율주행 차량의 보안성을 강화하기 위해 생성형 인공지능을 활용한 악성 트래픽 데이터 생성 방법을 제안하였다. 특히, 스푸핑 공격, 데이터 조작 공격, 통신 방해 공격, 서비스 거부 공격 등 네 가지 주요 보안 위협 유형을 정의하고, 각 공격 유형에 따른 주요 파라미터를 설정하였다. 이를 기반으로 CARLA 시뮬레이션 환경을 이용하여 실제 데이터와 유사한 최소한의 악성 트래픽 데이터 생성 및 CTGAN 모델의 설계를 제안하였다. 제안한 방법은 정상 트래픽 데이터와 악성 트래픽 데이터 간의 불균형 문제를 효과적으로 해결하여 인공지능 분류 모델을 활용한 자율주행 보안 시스템 설계에 활용될 수 있을 것으로 기대된다. 그러나 본 연구는 아직 실험을 통해 제안한 방법의 실효성을 검증하지 못한 상태이기에 향후 연구에서는 실제로 실험을 진행하고 정의된 네 가지 공격 유형 외에도 더욱 다양한 보안 위협 시나리오를 포함하여 악성 트래픽 데이터의 범위를 확장할 필요가 있다. 또한 실제 자율주행 차량에 적용하여 생성된 악성 트래픽 데이터가 실제로 수집된 데이터와 비교했을 때의 일반화 능력을 검증하는 실험을 수행해야 할 것이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임(IITP-2024-RS-2020-II201834)

참 고 문 헌

[1] 권순홍, 이종혁. “자율 주행 자동차 보안 위협 및 기술 동향. 정보보호학회지”, 30(2), 31-39, 2020.

[2] 김채현, 이진규 외 4명, “자율주행을 위한 적대적 공격 및 방어 딥러닝 모델 연구”, Nov. 2022.

[3] Kosmanos, D., Xenakis, A., Chaikalis, C., “The Impact of Spoofing Attacks in Connected Autonomous Vehicles under Traffic Congestion Conditions”. Telecom 2024, 5.

[4] K. M. A. Alheeti, A. Alzahrani and D. Al Dosary, "LiDAR Spoofing Attack Detection in Autonomous Vehicles," IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2022.

[5] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni, “Modeling Tabular data using Conditional GAN”, Jul. 2019.