Flow-to-Text 기반 IoT 침입 탐지를 위한 텍스트 표현 방식 비교 연구*

김유담[01], 안석현[2], 박수현[3], 최희수[1], 조성제[1]
[1]단국대학교 소프트웨어학과, [2]단국대학교 인공지능융합학과 [3]단국대학교 컴퓨터학과
{kyd1012kr, seokhyun, parksh, cv3686, sjcho}@dankook.ac.kr

A Comparative Study of Text Representations for Flow-to-Text-Based IoT Intrusion Detection

Yudam Kim[01], Seokhyun Ann[2], Suhyeon Park[3], Heesu Choi[1], Seong-je Cho[1]
[1]Dept. of Software Science, [2]Dept. of AI-based Convergence, [3]Dept. of Computer Science, Dankook University
{kyd1012kr, seokhyun, parksh, cv3686, sjcho}@dankook.ac.kr

요 약

본 논문은 IoT 네트워크 플로우 데이터를 사전학습 언어모델에 적용하기 위한 Flow-to-Text 변환 방식에서, 텍스트 표현 구조가 침입 탐지 성능에 미치는 영향을 분석한다. 이를 위해 IoT-23 데이터셋의 Zeek conn.log 기반 플로우 특징 12개를 사용하였으며, 동일한 플로우 데이터를 자연어 서술형(T1), 키=값 구조형(T2), 확장 자연어형(T3)의 세 가지 텍스트 템플릿으로 변환하였다. 변환된 텍스트는 RoBERTa 모델에 입력하여 이진 분류와 다중 분류 성능을 평가하였고, Random Forest, XGBoost, MLP와 비교하였다. 실험 결과, RoBERTa-T2는 이진 분류에서 Macro F1-score 90.54%로 전체 모델 중 가장 높은 성능을 보였으며, 다중 분류에서도 94.37%의 Macro F1-score를 기록하여 Random Forest의 94.71%와 유사한 수준의 성능을 달성하였다. 특히 T2는 세 가지 텍스트 표현 방식 중 이진 및 다중 분류 모두에서 가장 높은 성능을 보여, 네트워크 플로우 정보를 단순한 자연어 문장으로 변환하기보다 특징 명칭과 값을 명시적으로 보존하는 구조형 표현이 언어모델의 입력 정보 인식에 효과적임을 확인하였다.

1. 서 론

IoT 기기의 확산으로 다양한 네트워크 기반 공격이 증가하면서, 비정상 트래픽을 탐지하기 위한 침입 탐지 시스템(Intrusion Detection System, IDS)의 중요성이 커지고 있다. 기존 IDS 연구에서는 네트워크 플로우에서 추출한 수치형 특징을 기반으로 Random Forest, XGBoost, MLP 등 다양한 머신러닝 및 딥러닝 모델을 적용해 왔다[1, 2]. 이러한 모델들은 높은 탐지 성능을 보이지만, 입력 특징 설계에 크게 의존하며 새로운 공격 패턴에 대한 일반화 측면에서 한계가 존재한다.

최근에는 보안 로그나 네트워크 플로우와 같은 구조적 데이터를 텍스트로 변환한 뒤, 사전학습 언어모델(Pre-trained Language Model, PLM)에 입력하는 연구가 증가하고 있다. 언어모델은 입력 토큰 간의 문맥적 관계를 학습할 수 있으므로, 네트워크 플로우를 텍스트 시퀀스로 표현하면 침입 탐지에도 활용될 수 있다. 예를 들어 LogBERT는 시스템 로그를 BERT 기반 모델에 입력하여 이상 탐지를 수행하였으며[3], Mehavilla 등은 Zeek 플로우 데이터를 대형 언어모델에 적용하고 ML/DL 기반 모델과 성능을 비교하였다[4].

그러나 기존 연구는 주로 언어모델의 적용 가능성이나 모델 간 성능 비교에 초점을 두었으며, 동일한 네트워크 플로우 정보를 어떤 텍스트 구조로 표현하는 것이 효과적인지에 대한 분석은 상대적으로 부족하다. 이에 본 논문은 IoT-23 데이터셋을 대상으로 네트워크 플로우 데이터를 세 가지 텍스트 표현 방식으로 변환하고, RoBERTa[5] 기반 침입 탐지 성능을 비교한다. 본 논문의 목적은 Flow-to-Text 기반 침입 탐지에서 텍스트 표현 구조가 모델 성능에 미치는 영향을 실험적으로 분석하는 것이다.

2. 관련 연구

네트워크 침입 탐지 분야에서는 다양한 머신러닝 및 딥러닝 기반 접근법이 제안되었다[1, 2]. Santhosh Kumar 등[1]은 IoT 환경에서 Random Forest, SVM 등 다양한 머신러닝 기법을 적용한 IDS 연구를 체계적으로 정리하였으며, Chen 등[2]은 고급 지속 위협(APT) 환경에서 ML 기반 IoT 보안의 한계와 이슈를 분석하였다.

구조적 데이터를 텍스트로 변환하여 언어모델에 적용하는 연구도 증가하고 있다. Guo 등[3]은 로그 데이터를 BERT 입력 형식으로 변환한 LogBERT를 제안하여 이상 탐지 성능을 향상시켰다.

네트워크 플로우 데이터에 대한 텍스트 기반 접근으로 Mehavilla 등[4]은 Zeek 플로우 데이터를 직렬화하여 GPT-2, LLaMA 등 대형 언어모델에 적용하고 ML/DL 모델과 성능을 비교하였다.

본 논문은 세 가지 텍스트 표현 방식으로 변환하는 Flow-to-Text 프레임워크를 설계하고, RoBERTa 모델을 fine-tuning하여 표현 방식에 따른 침입 탐지 성능 차이를 이진 분

* 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학석사연계ICT핵심인재양성 지원을 받아 수행된 연구임(IITP-2026-RS-2023-00259867) . 또한, 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(2024-0-00035)

류 및 다중 분류 실험을 통해 비교 분석한다는 점에서 기존 연구와 차별점을 가진다.

3. 제안 기법

3.1 데이터셋 및 입력 특징

본 논문에서는 Stratosphere Laboratory에서 공개한 IoT-23 데이터셋[6]을 사용한다. IoT-23은 실제 IoT 기기에서 수집된 정상 트래픽과 악성코드가 실행된 IoT 기기에서 수집된 악성 트래픽으로 구성된 데이터셋으로, 다양한 IoT 봇넷 및 정상 기기 트래픽을 포함한다.

입력 데이터는 Zeek의 conn.log에서 추출한 네트워크 플로우 정보를 사용한다. 본 논문에서는 proto, duration, orig_bytes, resp_bytes, orig_pkts, resp_pkts, conn_state, missed_bytes, service, id.orig_p, id.resp_p, history의 총 12개 특징을 사용한다.

다중 분류 실험의 레이블은 Benign, PortScan, DDoS, Okiru, C&C (Command & Control)의 5개 클래스로 구성한다. 세부 C&C 계열 레이블은 샘플 수가 적고 동일한 공격 범주에 속하므로 C&C 클래스로 통합하였다. 이진 분류 실험에서는 Benign을 정상 클래스로, 나머지 공격 유형을 Malicious 클래스로 통합하였다. 데이터는 학습 집합과 평가 집합을 8:2 비율로 분할하였으며, 동일한 플로우가 양쪽 집합에 중복 포함되지 않도록 group 기반 분할을 적용하였다.

3.2 Flow-to-Text 변환

사전학습 언어모델은 텍스트 입력을 기반으로 동작하므로, 본 연구에서는 네트워크 플로우의 12개 특징을 텍스트 시퀀스로 변환하는 Flow-to-Text 방식을 적용하였다. 텍스트 표현 방식이 성능에 미치는 영향을 분석하기 위해 다음 세 가지 템플릿을 설계하였다.

T1은 자연어 서술형 표현으로, 각 플로우 특징을 문장 형태로 변환한다. 예를 들어 프로토콜, 지속 시간, 송수신 바이트 수 등을 자연어 문장으로 연결한다. T2는 키=값 구조형 표현으로, 각 특징을 feature=value 형태로 나열한다. T3는 확장 자연어형 표현으로, T1에 각 특징의 의미를 설명하는 추가 문맥을 포함한다. 세 가지 텍스트 표현 방식의 차이를 비교하기 위해, 동일한 네트워크 플로우 데이터를 각 템플릿으로 변환한 예시는 표 1과 같다. T1과 T3는 자연어 표현에 가깝지만 피처명과 값의 경계가 약해질 수 있다. 반면 T2는 자연어 문장성은 낮지만 피처명과 값을 명시적으로 보존하므로, 구조적 데이터의 정보 전달에 유리할 수 있다.

표 1. Flow-to-Text 템플릿 예시

템플릿	예시
T1	A network flow was observed using the tcp protocol. The communication lasted unknown seconds...
T2	Flow summary: protocol=tcp; duration=unknown seconds; source_port=39000; ...
T3	This network session reflects communication behavior over tcp. A device initiated traffic...

4. 실험 설정

본 논문에서는 구조적 데이터를 직접 입력으로 사용하는 기준 모델과 Flow-to-Text 기반 RoBERTa 모델을 비교하였다. 기준 모델로는 Random Forest, XGBoost, MLP를 사용하였다. 이들은 네트워크 침입 탐지 연구에서 널리 사용되는 모델이며, 수치형 플로우 특징을 직접 입력으로 사용한다. IoT-23 데이터셋은 클래스 간 샘플 수 불균형이 존재하므로, 각 모델에는 클래스 불균형을 완화하기 위한 클래스 가중치를 적용하였다. Random Forest와 XGBoost는 트리 수, 최대 깊이 등 주요 하이퍼파라미터를 조정하였으며, MLP는 은닉층 구조, 학습률, epoch 등을 데이터셋 특성에 맞게 설정하였다.

텍스트 기반 모델로는 RoBERTa를 사용하였다[5]. RoBERTa는 BERT의 사전학습 절차를 개선한 모델로, 언어 이해 및 텍스트 분류 작업에서 널리 활용된다. RoBERTa 논문은 BERT의 학습 절차와 하이퍼파라미터를 재검토하여 성능 개선을 제시한 연구이다. 본 논문에서는 roberta-base를 기반으로 T1, T2, T3 템플릿으로 변환한 텍스트를 각각 입력하여 fine-tuning하였다. 주요 하이퍼파라미터는 learning rate 2e-5, epochs 5, batch size 8, max length 256으로 설정하였다. 학습 과정에서는 Macro F1-score를 기준으로 조기 종료를 적용하였으며, patience는 2로 설정하였다.

모델 성능은 클래스 불균형을 고려하여 Macro F1-score를 주요 평가 지표로 사용하였다. Macro F1-score는 각 클래스의 F1-score를 동일한 비중으로 평균하므로, 다수 클래스의 성능에 의해 전체 성능이 편향되는 문제를 완화할 수 있다. 또한 전체적인 분류 정확도를 확인하기 위해 Accuracy를 보조 지표로 사용하였으며, 다중 분류 실험에서는 클래스별 F1-score를 추가로 분석하였다.

5. 실험 결과 및 분석

본 장에서는 이진 및 다중 분류 실험 결과를 제시하고, 모델별 탐지 성능과 클래스별 세부 특성을 비교 분석한다. 5.1절에서는 Macro F1-score와 Accuracy를 중심으로 전체적인 모델 성능을 비교하며, 5.2절에서는 다중 분류의 클래스별 F1-score 세부 분석 결과를 다룬다.

5.1 이진 및 다중 분류 실험 결과

모델별 성능 분석 결과가 표 2에 나타나 있다. 이진 분류에서 RoBERTa-T2가 Macro F1-score 90.54%로 가장 높은 성능을 보였으며, XGBoost(89.93%)와 Random Forest(88.73%) 순으로 높은 탐지 효율을 나타냈다. MLP는 76.99%로 가장 낮은 성능을 나타냈다. RoBERTa-T2는 XGBoost 대비 0.61%p, RF 대비 1.81%p 높은 성능을 기록하였으며, MLP와는 13.55%p의 큰 차이를 보였다.

표 2. 모델별 이진 및 다중 분류 성능 비교

Model	이진 분류		다중 분류	
	Acc	F1	Acc	F1
RF	0.9383	0.8873	0.9387	**0.9471**
XGBoost	0.9439	0.8993	0.8839	0.8759
MLP	0.8667	0.7699	0.7851	0.7721
RoBERTa-T1	0.9453	0.8953	0.8793	0.9036
RoBERTa-T2	**0.9495**	**0.9054**	**0.9390**	**0.9437**
RoBERTa-T3	0.9446	0.8938	0.8796	0.9040

다중 분류에서는 RF가 Macro F1-score 94.71%로 가장 높은 성능을 보였으며, RoBERTa-T2(94.37%)는 0.34%p 차이로 두 번째

높은 성능을 보였다. XGBoost는 87.59%로 RF 대비 7.12%p 낮은 성능을 보였으며, MLP는 77.21%로 가장 낮은 성능을 나타냈다.

RoBERTa 템플릿 간 비교에서는 이진 및 다중 분류 모두에서 T2가 가장 높은 성능을 보였으며, T1과 T3는 유사한 수준을 나타냈다. 이는 키=값 구조형 텍스트 표현이 모델의 입력 이해에 유리하게 작용함을 시사한다.

또한 다중 분류에서 RoBERTa-T1과 T3의 경우 Accuracy는 비교적 낮으나 Macro F1-score는 더 높은 값을 보였다. 이는 클래스 불균형 환경에서 소수 클래스에 대한 예측 성능이 상대적으로 우수함을 의미한다.

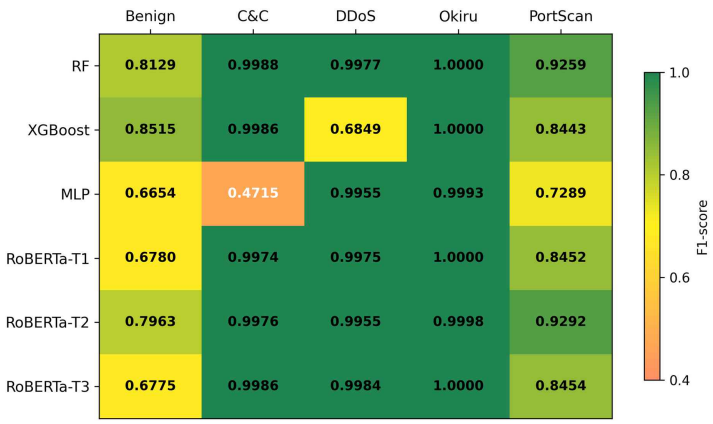
5.2 클래스별 F1-score 분석

그림 1의 히트맵은 다중 분류 실험의 클래스별 F1-score를 보여준다. Okiru 클래스는 모든 모델에서 F1-score 99.9% 이상으로 타 클래스와 명확히 구별되는 고유한 네트워크 패턴이 존재함을 시사한다. 반면 XGBoost에서는 DDoS의 F1-score가 68.49%로 비교적 낮았으며, 이는 PortScan 샘플 일부가 DDoS로 오분류된 영향으로 분석된다. MLP에서는 C&C의 F1-score가 47.15%로 크게 저하되었으며, 다수의 PortScan 샘플이 C&C로 오분류된 결과로 볼 수 있다.

RoBERTa 템플릿 간 비교에서는 T2가 전 클래스에 걸쳐 가장 안정적인 성능을 보였으며, 특히 PortScan에서 T2(92.92%)가 T1(84.52%), T3(84.54%) 대비 8.40%p 높은 성능을 나타냈다.

그림 1의 히트맵을 통해 모델별 클래스 성능 차이를 직관적으로 확인할 수 있으며, RoBERTa-T2가 전반적으로 균형 잡힌 성능을 보이는 반면, 일부 모델은 특정 클래스에서 성능 편차가 크게 나타남을 확인할 수 있다.

그림 1. 모델별 클래스 F1-score 히트맵



6. 결론 및 향후 연구

본 논문은 IoT-23 데이터셋을 기반으로 네트워크 플로우 데이터를 텍스트 시퀀스로 변환하는 Flow-to-Text 방식을 적용하고, 텍스트 표현 구조가 RoBERTa 기반 침입 탐지 성능에 미치는 영향을 분석하였다. 이를 위해 동일한 Zeek conn.log 플로우를 자연어 서술형(T1), 키=값 구조형(T2), 확장 자연어형(T3)의 세 가지 방식으로 변환하였으며, 이진 분류와 다중 분류 실험을 통해 성능을 비교하였다.

실험 결과, 키=값 구조형 표현인 T2가 RoBERTa 기반 모델 중 이진 및 다중 분류 모두에서 가장 높은 성능을 보였다. 특히 이진 분류에서는 RoBERTa-T2가 Macro F1-score 90.54%로 전체 비교 모델 중 가장 우수한 성능을 기록하였으며, 다중 분류에서는 94.37%의 Macro F1-score를 달성하여 Random Forest와 유사한 수준의 성능을 보였다. 이러한 결과는 네트워크 플로우와 같은 구조적 데이터에서는 자연어적 문장성보다 피처명과 값을 명시적으로 보존하는 표현 방식이 언어모델의 학습에 더 효과적일 수 있음을 시사한다.

본 논문은 IoT-23 단일 데이터셋과 RoBERTa 모델을 중심으로 실험을 수행하였다는 점에서 한계가 있다. 또한 세 가지 텍스트 템플릿만을 비교하였으므로, 템플릿 구조와 입력 길이, 피처 순서, 피처 선택 방식 등이 성능에 미치는 영향을 추가적으로 분석할 필요가 있다. 향후 연구에서는 다양한 IoT 침입 탐지 데이터셋으로 일반화 성능을 검증하고, BERT 계열 PLM뿐 아니라 경량 언어모델 및 생성형 언어모델과의 비교를 수행할 계획이다. 또한 attention 분석이나 feature masking 실험을 통해 Flow-to-Text 표현 방식이 모델의 판단 과정에 미치는 영향을 해석하는 연구로 확장할 수 있다.

참고문헌

[1] S. V. N. Santhosh Kumar, M. Selvi, and A. Kannan, "A comprehensive survey on machine learning-based intrusion detection systems for secure communication in Internet of Things," Comput. Intell. Neurosci., vol. 2023, Art. 8981988, 2023. doi: 10.1155/2023/8981988

[2] Z. Chen, J. Liu, Y. Shen, M. Simsek, B. Kantarci, H. T. Mouftah, and P. Djukic, "Machine learning-enabled IoT security: Open issues and challenges under advanced persistent threats," ACM Comput. Surv., vol. 55, no. 5, Art. 105, 2022. doi: 10.1145/3530812

[3] H. Guo, S. Yuan, and X. Wu, "LogBERT: Log anomaly detection via BERT," in Proc. IJCNN, 2021, pp. 1-8. doi: 10.1109/IJCNN52387.2021.9534113

[4] L. Mehavilla, M. Rodríguez, J. García, and Á. Alesanco, "Evaluating large language models effectiveness for flow-based intrusion detection: a comparative study with ML and DL baselines," Artificial Intelligence Review, vol. 59, Article 50, 2026. doi: 10.1007/s10462-025-11432-2

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019

[6] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0)," [Data set]. Zenodo, 2020. http://doi.org/10.5281/zenodo.4743746