Bi-LSTM VAE 기반 차량 CAN 침입 탐지 시스템

김용수[1], 강효은[2], 김호원*
부산대학교 정보융합공학과
{yongsu, hyoeun0915, howonkim}@pusan.ac.kr

Bi-LSTM VAE based Intrusion Detection System for In-Vehicle CAN

Yong-Su Kim[1], Hyo-Eun Kang[2], Ho-Won Kim*
Pusan National University

요 약

승차 공유, 카풀, 렌터카의 이용률이 증가하면서 많은 사용자가 동일한 차량에 로컬 액세스 할 수 있는 시나리오가 더욱 보편화됨에 따라 차량 네트워크에 대한 공격 가능성이 커지고 있다. 차량용 CAN Bus Network에 대한 DoS(Denial of Service), Fuzzy Attack 및 Replay Attack과 같은 공격은 일부 ECU(Electronic Controller Unit) 비활성 및 작동 불능 상태를 유발한다. 에어백, 제동 시스템과 같은 필수 시스템이 작동 불가 상태가 되어 운전자에게 치명적인 결과를 초래할 수 있다. 차량 네트워크 침입 탐지를 위하여 많은 연구가 진행되고 있으나, 기존 화이트리스트를 이용한 탐지 방법은 새로운 유형의 공격이 발생하거나 희소성이 높은 공격일 때 탐지하기 어렵다. 본 논문에서는 인공신경망 기반의 CAN 버스 네트워크 침입 탐지 기법을 제안한다. 제안하는 침입 탐지 기법은 2단계로 나누어진다. 1단계에서 정상 패킷 분포를 학습한 VAE 모형이 이상 탐지를 수행한다. 이상 패킷으로 판정될 경우, 2단계에서 인코더로부터 추출된 잠재변수와 VAE의 재구성 오차를 이용하여 공격 유형을 분류한다. 분류 결과의 신뢰점수(Confidence score)가 임계치보다 낮을 경우 학습하지 않은 공격으로 판단한다. 본 연구 결과물은 정보보호 연구·개발 데이터 챌린지 2019 대회의 차량 이상징후 탐지 트랙에서 제공하는 정상 및 3종의 차량 공격시도 패킷 데이터를 대상으로 성능을 평가하였다. 실험을 통해 자동차 제조사의 규칙이나 정책을 사전에 정의하지 않더라도 낮은 오탐율로 비정상 패킷을 탐지해 낼 수 있음을 확인할 수 있다.

1. 서론

최근 공유 모빌리티 산업이 크게 발달함에 따라 차량공유 시장은 가파른 성장세를 유지하고 있다. 글로벌 시장조사업체인 맥킨지글로벌 연구소에 따르면 차량공유의 확산으로 2030년에는 일반소비자 자동차 구매가 현재보다 최대 연간 400만 대 감소하고 차량공유용 판매는 200만대 증가할 것으로 전망했다. 삼정KPMG는 차량공유 시장 규모는 2025년 약 238조 원을 기록한 이후 2050년 약 475조 원을 초과할 것으로 전망하고 있다[1].

이와 관련하여 자동차 산업의 보안 위협도 증가하는 추세이다. 특히, 차량 공유 서비스를 통해 많은 사용자가 같은 차량에 접근할 수 있어 공유 차량을 대상으로 하는 위협 시나리오가 증가하고 있다. 이러한 공격은 차량 내 장비와 시스템을 연결하고 통신할 수 있는 CAN 네트워크 프로토콜을 악용하여 이루어진다.

CAN(Controller Area Network)은 차량 내부 기기 간 통신을 위해 설계된 차량 네트워크 표준 프로토콜이다. CAN 네트워크 모든 ECU가 브로드캐스트 방식으로 메시지를 수신하며, 자신에게 필요한 메시지일 경우 수신하고 아닐 경우 무시한다. CAN 통신은 별도의 인증 또는 액세스 제어 체계가 존재하지 않아 각 수신기는 발신자를 식별할 수 없으며, 수신된 패킷에 대한 적법 여부를 확인할 수 없다. CAN 버스에 대한 로컬 액세스 공격은 이와 같은 CAN 설계 취약점을 이용하여 이루어진다. CAN 표준은 공격자가 절대로 CAN 버스에 대한 무단 액세스를 얻지 못한다는 가정하에 구현되었다. 이로 인해 시스템 내에서 CAN 버스의 모든 데이터는 신뢰되므로, CAN에서는 진짜 오류 메시지와 사이버 공격자에 의해 만들어진 가짜 오류 메시지를 구별할 수 없게 된다. 따라서 CAN 네트워크에 위조 패킷이 주입되거나 패킷 조작과 같은 악의적인 공격이 발생하면 운전자의 생명에 치명적인 위협을 초래할 수

있다. 공격자가 차량에 직접적인 접근을 하는 것은 어렵지만, 차량 공유 서비스가 증가함에 따라 다수가 같은 차량에 접근할 수 있어 로컬 액세스 공격에 대한 위협이 증가하고 있다.

본 논문에서는 차량용 CAN 네트워크의 침입 탐지를 위한 인공신경망 알고리즘 기반의 탐지 기술을 제안한다. 제안하는 모델은 2단계로 이루어져 있으며, VAE(Variational Autoencoder) cell 내부에 Bidirectional LSTM을 적용하고 정상 패킷으로만 학습한다. 새로운 패킷 데이터가 입력되면 재구성 오차를 이용하여 정상/이상을 판정한 후, 이상 패킷 데이터에 대하여 VAE의 인코더에서 출력된 잠재변수와 재구성 오차를 이용하여 세부 공격기법을 분류한다. 신뢰점수 (Confidence score)가 임계치보다 낮을 경우 알려지지 않은 공격으로 판단하며, 임계치보다 높을 경우 가장 높은 공격 클래스를 출력한다.

본 논문의 구성은 제2장에서 차량 네트워크 침입탐지 관련 선행 연구를 소개하고, 제3장에서는 본 연구에서 제안하는 침입 탐지 및 공격기법 분류를 수행하는 방법을 논한다. 제4장에서는 실험 결과 및 분석을 다루고 제5장을 끝으로 결론을 도출한다.

2. 관련연구

일반적으로 차량 CAN 통신 시스템 IDS에 관한 연구 논문은 CAN에 흐르는 신호를 통계적으로 학습하여 임계치를 벗어나는 신호에 대하여 공격받았다고 판단하는 구조이다. 이상 패킷을 식별하기 위해 CAN 트래픽의 타이밍 간격, 패킷 시퀀스의 주파수 등을 이용하고 있다. 고정된 시간 간격 및 주파수로 모든 수신기 ECU에 지속해서 브로드캐스팅하는 CAN 네트워크 특성을 이용하여 정상적인 트래픽과의 편차로 시스템의 공격을 식별할 수 있다.

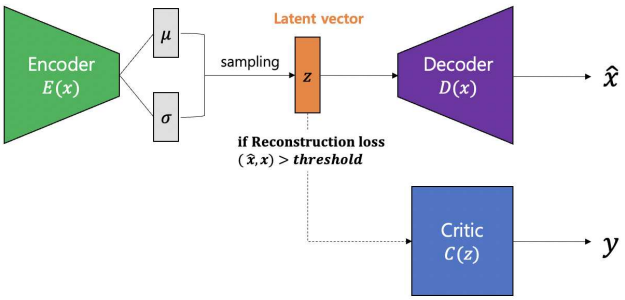
최근 머신러닝을 활용한 이상 탐지 연구도 활발히 이루어지고 있다. Müter[2] 등은 엔트로피 기반의 CAN 네트워크 내 이상 탐지 기법을 제안하였다. CAN 네트워크는 일반적인 컴퓨터 네트워크에 비해 패킷 유형이 정해져 있고 제한적인 트래픽을 가지고 있기 때문에 정상적인 데이터 분포의 불확실성이 매우 낮은 것을 확인할 수 있다. 이러한 특성 때문에 CAN 네트워크를 대상으로 한 공격은 정상 데이터 분포와 많이 떨어져 있으며, 엔트로피 값이 증가하기 때문에 침입 탐지를 원활하게 할 수 있다고 주장한다.

Taylor[3] 등은 LSTM(Long Short-Term Memory)를 이용한 이상 탐지 기법을 제안하였다. 이 방법은 정상적인 CAN 패킷 시퀀스에서 다음 값을 예측하도록 LSTM을 학습하여, 예측값과 많은 차이를 보이는 패킷이 나타나면 공격이 발생했다고 판단한다. 해당 기법은 낮은 오탐율을 보이지만 단일 타입의 CAN 패킷에서만 작동하는 단점이 있다.

GIDS[4]는 GAN(Generative Adversarial Networks) 기반으로 제안된 IDS 모델이다. GIDS는 정상 데이터 분포를 학습하여 분포를 벗어나는 데이터에 대해서 공격이 발생했다고 판단한다. 해당 기법은 알려지지 않은 공격에 대해서도 높은 정확도로 탐지할 수 있다는 장점이 있다.

3. 침입탐지 프레임워크

본 연구에서는 제안하는 차량 네트워크 침입탐지 프레임워크는 (그림 1)과 같이 2단계로 이루어진다. 1단계에서는 VAE[5]를 통해서 정상 패킷의 특성을 학습하여 정상 및 이상을 탐지한다. 여기서 $E(x)$는 VAE의 인코더, $D(x)$는 VAE의 디코더를 의미한다. 2단계는 공격 기법 추론 모델인 Critic $C(z)$로써, VAE에서 추출한 공격 패킷에 대한 특징과 재구성 에러를 입력받아 공격기법을 추론한다. 출력값의 신뢰 점수(Confidence score)가 임계점보다 낮으면 학습하지 않은 공격기법으로 판별한다.



(그림 1) Proposed architecture

본 연구에서는 연속적인(sequential) 시계열(time series) 문제에 대하여 우수한 성능을 보이는 Bidirectional LSTM[6]을 적용한 인코더 모델을 구성하였다. 이를 통해 인코더는 정상 신호 패턴의 특성을 학습할 수 있다. 인코더의 Bidirectional LSTM 층에서는 입력 시퀀스 $X=\{x_1, x_2, x_3, ..., x_T\}$가 순차적으로 forward 층의 각 스텝으로 입력되면서 순방향의 상태 벡터와 역방향의 상태 벡터를 획득한다.

인코더는 식 (1)과 같이 잠재변수 샘플링에서 사용

할 2개의 매개변수 μ와 σ를 출력한다. VAE 입력 벡터X에 대하여, 잠재변수의 평균을 μ, 분산을 σ라 할 때, 디코더의 입력 잠재변수 z는 아래와 같이 표현된다.

$$z(X) = \mu(X) + \sigma(X)\epsilon, \epsilon \sim N(0,1) \quad (1)$$

디코더는 잠재변수로부터 식 (2)와 같이, L개의 복원 확률의 평균을 계산하여 이상치 점수를 구한다. 이상치 점수가 임계점보다 크면, 이상치로 판정한다.

$$Anomaly\ Score = \frac{1}{\frac{1}{L}\sum_{l=1}^{L} p_\Theta(x|z^{(l)})} \quad (2)$$

4. 실험 결과

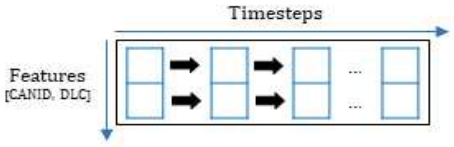
본 논문에서 제안하고 있는 모델의 우수성을 입증하기 위하여 다음과 같이 실험하고 그 결과를 비교 분석 하였다.

4.1 실험 환경

데이터세트는 2019년도 정보보호 R&D 데이터 챌린지 대회의 “자동차용 침입 탐지” 트랙에서 제공하는 차량 내부 네트워크 트래픽 데이터를 활용하였다.[7] 데이터세트는 3개 차종 (KIA Soul, CHEVOLET Spark, HYUNDAI Sonata)들의 정상 상태의 네트워크 트래픽과 공격 상태 (Flooding, Fuzzy, Malfunction, Replay)에 대한 네트워크 트래픽이며, 실제 차량에서 CAN 버스 네트워크 트래픽이다.

학습 데이터는 예선 데이터세트를 이용하였으며, 테스트세트 A는 본선 1차, 테스트세트 B는 본선 2차 데이터세트를 사용하였다. 테스트세트 A는 Normal, Fuzzy, Malfunction으로 이루어져 있으며, 테스트세트 B는 Malfunction, Replay로 이루어져 있다. 학습 데이터에는 Replay attack 패킷이 포함되어 있지 않기 때문에 본 연구에서는 알려지지 않은 공격으로 식별하였다.

인코더는 (그림 2)와 같이 Bidirectional LSTM을 cell을 가짐에 따라 Input 형태를 3차원 형태인 [sample size, time step, features]을 가지므로 time step은 30으로 지정하였다. 즉, 30개의 CAN 패킷마다 침입 탐지 여부를 판단하는 것이다.



(그림 2) Input shape of BiLSTM Encoder

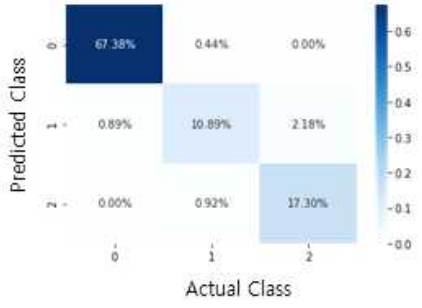
4.2 실험 결과

학습 데이터는 예선 데이터세트를 이용하였으며, 테스트세트 A는 본선 1차, 테스트세트 B는 본선 2차 데이터세트를 사용하였다. 테스트세트 A는 Normal, Fuzzy, Malfunction으로 이루어져 있으며, 테스트세트 B는 Malfunction, Replay로 이루어져 있다. 학습 데이터에는 Replay attack 패킷이 포함되어 있지 않기 때문에 본 연구에서는 알려지지 않은 공격으로 식별하였다. 정상 상태와 공격 상태에 대한 VAE 모델의 재구성 오차는 다음 <표1>과 같다.

<표 1> Average reconstruction error

Data type	Average reconstruction error
Normal	0.77
Flood attack	1.65
Fuzzy attack	4.94
Malfunction attack	0.06
Replay attack	1.31

실험을 통하여 알려진 공격기법에 대한 혼돈행렬 (Confusion matrix)은 다음 (그림 3)과 같다. 각 레이블은 학습한 공격 기법을 의미한다.



(그림 3) Confusion matrix
{0: Flood attack, 1: Fuzzy attack,
2: Malfunction attack}

5. 결론

본 논문에서 제안하는 탐지 모델의 구조는 2단계로 이루어지는 구조이다. 1단계 VAE 모델은 정상 패킷에 대해서만 학습하였기 때문에, 학습의 여부와 상관없이 알려지지 않은 공격에 대해서도 탐지할 수 있다.

또한, 공격 패킷 데이터의 분포가 불균형한 상태에서도 올바르게 분류한 결과를 통하여 VAE의 인코더 cell에 Bidirectional LSTM을 적용함으로써 패킷의 시계열 특성을 잘 반영할 수 있음을 확인하였다.

※ This work is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as 「Innovative Talent Education Program for Smart City」

References

[1] KPMG(2019.8.), TaaS 투자로 본 모빌리티 비즈니스의 미래

[2] M. Müter and N. Asaj, “Entropy-based anomaly detection for in-vehicle networks,” IEEE Intelligent Vehicles Symposium (IV), pp. 1110-1115, June. 2011.

[3] A. Taylor, S. Leblanc and N. Japkowicz, “Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks,” IEEE International Conference on Data Sceince and Advanced Analytics (DSAA), pp. 130-139, Oct. 2016.

[4] E. Seo, H. Song and H. Kim, “GIDS: GAN based Intrusion Detection System for In-Vehicle Network,” Proceedings of the 16th Annual Conference on Privacy, Security and Trust (PST), pp. 1-6, 2018.

[5] Kingma, Diederik P., and Max Welling. "An introduction to variational autoencoders." arXiv preprint arXiv:1906.02691 (2019).

[6] Graves, Alex, and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." Neural networks 18.5-6 (2005): 602-610.

[7] Mee Lan Han, Byung Il Kwak, and Huy Kang Kim. “Anomaly intrusion detection method for vehicular networks based on survival analysis.” Vehicular Communications 14 (2018): 52-63.