IoT 침입 탐지 시스템의 강건성 향상을 위한 GAN 기반 적대적 공격 생성-학습 프레임워크

최태준*, Segun Popoola**, 이현우***

*,***한국에너지공과대학교 (대학원생, 교수), **Anglia Ruskin Unversity (교수)

GAN-based Adversarial Attack Generation-Training Framework for Robust IoT Intrusion Detection Systems

Taejun Choi*, Segun Popoola**, Hyunwoo Lee***

*,***Korea Institute of Energy Technology (Graduate student, Faculty), **Anglia Ruskin University (Faculty)

요 약

딥러닝 기반 IoT 침입 탐지 시스템은 사물인터넷 환경에서 고도화되는 사이버 공격으로부터 대규모 네트워크 트래픽을 효과적으로 보호하기 위해 널리 활용되고 있다. 그러나 이러한 시스템은 작은 교란으로 오분류를 유발하는 적대적 공격에 취약하다. 기존 화이트박스 적대적 공격 기법은 도메인 제약을 충분히 반영하지 못해 현실적인 공격 데이터 생성에 한계가 있으며, 이는 적대적 학습의 효과를 저해한다. 본 연구에서는 도메인 제약을 보장하면서 현실적인 적대적 공격 데이터를 생성하는 GAN 기반 적대적 공격 생성 모델을 제안하고, 이를 적대적 학습 프레임워크에 통합한다. 실험 결과, 제안 방법은 평균 99.89%의 공격 성공률을 기록하여 다른 공격 기법들에 비해 높은 성공률을 보였으며, 침입 탐지 시스템의 성능을 크게 저하시켰다. 또한, 생성된 적대적 공격 데이터를 활용한 적대적 학습을 통해 모델의 견고성이 향상됨을 확인하였다.

I. 서론

사물인터넷(Internet of Things, IoT) 환경에서 침입 탐지 시스템(Intrusion Detection System, IDS)은 서비스 중단, 물리적 손상, 에너지 불안정 등을 유발하는 사이버 공격으로부터 대규모 실시간 네트워크 트래픽을 보호하는 데 필수적인 역할을 수행한다. 최근에는 탐지 성능 향상을 위해 대규모 네트워크 데이터로부터 복잡하고 비선형적인 패턴을 효과적으로 학습할 수 있는 딥러닝(Deep Learning, DL) 기반 IoT IDS가 널리 활용되고 있다 [1, 2].

그러나 DL 기반 IDS는 입력 데이터에 미세한 교란(perturbation)을 추가하여 오분류를 유도하는 적대적 공격(adversarial attack)에 취약하다. 적대적 공격은 정상 트래픽과의 유사성을 유지함으로 인해 탐지 오류를 유발하여 IDS 성능을 크게 저하시킬 뿐만 아니라, 핵심 인프라의 안정성을 심각하게 위협할 수 있다 [3, 4]. 이러한 문제를 완화하기 위한 다양한 방어 기법 중, 적대적 공격 데이터를 모델 학습 과정에 포함시켜 탐지 모델의 강건성을 향상시키는 적대적 학습(adversarial training)이 효과적인 방법으로 알려져 있다 [1, 3].

적대적 학습의 핵심 과제는 실제 환경을 반영한 현실적인 적대적 공격 데이터를 생성하는 데 있다. Fast Gradient Sign Method (FGSM), Jacobian-based Saliency Map Attack (JSMA), Projected Gradient Descent (PGD), Basic Iterative Method (BIM), Carlini & Wagner (C&W)와 같은 기존 화이트박스 공격 기법들은 네트워크 트래픽의 도메인 제약 조건을 충분히 반영하지 못하는 한계를 가진다. 그 결과, 생성된 적대적 공격 데이터가 실제 IoT 네트워크 환경을 충분히 반영하지 못하며, 이는 적대적 학습의 효과를 저해하는 요인으로 작용한다 [1].

이러한 한계를 극복하기 위해, 본 논문에서는 현실성과 도메인 유효성을 동시에 만족하는 적대적 네트워크 트래픽 데이터를 생성하기 위한

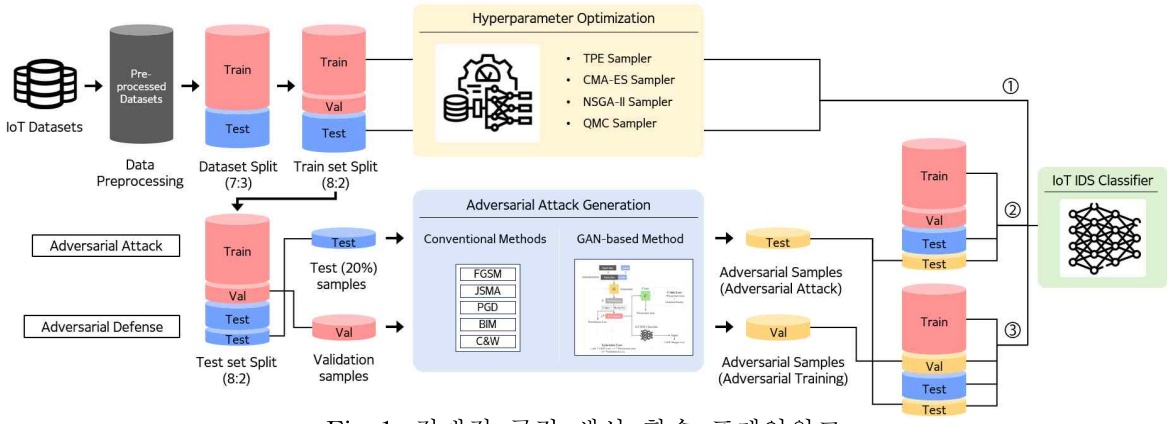


Fig 1. 적대적 공격 생성-학습 프레임워크

생성적 적대 신경망(Generative Adversarial Network, GAN) 기반 적대적 공격 생성 모델을 제안한다. 제안하는 모델은 범주형 특징과 특징 간 상관관계를 보존함과 동시에, 각 특징의 유효 범위를 유지함으로써 IoT 네트워크 환경에서의 데이터 타당성을 확보한다.

실험 결과, 제안한 GAN 기반 방법으로 생성된 적대적 공격 데이터는 평균 99.89%의 공격 성공률을 기록하여 기존 화이트박스 공격 기법에 비해 높은 공격 성공률을 보이며 IDS 성능을 효과적으로 저하시킴을 확인하였다. 또한, 이러한 데이터를 활용한 적대적 학습을 IDS에 적용할 경우, 감소된 탐지 성능이 회복되어 모델의 강건성이 향상됨을 확인하였다.

본 연구의 주요 기여는 다음과 같다.

1. 도메인 제약 조건을 반영하여 현실적인 적대적 공격 데이터를 더 효율적으로 생성하는 GAN 기반 공격 생성 모델을 제안한다.
2. 4개의 IoT 데이터셋과 DL 기반 IDS 모델을 대상으로 기존 화이트박스 공격 기법과 제안 방법의 성능을 비교·분석한다.
3. GAN 기반 적대적 공격 데이터를 활용한 적대적 학습이 IoT IDS의 강건성을 효과적으로 향상시킴을 실험적으로 입증한다.

II. 기존 적대적 공격 방법

본 연구에서는 공격자가 IoT IDS 모델의 구조와 파라미터를 포함한 모든 정보를 알고 있으며, 입력 특징에 대해 임의로 교란을 가할 수 있는 화이트박스 적대적 공격 환경을 가정한다. 대표적인 화이트박스 공격 기법으로는 FGSM, BIM, PGD, JSMA, C&W가 있다.

FGSM은 손실 함수의 그래디언트 부호 방향으로 단일 단계의 교란을 적용하여 효율적으로 오분류를 유도하는 기법이다. 이를 확장한 BIM은 FGSM을 반복적으로 적용하여 제한된 범위 내에서 교란을 점진적으로 누적함으로써 보다 강력한 적대적 공격 데이터를 생성한다. PGD 역시 반복적 업데이트를 기반으로 하며, 교란을 제약 집합 내로 투영하는 방식을 통해 강력한 화이트박스 공격 성능을 보인다. 한편, JSMA는 자코비안 기반 saliency map을 활용하여 영향력이 큰 소수의 특징만을 선택적으로 교란하는 기법이다. C&W 공격은 적대적 공격 데이터 생성을 최적화 문제로 정식화하여 최소한의 변화로 높은 공격 효과를 달성하는 기법이다.

이러한 방법들은 본 연구에서 제안하는 GAN 기반 적대적 공격 생성 모델의 성능을 평가하기 위한 기준 공격 방법으로 사용된다.

III. 적대적 공격 생성 모델 설계

본 연구에서는 Fig. 1에 나타낸 바와 같이, 강건한 IoT IDS을 위한 적대적 공격 생성-학습 프레임워크를 제시한다. 전체 프레임워크는 정상 데이터로 IDS를 학습한 이후, 기존 화이트박스 공격 기법과 GAN 기반 적대적 공격 생성 모델을 활용하여 적대적 공격 기법을 통해 생성된 공격 데이터를 활용하여 모델의 취약성을 분석하고, 이를 다시 학습 과정에 반영하는 구조로 구성된다.

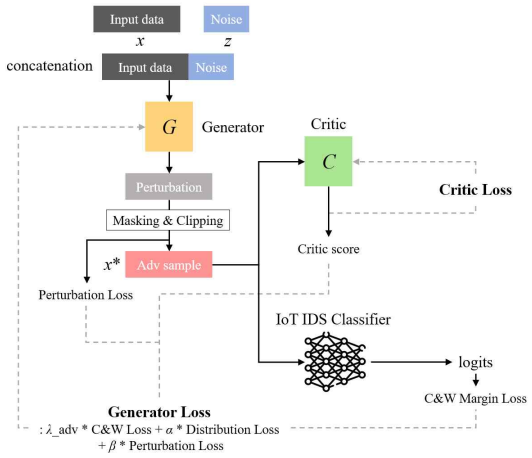


Fig. 2. GAN 기반 적대적 공격 생성 모델

Fig. 2와 같이, 제안하는 적대적 공격 생성 모델은 생성자(generator), 판별자(critic), 그리고 실제 데이터로 학습된 타깃 IDS로 구성된 GAN 기반 구조를 따른다. 생성자는 입력 데이터와 랜덤 노이즈를 결합하여 다양한 형태의 적대적 교란을 생성한다. 이를 통해 동일한 입력에서도 다양한 적대적 샘플을 생성해 공격의 다양성과 전이성을 향상시키고, 특정 패턴에 대한 과적합을 방지한다. 기존 공격 방식과 달리 도메인 특성을 유지하기 위해 이진 마스크를 적용해 범주형 또는 상관관계가 있는 특징의 변경을 제한한다 [4]. 또한 생성된 데이터가 클래스별 특징 범위 내에 있도록 제약을 두어 현실성을 유지한다.

생성된 적대적 공격 데이터는 판별자와 타깃 IDS에 동시에 입력되며, 타깃 IDS는 생성자가 오분류를 유도하는 방향으로 학습되도록 한다. 특히, 작은 교란에서도 안정적이고 효과적인 공격 생성을 위해 logit 기반의 C&W 마진 손실을 적용하여 생성된 샘플이 대상 모델에서 오분류되도록 유도하였다. 이는 학습의 수렴 속도와 안정성을 향상시키며, IoT 네트워크와 같이 이산적이고 상호 의존적인 특징을 갖는 데이터 환경에서 특히 효과적이다 [5]. 더불어, 분포 손실을 통해 생성 데이터와 실제 데이터 간의 분포 유사성을 유지하고, 교란 손실을 통해 교란의 크기를 제한하여 최소한의 변화로 현실적인 데이터 생성을 유도한다. 이 손실 함수들을 공동으로 최적화함으로써, 제안된 모델은 도메인 의미를 유지하면서도 IDS를 효과적으로 우회하는 현실적인 적대적 공격 데이터를 생성한다.

IV. 평가

실험 환경. IDS 성능 평가를 위해 NF-ToN-IoT-v3, Edge-IIoT, X-IIoTID, WUSTL-IIoT-2021의 네 가지 IoT 벤치마킹 데이터셋을 사용하였다. 전처리 후, 각 데이터셋에 대해 최적의 IDS 성능을 달성하기 위한 하이퍼파라미터 구성을 도출하기 위해 Optuna 기반 하이퍼파라미터 최적화를 수행하였다. 도출된 최적 하이퍼파라미터를 바탕으로 IDS 분류기를 구성하고, 전체 데이터셋의 56%를 학습 데이터로 사용하여 모델을 학습하였다. 학습된 IDS 모델은 기준 성능을 확보하기 위해 전체 데이터의 30%로 구성된 테스트셋에서 평가되었다 ①. 이후, FGSM, JSMA, PGD, BIM, C&W와 같은 기존 화이트박스 공격 기법과 제안된 GAN 기반 모델을 이용하여 각각 적대적 샘플을 생성하고, 이를 테스트셋의 20%에 해당하도록 구성하여 IDS의 성능을 평가하였다 ②. 마지막으로, 두 방식으로 생성된 추가 적대적 데이터를 활용하여 적대적 학습을 수행하고, 이에 따른 성능 변화를 비교·분석하였다 ③.

기존 대비 성능 비교. Fig. 3은 적대적 공격이 IDS에 의해 정상으로 오분류되는 비율, 즉 공격 성공률을 비교한 결과를 나타낸다. GAN 기반 적대적 공격은 평균 99.89%의 공격 성공률을 기록하여 기존 화이트박스 공격들의 평균 수치(FGSM 69.45%, JSMA 82.00%, PGD 92.85%, BIM 89.26%, C&W 70.65%)에 비해 가장 높은 수치를 보였다. Edge-IIoT 데이터셋에서는 99.74%에 도달하였고, 나머지 데이터셋에서도 90% 이상의 높은 수준을 유지하였다. 또한, Fig.4 하단 좌측의 혼동 행렬에서도 확인되듯이, 적대적 공격 적용 시 다수의 공격 데이터가 정상 레이블에 집중적으로 오분류되는 경향이 뚜렷하게 나타났다. 이러한 결과는 GAN 기반 공격이 데이터셋 전반에서 일관되게 높은 공격 성공률을 보인다는 점에서 공격 생성의 일반화 측면에서의 효과적인 접근 방식임을 시사한다.

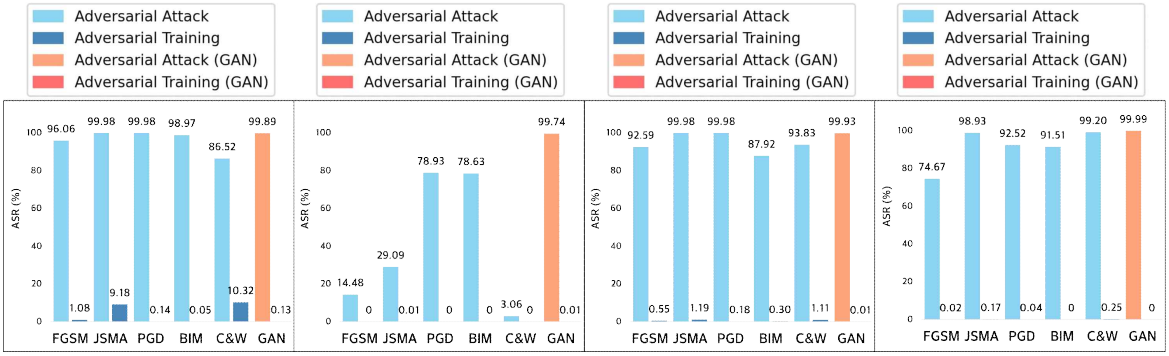
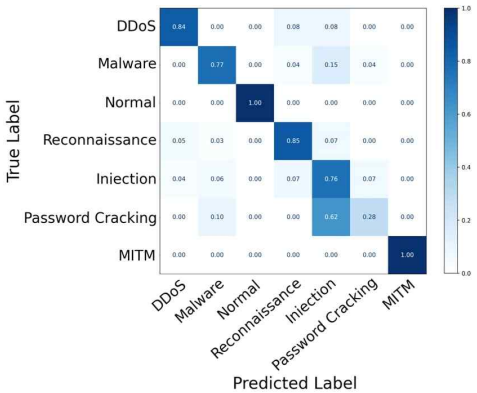
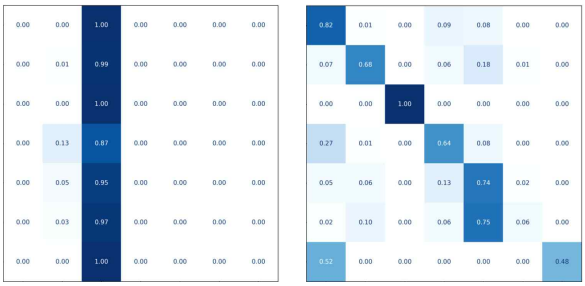


Fig 3. 공격 성공률 비교 (NF-ToN-IoT-v3, Edge-IIoT, X-IIoTID, WUSTL-IIoT-2021)



원본 데이터에 대한 분류 결과 ①



적대적 공격 후 ②　　적대적 학습 후 ③

Fig. 4. 혼동 행렬 (Edge-IIoT 데이터셋 사례)

강건성. 적대적 학습을 적용한 후에는 오분류율이 현저히 감소하여 대부분의 경우 10% 이하로 낮아졌으며, 일부 경우에는 0%에 근접한 수준을 보였다. Fig.4 하단 우측의 혼동 행렬에서도 적대적 공격 데이터가 원래의 레이블로 정확히 분류되는 비율이 크게 증가함을 확인할 수 있다. 이 결과는 GAN 기반 공격이 높은 공격 성능을 보이더라도, 적대적 학습을 통해 IDS의 강건성을 향상시킬 수 있음을 보여준다.

V. 결론

본 연구는 GAN 기반 적대적 공격 생성 모델로 생성된 데이터를 학습에 포함시켜 IoT IDS의 적대적 공격 강건성을 향상시켰다. 향후 연구에서는 보다 실제 환경에 가까운 공격 시나리오를 반영하기 위해 블랙박스 환경에서의 적대적 공격 및 방어 기법으로 확장할 계획이다.

[감사의 글]

이 연구는 2025년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(과제번호 RS-2025-02653102).

[참고문헌]

[1] K. He et al., “Adversarial machine learning for network intrusion detection systems: A comprehensive survey,” IEEE Communications Surveys Tutorials, Vol. 25, no. 1, 2023.

[2] X. Yuan et al., “A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system,” Computers Security, vol. 137, 2024.

[3] E. De˘girmenci et al., “Adversarial attack detection approach for intrusion detection systems,” IEEE Access, vol. 12, 2024.

[4] B.-E. Zolbayar et al., “Generating practical adversarial network traffic flows using nidsgan,” 2022.

[5] N. Carlini et al., “Towards evaluating the robustness of neural networks,” 2017.