Journal of KIIT. Vol. 22, No. 12, pp. 175-182, Dec. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.12.175

GPT를 활용한 MQTT 기반 IDS 우회 방법

이소원*[1], 김병찬*[2], 최선오**

MQTT-based IDS Evasion Method using GPT

Sowon Lee*[1], Byeongchan Kim*[2], and Sunoh Choi**

이 논문은 2024년도 교육부의 재원으로 한국 연구재단 기초연구사업 (RS-2023-00237159)과
지자체-대학 협력기반 지역혁신 사업 (2023RIS-008)의 지원을 받아 수행된 연구임

요 약

사물인터넷(IoT)은 경량 프로토콜인 MQTT(Message Queuing Telemetry Transport)를 통해 기기 간의 원활한 통신을 지원하며, 효율성과 확장성 등의 장점으로 인해 다양한 분야에서 활용되고 있다. 그러나 IoT가 확산되며 네트워크 보안 문제도 야기되고 있으며, 특히 MQTT 트래픽의 경량성과 비표준화된 특성, 그리고 공격 기법의 지속적인 진화로 인해 기존 침입 탐지 시스템(IDS)은 이러한 트래픽을 효과적으로 탐지하는 데 한계를 드러내고 있다. 본 연구에서는 IDS의 탐지 성능을 향상시키는데 활용할 수 있도록 GPT(Generative Pre-trained Transformers) 같은 생성 인공지능(AI)를 활용하여 IDS 탐지를 회피할 수 있는 MQTT 데이터를 생성하는 방법을 제안한다. 전처리된 정상 MQTT 데이터를 기반으로 GPT 모델을 학습시켜 데이터를 생성하고, 악성 MQTT 데이터를 merge 함으로써 IDS 탐지를 우회할 수 있는 데이터를 생성하였다. 생성된 데이터는 IDS의 취약점을 분석하고, 탐지 성능 개선을 위한 학습 데이터 강화 및 모델 설계에 활용될 것을 기대한다.

Abstract

The Internet of Things(IoT) supports seamless communication between devices through the lightweight protocol Message Queuing Telemetry Transport(MQTT), which is widely applied across various fields due to its advantages in efficiency and scalability. However, the rapid expansion of IoT has raised significant network security concerns. In particular, the lightweight and non-standardized characteristics of MQTT traffic, coupled with the continuous evolution of attack techniques, have exposed the limitations of existing Intrusion Detection Systems(IDS) in effectively detecting such traffic. This study proposes a method for generating MQTT data capable of evading IDS detection by leveraging Generative AI, specifically Generative Pre-trained Transformers(GPT), to enhance IDS detection performance. By training a GPT model on preprocessed normal MQTT data, malicious MQTT data were merged and synthesized to create datasets capable of bypassing IDS detection. The generated data are expected to facilitate the analysis of IDS vulnerabilities and contribute to improving detection performance through the enhancement of training data and model design.

Keywords
AI, GPT, MQTT, IDS, evasion

* 전북대학교 소프트웨어공학과 학사과정
- ORCID[1]: https://orcid.org/0009-0008-7215-7070
- ORCID[2]: https://orcid.org/0009-0007-2254-0111
** 전북대학교 소프트웨어공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0654-7109

· Received: Nov. 30, 2024, Revised: Dec. 13, 2024, Accepted: Dec. 16, 2024
· Corresponding Author: Sunoh Choi
Dept. of Software Engineering Jeonbuk National University Korea
Tel.: +82-63-270-4784, Email: suno7@jbnu.ac.kr

Ⅰ. 서 론

사물인터넷(IoT, Internet of Things) 사용량은 점점 증가하고 있다. 보안 칼럼에 따르면 IoT 증가 추이로 판단하였을 때 2025년에는 1조 대 이상의 IoT 기기가 사용될 것으로 예상된다[1].

IoT 기기 사용량이 증가함에 따라 보안 위협 또한 증가하고 있다. 그러나 단순히 IoT 기기의 증가에 따른 보안 위협이 증가한다고 볼 수 없다. IoT의 특성상 주기적인 업데이트, 비밀번호 변경 등 보안 조치가 제대로 이루어지지 않아 공격자로부터의 공격에 노출되어 있다[2]. 또한 IoT는 다양한 물리적 구성요소(장치, 센서, 게이트웨이 등)와 통신 기술, 서비스 API 기술, 사용자 인터페이스 기술 등 요소들의 결합 지점에서 취약점이 발생하고 있다[3].

IoT는 일반 컴퓨터에 비해 컴퓨팅 성능이 낮아 경량화 프로토콜을 사용한다. 시스템 아키텍쳐 계층에 따라 필요한 IoT 프로토콜 유형이 사용되고 특히 애플리케이션 계층에서는 MQTT(Message Queuing Telemetry transport, AMQP(Advanced Message Queuing Protocol), CoAP(Constrained Application Protocol) 등 프로토콜이 사용된다. 이 중 MQTT는 효율성과 확장성으로 인해 IoT 환경에서 가장 널리 사용되는 프로토콜이다. MQTT는 머신 대 머신 통신에 사용되는 표준 기반 메시징 프로토콜이다. MQTT는 발행/구독 패턴을 사용하여 메시지 발신자(발행자)와 메시지 수신자(구독자)를 분리하고, 메시지 브로커라는 제3의 구성요소를 통해 게시자와 구독자 간의 통신을 처리한다.

MQTT는 제한된 리소스를 가진 환경에서도 효율적인 통신이 가능하다. 하지만 경량성을 유지하며 보안을 최소화하는 통신 환경을 제공하기에 약점이 존재한다. 이러한 문제에 대응하기 위해 기존 IoT 침해에 관련된 기존 보안 연구들은 효율적이거나 정교하게 보안 위협을 탐지하고 예방하기 위한 연구가 진행되고 있다. 슬라이딩 윈도우 기법을 적용한 LSTM 기반 딥러닝 모델에 MQTT 트래픽의 시간적 패턴을 학습하여 악성 트래픽 탐지 정확도를 향상하거나[4] IoT 환경에서 발생하는 악성 코드 탐지를 위해 IoT 기기의 제한된 리소스와 다양한 아키텍처를 고려한 저비용 패킷 기반 탐지 시스템을 설계하는 연구가 진행되었다[5]. 또한 샌즈랩은 IoT 기기의 리소스 한계를 극복하기 위한 기술 개발 및 특허를 획득하였다[6]. 이를 통해 공격을 실시간으로 탐지하고, 탐지 성능을 높이거나 솔루션을 제공하여 효율적인 탐지 방안을 제안하는 연구가 대부분임을 확인할 수 있다.

따라서 본 연구에서는 악성 MQTT 프로토콜에 대한 보안 위협을 탐지하는 블랙박스 모델을 회피하는 연구를 진행하였다. 블랙박스 모델을 회피하기 위한 데이터 생성을 위해 전처리 된 MQTT 데이터를 학습한 3가지 생성형 AI(Artificial Intelligence) GAN(Generative Adversarial Networks)[7], GPT-2(Generative Pre-trained Transformer-2)[8]와 GPT-4o (Generative Pre-trained Transformer-Omni)[9]를 기반으로 생성된 데이터를 통해 실제로는 악성 데이터의 특성을 갖지만, 정상으로 판별되는 데이터를 생성하는 방법을 제안 및 평가하였다. GAN은 딥러닝 모델의 탐지를 회피하기 위한 적대적 공격을 수행하기 위하여 제안되었다. 그리고 트랜스포머 모델을 기반으로 GPT-2가 생성형 AI 모델로 제안되었고 GPT-4o는 자연어 생성뿐만 아니라 멀티모달을 위한 생성형 AI 모델로 제안되었다.

생성형 AI를 학습시키기 위한 데이터셋으로는 Kaggle에서 제공하는 MQTT 프로토콜 데이터셋인 MQTTset을 기반으로 S. Choi et al.의 논문에서 악성 공격의 특성에 따라 전처리한 MQTT 패킷 데이터를 사용하여 학습하였다[10]. 생성된 데이터의 정교함과 블랙박스의 탐지율을 평가하고 생성형 AI 간의 성능 비교를 위해 LSTM(Long Short-Term Memory) 기반의 Black Box 모델을 구축 및 TPR(True Positive Rate) 성능 평가 지표를 활용하여 생성된 데이터가 블랙박스 모델을 속이는지에 대해 평가하고, 생성 AI 모델의 성능을 비교하였다.

이를 통해 생성형 AI로 MQTT 데이터를 생성하고 악의적인 활동을 탐지 및 모니터링하는 도구인 IDS(Intrusion Detection System, 침입 탐지 시스템) 우회 방법을 제시하였고 LSTM 기반 Black Box로 판별하는 실험을 통해 생성형 AI인 GPT의 MQTT 데이터 생성 성능이 우수함을 입증하였다.

본 논문의 구성은 다음과 같다. 2장은 AI를 활용한 악용 사례와 기존 연구에서 진행된 IDS 탐지 연

구 사례를 소개하고 3장에서는 본 연구 배경 및 데이터 생성 모델인 GPT와 탐지 모델에 쓰인 LSTM 소개한다.

이후 AI 모델별 데이터 생성 과정을 설명하고 생성된 데이터를 LSTM 기반 블랙박스 모델로 판별하여 성능을 평가하였다.

II. 관련 연구

2.1 AI 악용 사례

최근 러시아-우크라이나 전쟁에서 딥페이크와 같은 AI 기술이 정보 전쟁 도구로 악용되고 있다. 우크라이나는 러시아 국민들에게 전쟁의 현실을 알리기 위해 딥페이크와 사이버 공격을 결합했으며, 러시아는 이를 왜곡해 우크라이나 피해 이미지를 조작된 것으로 주장하며 선전에 활용하였다. 이러한 AI 기술의 남용은 양측에서 여론 혼란과 정보 조작을 확대하는 데 기여하고 있다[11].

딥페이크를 이용한 금융 및 신뢰 사기가 급증하고 있다. 딥페이크로 유명인의 얼굴과 목소리를 합성해 가짜 가상 자산 투자 홍보 영상을 제작한 사례가 보고 되고 있다. 이 수법은 투자자를 유인하여 큰 금액을 투자하게 하고, 가짜 자산을 다른 코인으로 환전한 뒤 수익을 챙기는 방식이다. 이로 인해 피해자들은 금전적인 손해를 입고 있다[12][13].

자동차 업계에서는 AI 기술을 활용한 사이버 공격 사례가 우려되고 있다. 특히 자율주행차의 인공지능 시스템을 오작동 시키거나, 해커들이 차량 시스템을 장악하여 운전자와 승객의 안전을 위협할 수 있는 위험성이 제기되고 있다[14].

2.2 효율적인 IDS 탐지 방안 연구

악성 트래픽을 탐지하기 위한 IDS의 연구가 진행되고 있다. B. Isong et al.은 IoT 환경에서 IDS의 탐지율을 높이기 위해 시그니처 기반과 이상 탐지 기반을 결합한 모델을 제안하였고, 95% 이상의 정확도와 낮은 오탐률을 달성하였다. 비지도 학습 기반 IDS는 네트워크 변화에 적응하며 제로데이 공격을 90% 이상의 정확도로 탐지했다. 블록 체인을 IDS에 통합하여 데이터 무결성을 보장하며 신뢰성과 탐지 성능을 향상시켰다[15].

N. Butt et al.은 스마트 홈 네트워크에서 기존 모델이 과적합 문제를 겪는 것을 해결하기 위해 하이브리드 딥러닝 및 머신러닝 모델을 제안하였다. 데이터 정규화와 하이퍼파라미터 튜닝을 통해 탐지 성능을 향상하는 데 초점을 맞췄으며, 특히 실시간 데이터셋을 활용하여 현실적인 탐지 성능을 입증했다[16].

2.3 탐지 모델 우회 관련 연구

T. Fladby et al.은 IDS의 탐지율을 낮추기 위해 네트워크 흐름 크기, 연결 시간, 주기성 등의 매개변수를 변경하는 기법을 연구하였다. 이를 통해 악성 패킷의 본질을 유지하면서도 탐지를 회피할 수 있음을 실험적으로 입증했으며, 이러한 매개변수 변경은 Stochastic Optimization 기법을 활용하여 최적화되었으며, 탐지율을 최대 20% 감소시켰다[17].

S. Choi et al.은 GAN을 활용하여 기존 AI 기반 악성 PowerShell 탐지 시스템의 탐지율을 낮추는 적대적 공격 기법을 연구하였다. 정상 데이터와 유사한 가짜 PowerShell 데이터를 생성하며, 생성 과정에서 원래 악성 데이터의 본질을 유지하면서도 탐지를 회피할 수 있음을 입증하였다. GAN의 추가 토큰 길이(L) 조정을 통해 탐지율 감소 효과를 확인하였다. 추가된 토큰 길이에 대해 L=4로 설정한 경우 탐지율이 0%로 떨어지는 결과를 얻었고, 이를 통해 GAN이 정상 데이터와 유사한 가짜 데이터를 생성하며 기존 탐지 시스템을 우회할 수 있음을 입증하였다[18].

III. IDS 탐지 성능 향상을 위한 MQTT 데이터 생성형 AI 기법 및 비교 분석

3.1 연구 배경 및 방법

기존 IDS는 제한된 데이터셋으로 인해 다양한 MQTT 트래픽 패턴을 충분히 학습하지 못하며, 이는 IDS의 한계를 초래한다. 따라서 본 논문에서는 생성형 AI 기법인 GAN과 GPT-2, GPT-4o를 활용하여

MQTT 데이터를 생성함으로써, IDS 학습 데이터의 다양성을 확보하여 탐지 성능을 향상시키고자 한다. 이러한 접근은 IDS의 일반화 능력을 강화한다.

이를 평가하기 위해 LSTM 기반의 Black Box 모델을 이용하여 판별한다. 평가 지표로는 판별하고자 하는 데이터 중 실제 올바르게 판별한 비율을 나타내는 분류 성능 지표인 TPR(True Positive Rate)에 따라 성능을 비교한다.

3.2 LSTM

LSTM(Long Short-Term Memory)은 시퀀스 데이터의 장기 의존성을 학습할 수 있는 RNN(Recurrent Neural Network, 순환 신경망)의 변형으로 기억 셀과 게이트 구조를 통해 장기 정보와 단기 정보를 효과적으로 조합하는 특징을 가진다. 본 연구에서는 LSTM 기반의 Black Box 모델을 IDS로 활용하여 정상 데이터와 악성 데이터를 학습시킨 후, GAN과 GPT가 생성한 MQTT 데이터를 IDS 관점에서 판별하였다. 모델이 판별한 결과는 TPR을 사용하여 공격을 탐지한 비율로 나타낸다. TPR은 수식(1)과 같이 정의되며 모델이 positive라고 예측했는데 실제 정답이 positive인 경우(True Positive, TP)와 모델이 negative라고 예측했는데 실제 정답이 positive인 경우(False Negative, FN) 중 실제 정답이 positive(TP)인 비율이다. 이는 올바르게 탐지한 경우의 수를 전체의 경우의 수로 나누어 계산한다. 따라서 생성된 데이터가 Black Box 모델에 탐지되지 않았는지, 즉 IDS의 탐지를 우회하였는지 평가할 수 있다.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

3.3 GPT

해당 논문에서는 생성형 AI 기법인 GPT(Generative Pre-trained Transformer) 모델을 활용하여 MQTT 데이터 학습 및 생성한다. GPT는 사전 학습된 변환기 기반 언어 모델로 트랜스포머 모델의 디코더 구조를 기반으로 한다. 텍스트와 시퀀스 패턴을 학습하며 Self-Attention 메커니즘을 통해 입력 시퀀스의 각 요소 간의 관계를 이해하여 자연스러운 시퀀스 데이터를 생성한다[7][8]. 본 연구에서는 기존 데이터의 시퀀스 특성을 유지하는 데이터를 생성하기 위해 GPT 모델에 MQTT 시퀀스 데이터를 학습한 후 데이터를 생성한다. GPT-2와 GPT-4o 두 가지 모델을 각각 활용하여 MQTT 데이터를 생성하였다.

GPT-2(Generative Pre-trained Transformer 2)는 Open AI에서 개발한 대규모 Transformer 기반의 언어 모델로, 사전 학습된 거대한 파라미터(최대 15억 개)와 40GB 이상의 방대한 데이터셋을 활용하여 문맥 이해 및 텍스트를 생성한다. GPT-2는 Transformer의 Decoder를 활용하는 이전 GPT-1의 구조와 거의 유사하나, 모델 크기와 데이터 크기를 크게 확장하여 자연어 처리에서 뛰어난 성능을 보인다. 특히, GPT-2는 GPT-1에 비해 제로샷 학습을 기준으로 많은 자연어 처리 작업에서 10~30% 이상의 성능 향상을 보였다.

GPT-2는 OpenAI에서 공개한 GPT-2의 모델을 기반으로 파인튜닝을 통해 MQTT 데이터에 최적화되도록 수정하였다. 이를 통해 모델은 MQTT 데이터의 특성을 학습하며 해당 도메인에 적합한 데이터 생성 및 문맥적 이해를 수행할 수 있도록 설계하였다.

GPT4o(Generative Pre-trained Transformer 4 Omni)는 GPT-4 터보 모델을 기반으로 구축된 현재 기준 OpenAI의 가장 최신 모델이다. GPT-4o는 다양한 벤치마크 테스트에서 우수한 성능을 보였으며, GPQA에서 53.6%, MGSM에서 90.5%의 높은 성능을 기록했다. GPT-4o는 기존 모델보다 텍스트와 이미지 모두 비영어권 언어에서도 GPT-4보다 더 높은 성능을 보였으며, 비전 기능을 갖춘 GPT-4 Turbo보다 비전 이해, 처리 및 분석 성능이 높음을 보인다[9].

GPT-4o는 누구나 접근에 용이한 ChatGPT에서 제공하는 GPT-4o를 통해 MQTT 데이터를 생성하였다. 데이터를 생성하기 위해 MQTTset 데이터에 전처리 과정만 거친 원본 정상 데이터(normal data)와 데이터셋 구조에 대한 특성을 알려주었다.

3.4 GPT를 이용한 IDS 우회 방법

본 연구에서는 normal data를 GPT 모델에 학습하

여 AI가 생성한 데이터(fake data)와 MQTTset 데이터에 전처리 과정만 거친 원본 악성 데이터(mal data)를 비율(Ratio)에 따라 merge하여 merged data를 생성한다. Black Box IDS의 탐지에 혼란을 주어 악성 데이터의 특성을 분류하지 못하도록 하는 merged data 생성 방안을 제안한다.

merge 방식은 데이터의 시퀀스적 특성을 고려하여 설계하였다. Ratio에 따라 두 데이터를 균등하게 분배하고, 남은 데이터를 분산시켜 데이터들이 최대한 혼합되도록 하였다. 또한, 데이터를 하나의 패킷 단위를 고려한 5개 단위로 그룹화하여 병합을 진행하였다.

MQTTset 데이터에 전처리 과정만 거친 정상 데이터를 GPT 모델에 학습하고 제안한 merge 방식을 거쳐 생성된 IDS 우회를 위한 merged data가 IDS인 LSTM Black Box에서 판별 후 결과에 대한 과정은 그림 1과 같다.

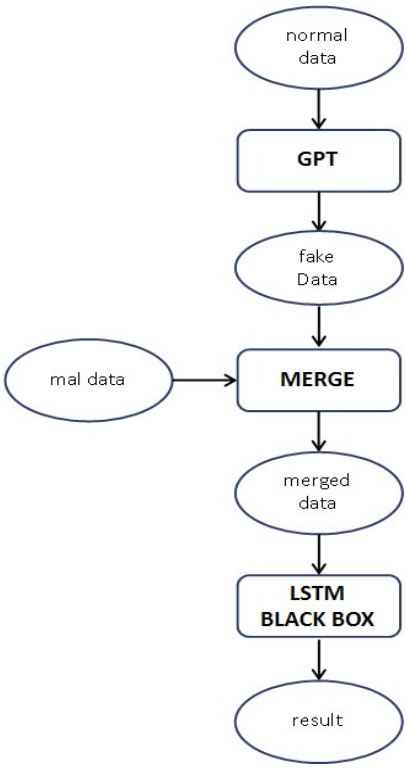


그림 1. GPT 모델 아키텍처
Fig. 1. GPT model architecture

3.5 실험 결과

3.5.1 데이터 및 데이터 전처리

본 연구에서 생성형 AI 모델을 학습시키기 위한 데이터로 MQTTset 데이터를 악성의 특성을 고려하여 이전 연구에서 제안한 전처리한 데이터를 사용하였다. 해당 연구에서는 5가지 악성 공격 기법(DoS Attack, Flood Attack, SlowITe Attack, Malformed Attack, Brute-Force Attack) 특성을 고려한 전처리 방안을 제안한다[10].

정상 데이터, 악성 데이터는 각 1,000개씩 구성되어있다. 전처리 된 데이터는 한 행에 1개, 공백을 기준으로 80개로 나뉘어 각각마다 특징을 갖는 데이터이다. 80개의 메시지 중 5개마다 하나의 IP에서 전송된 데이터를 의미하며, 이는 한 데이터에 총 16개의 IP에서 전송된 데이터를 의미한다.

AI로 fake data를 생성하고 이후 합치는 과정에서도 데이터의 특징을 고려하여, 16개의 데이터 그룹을 유지하였다. mal data를 합치는 Ratio을 0~15로 다르게 하여 각 Ratio에 따라 merged data를 생성하였다. Ratio 값에 따른 fake data와 mal data의 Ratio는 표 1과 같다.

표 1. Ratio에 따른 데이터 개수
Table 1. Number of data by ratio

Ratio	Fake data	Mal data
0	0	16
1	1	15
2	2	14
...	...	...
13	13	3
14	14	2
15	15	1

3.5.2 GPT 모델 및 학습

코드가 공개되어있는 GPT-2와 달리, GPT-4o는 현재 API만 제공한다. 따라서 두 버전에 따른 학습 방법을 달리하여 데이터를 생성하였다. GPT-2는 모델 학습을 위해 NVIDIA GeForce RTX 3050, Windows 11 64-bit OS에서 Python 3.12, Pytorch 3.12

및 Hugging Face의 transformers 라이브러리, Trainer, TrainingArguments를 이용하였다.

IDS를 우회하는 fake data를 생성하기 위해 파인 튜닝을 기반으로 GPT-2 모델을 수정 후 normal data를 학습하였다. 기존의 Encoder Decoder Tokenizer는 sub-tokenizer 방식을 사용한다. 이 인코딩 방식은 -1의 경우 '-'와 1을 분리하여 토큰화함으로 MQTT 데이터에 적용하기 적합하지 않다.

따라서 공백을 기준으로 숫자 시퀀스를 고유한 토큰으로 간주하도록 토크나이징을 하였다. 모델 학습 전 숫자를 토큰 ID로 매핑하고, 이후 생성된 데이터의 토큰 ID를 역매핑하여 원래 숫자 시퀀스로 복원하였다.

생성 데이터의 길이는 80으로 설정하여 생성되는 데이터가 80개가 원본 데이터와 동일한 80개가 되도록 하였다. prompt에 생성되는 데이터가 "1001 0 1 -1 -1"과 같이 시작하도록 값을 주어 생성된 시퀀스를 디코딩하여 가상 데이터를 CSV 형식의 파일로 저장하였다. fake 데이터 생성 시 Top-k. Top-p, temperature을 적용하여 생성되는 데이터의 다양성과 일관성을 조정하였다.

GPT-4o는 chatGPT를 활용하여 명령 프롬프트를 통해 작업을 지시하고 데이터를 생성하였다. fake data를 생성하기 위해 normal data와 하나의 데이터 길이가 80임을 입력하고, 생성되어야 하는 데이터의 시작 값의 형식, 80개의 데이터의 구성과 같이 데이터의 특성을 설명한 후 최종적으로 CSV 형식으로 데이터를 생성하도록 하였다.

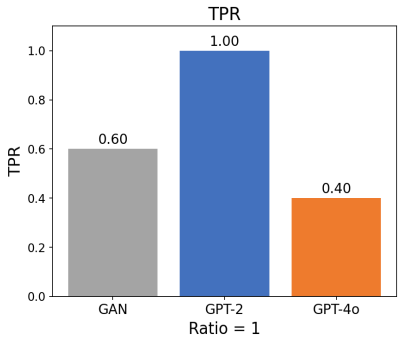
3.5.3 실험 결과

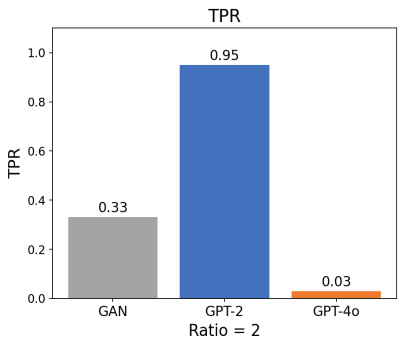
GPT를 통해 만들어진 데이터를 비교하기 위하여 적대적 생성 모델인 GAN을 이용해 데이터를 생성하였다. 생성한 데이터들을 LSTM 기반 Black Box로 탐지 회피율을 확인해 본 결과 GPT-4o, GAN, GPT-2 순으로 회피율이 높음을 확인하였다. fake data가 섞이는 비율인 Ratio가 0, 1, 2, 4, 8, 12일 때 탐지율은 표 2와 그림 2와 같다. 각 탐지율에 대한 값은 소수 넷째 자리에서 반올림하였다.

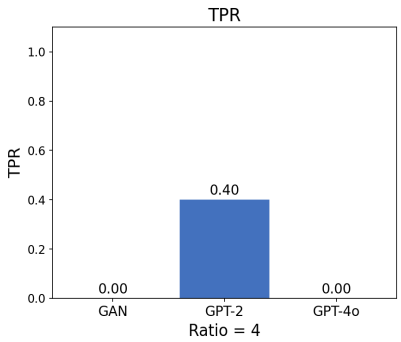
Ratio가 0일 때는 mal data만 있는 경우이며 100% (탐지율 1.0) 탐지되는 것을 통해 LSTM 기반 Black Box 모델의 성능을 검증하였다. Ratio가 1일 때는 GAN과 GPT-4o 두 모델은 각각 60%와 40% 탐지되었고, Ratio = 2부터 GPT-4o는 0.03%로 이전 대비 낮은 탐지율을 보였다. Ratio = 4부터는 GAN과 GPT-4o 두 모델 모두 100% 회피 (탐지율 0.0)하며 GPT-2는 40% 탐지되었다.

표 2. 각 모델의 TPR
Table 2. TPR for each model

Ratio	GAN	GPT-2	GPT-4o
0	1.0	1.0	1.0
1	0.600	1.0	0.400
2	0.330	0.950	0.030
4	0.000	0.400	0.0
8	0.010	0.190	0.0
12	0.000	0.100	0.0







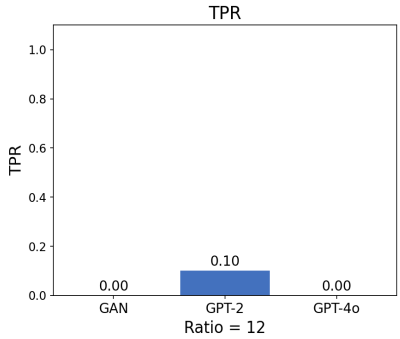


그림 2. 각 모델의 TPR
Fig 2. TPR for each model

Ⅳ. 결론 및 향후 과제

본 연구에서는 IoT에 사용되는 대표적인 프로토콜인 MQTT를 통해 악성 탐지 Black Box 모델의 탐지를 회피하는 MQTT 데이터 생성 방안을 제안하였다. 악성 데이터의 특성을 고려하여 전처리한 데이터를 모델에 학습시켰으며, GAN, GPT-2 그리고 GPT-4o 생성형 AI에 학습 및 실제 데이터와 유사한 fake data를 생성하였다. 각 딥러닝 모델에 대하여 생성된 fake data에 대해 LSTM 기반 블랙박스로 탐지하였고, TPR을 통해 각 생성된 fake data를 평가하였다.

GPT-2 모델의 경우는 Ratio = 4부터 40% 탐지율을 보였고 그 이후로 점점 탐지율이 낮아지는 추이를 보였다. GPT-4o 모델의 경우, Ratio = 2부터 3%의 낮은 탐지율을 보였고, 이후로는 완전히 탐지를 회피하였다.

본 논문에서 제안한 방안으로 생성된 데이터는 이후 IDS의 탐지율 성능 향상을 위한 정밀한 데이터로 활용될 수 있을 것을 기대한다.

그러나 chatGPT에서 제공하는 GPT-4o를 제외한 GAN과 GPT-2의 학습 및 데이터 생성 속도를 고려하였을 때 GPT-2의 속도가 현저히 느렸다는 점에서 코드 개선이 필요하다는 점과 GPT-4o를 통한 데이터 생성 환경이 다르다는 점에서 본 연구의 한계를 찾을 수 있다. 따라서 이러한 한계를 극복하기 위한 코드 최적화 연구 및 GPT-4o 데이터 생성에 관하여 chatGPT가 아닌 API를 활용한 데이터 생성 연구 방안을 향후 과제로 수행할 수 있을 것이다.

또한, 한편으로 Deep Fake와 같은 적대적 공격을 막기 위한 연구들도 진행되고 있는데, 우리의 연구 또한 악성 트래픽을 탐지하기 위한 딥러닝 모델에 대한 적대적 공격 수단으로 사용될 수 있다. 따라서, 악성 트래픽 탐지연구에서 생성형 AI를 위한 적대적 공격을 방지하거나 완화하는 연구를 향후 연구로 수행하려고 한다.

References

[1] IoT vulnerability, https://www.etnews.com/20220913000181 [accessed: Nov. 28, 2024]

[2] IoT security risks, https://www.boancloud.co.kr/security-issues/49/ [accessed: Nov. 28, 2024]

[3] Y. H. Jeon, "A Study on the Security Modeling of Internet of Things (IoT)", Journal of KIIT, Vol. 15, No. 12, pp. 17-27, Dec. 2017. https://doi.org/10.14801/jkiit.2017.15.12.17.

[4] D. Lee, S. Im, and S. Choi, "Malicious traffic detection method using LSTM and sliding window in MQTT-based IoT environment," Journal of KIIT, Vol. 21, No. 5, pp. 111-120, May. 2023. https://doi.org/10.14801/jkiit.2023.21.5.111.

[5] S. Y. Shin, D. H. Lee, and S. J. Lee, "Design Method of Things Malware Detection System(TMDS)", Journal of Information Security Studies, Vol. 33, No. 3, pp. 459-469, Jun. 2023. https://doi.org/10.13089/JKIISC.2023.33.3.459.

[6] SandsLab technology exports, https://www.sandslab.io/bbs/board.php?bo_table=m06_01&wr_id=75&page=1 [accessed: Nov. 28, 2024]

[7] I. J. Goodfellow, 나J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengid, "Generative Adversarial Networks", Machine Leaning, Jun. 2014. https://doi.org/10.48550/arXiv.1406.2661.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners", OpenAI Technical Report, Vol. 1, No. 8, Feb. 2019.

[9] OpenAI, "GPT-4 introduction", https://openai.com/index/hello-gpt-4o [accessed: Nov. 28, 2024]

[10] S. Choi and J. H. Cho, "Novel Feature Extraction Method for Detecting Malicious MQTT Traffic Using Seq2Seq", Applied Sciences, Vol. 12, No. 23, pp. 12306, Dec. 2022. https://doi.org/10.3390/app122312306.

[11] Deepfakes in Russo-Ukrainian information warfare, https://theconversation.com/les-deepfakes-au-coeur-de-la-guerre-informationnelle-russo-ukrainienne-240945 [accessed: Nov. 28, 2024]

[12] Security threats caused by AI, including deepfake crimes, https://m.boannews.com/html/detail.html?idx=132580 [accessed: Nov. 28, 2024]

[13] Rising AI hacking threats: Criminal AIs in 2024, https://www.edaily.co.kr/news/read?newsId=03011046635835568 [accessed: Nov. 28, 2024]

[14] Cybersecurity threats in 2024: Hacktivism and AI, https://www.segye.com/newsView/20240123514260 [accessed: Nov. 28, 2024]

[15] B. Isong, O. Kgote, and A. Abu-Mahfouz, "Insights into Modern Intrusion Detection Strategies for Internet of Things Ecosystems", Electronics, Vol. 13, No. 12, pp. 2370, Jun. 2024. https://doi.org/10.3390/electronics13122370.

[16] N. Butt and A. Shahid, "Intelligent Deep Learning for Anomaly-Based Intrusion Detection in IoT Smart Home Networks", Mathematics, Vol. 10, No. 23, pp. 4598, Dec. 2022. https://doi.org/10.3390/math10234598.

[17] T. Fladby, H. Haugerud, S. Nichele, K. Begnum, and A. Yazidi, "Evading a Machine Learning-based Intrusion Detection System through Adversarial Perturbations", Proc. of the International Conference on Research in Adaptive and Convergent Systems, Gwangju, Korea, pp. 161-166, Oct. 2020. https://doi.org/10.1145/3400286.3418252.

[18] S. O. Choi, "Malicious PowerShell Detection using Attention against Adversarial Attacks", Electronics, Vol. 9, No. 11, pp. 1817, Nov. 2020. https://doi.org/10.3390/electronics9111817.

저자소개

이 소 원 (Sowon Lee)

2020년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 학사과정
2023년 1월 ~ 현재 : 전북대학교 지능형보안연구실 연구원
관심분야 : 지능형보안, 인공지능, 데이터 분석

김 병 찬 (Byeongchan Kim)

2020년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 학사과정
2023년 3월 ~현재 : 전북대학교 지능형보안연구실 연구원
관심분야 : 지능형보안, 인공지능

최 선 오 (Sunoh Choi)

2005년 2월 : 고려대학교 컴퓨터학과 (이학사)
2014년 5월 : Purdue Univ. 컴퓨터공학부 (공학박사)
2014년 8월 ~ 2019년 2월 : ETRI 정보보호연구본부 선임연구원
2021년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 부교수
관심분야 : 자동차보안, 원자력보안, 지능형보안