차세대 AI 기반 네트워크 침입 탐지 시스템(NIDS) 연구 동향: 실무적 보안 태스크를 중심으로

Research Trends in Next-Generation AI-Based Network Intrusion Detection Systems (NIDS): A Task-Oriented Perspective

유예찬[1] · 홍석원[2] · 최지현[2] · 최석환[3]
Ye-Chan Yu and Seok-Won Hong and Ji-Hyun Choi and Seok-Hwan Choi

[1]연세대학교 소프트웨어학부
[2]연세대학교 전산학과
[3]연세대학교 소프트웨어학부
E-mail: {zkfla2773, sw.hong, jihyun.choi, sh.choi}@yonsei.ac.kr

요 약

본 논문은 인공지능 기술이 적용된 최신 네트워크 침입 탐지 프레임워크들을 실무적 보안 태스크를 기준으로 재분류하여 심층 분석을 수행한다. 구체적으로 미지의 위협을 식별하는 이상 탐지(Anomaly Detection), 위협의 종류를 세분화하는 다중 클래스 공격 분류(Multi-class Attack Classification), 능동적 방어를 위한 위협 예측 및 대응(Threat Prediction & Proactive Response) 역량을 중점적으로 평가한다. 나아가 연산 효율성 저하, 데이터 품질 의존도, 해석 가능성 부족 등 실환경 배포 시 발생하는 기술적 한계점을 도출하며, 연합 학습 및 경량화 모델 설계를 통한 향후 차세대 보안 아키텍처의 발전 방향성을 제시한다.

키워드 : Network Intrusion Detection Systems, Anomaly Detection, Multi-class Attack Classification, Threat Prediction, Large Language Models, Graph Neural Networks, Explainable AI

1. 서론

사물인터넷(IoT) 기기의 급격한 확산과 네트워크 인프라의 고도화로 인해 DDoS, 데이터 유출, 랜섬웨어 등 사이버 위협은 갈수록 지능화되고 있다. 이러한 디지털 생태계를 보호하기 위해 실시간 트래픽 모니터링 및 위험 식별을 수행하는 네트워크 침입 탐지 시스템(NIDS)의 역할이 어느 때보다 중요해졌다. 그러나 초기 시그니처 기반 기법은 사전에 정의된 패턴에만 의존하여 변종 공격 대응에 한계를 보이며, 기존의 이상 탐지 기반 기법 역시 높은 오탐률로 인한 알람 피로도를 유발하는 등 실무적 결함을 드러내고 있다. 따라서 변화하는 위협 환경에 유연하게 대응하기 위해서는 단순한 이상 징후 식별을 넘어, 위협을 정밀하게 분류하고 능동적으로 방어할 수 있는 차세대 보안 체계의 구축이 절실한 시점이다.

최근 사이버 보안 분야에서는 대형 언어 모델(LLM), 그래프 신경망(GNN), 설명 가능한 인공지능(XAI) 등 최신 기계학습 패러다임을 NIDS에 접목하여 이러한 한계를 극복하려는 시도가 활발히 이루어지고 있다. 본 논문은 이러한 최신 기술들을 이상 탐지, 다중 클래스 공격 분류, 위협 예측 및 능동적 대응이라는 세 가지 핵심 태스크를 중심으로 분석한다. 이를 통해 탐지 모델이 실제 보안 현장에서 달성해야 하는 실무적 목적에 초점을 맞추어 현재의 기술적 한계점을 진단하고 미래 지향적인 발전 과제를 도출하고자 한다.

본 연구는 상대정글한테 108갱 당하고 미드차이로 따잇당해서 서글프게 울부짖는 미드라이너가 부득이하게 정글차이를 외치는 연구다.

2. 본론

2.1 이상 탐지 (Anomaly Detection)

정상적인 트래픽 패턴에서 벗어나는 제로데이(Zero-day) 공격을 식별하는 이상 탐지 태스크는 NIDS의 가장 기본적인 역할이다. 해당 과제에서는 다양한 포맷의 네트워크 데이터를 얼마나 빠르고 정확하게 정상/이상으로 이진 분류(Binary classification)해 내는지가 핵심 관건이다. 하지만, 기존의 단순 통계나 얕은 기계학습 기법으로는 고차원적이고 비선형적인 현대 네트워크 트래픽의 숨겨진 패턴을 파악하는 데 한계가 존재하였다. 이에 따라 방대한 데이터 내부에 내재된 복잡한 특징(Feature)을 스스로 학습하고 추출할 수 있는 고도화된 인공지능 아키텍처의 도입이 본격화되었다. 일례로, 트랜스포머를 활용한 TabTransformer 연구 [7]는 숫자형 및 범주형 데이터로 섞여 있는 표 형태의 네트워크 트래픽 특성을 셀프 어텐션(Self-attention) 메커니즘으로 추출하여 탐지 정밀도를 높였다.

양질의 라벨링 데이터가 부족한 실무적 한계를 극복하기 위한 연구도 이어지고 있다. TS-IDS [4] 모델은 네트워크 흐름을 노드와 엣지로 모델링하는 그래프 구조에 자기 지도 학습(Self-supervised learning)을 결합하여, 정답 라벨 없이도 네트워크의 구조적 이상 징후를 포착하는 뛰어난 성능을 입증하였다. 나아가 Yang 등 [8]은 지식 증류(Knowledge Distillation)를 활용해 거대 LLM 교사 모델의 지식을 경량 합성곱 신경망(CNN) 학생 모델로 이전시킴으로써, 배포 환경의 제약을 극복하는 훈

련 방법론을 제시했다. 이처럼 최신 이상 탐지 모델들은 정형 데이터의 피처 추출(TabTransformer)에서부터 미라벨 데이터 구조 분석(TS-IDS)에 이르기까지 구체적인 실무 제약을 해결하는 방향으로 발전하고 있다.

2.2 다중 클래스 공격 분류 (Multi-class Attack Classification)

단순한 이상 징후 포착을 넘어, 식별된 위협이 분산 서비스 거부(DDoS)나 멀웨어(Malware) 등 어느 것에 해당하는지 구체적으로 분류하는 태스크는 신속한 사고 대응을 위해 필수적이다. 공격의 유형을 정밀하게 나누기 위해서는 개별 패킷뿐만 아니라 네트워크 전반의 연결 구조와 동작 맥락을 이해해야 한다.

Wang 등 [5]이 제안한 BS-GAT 모델은 서브넷 유사성과 트래픽 동작 규칙을 엣지(Edge)의 속성으로 정의하여 동작 유사성 기반 그래프를 생성한다. 이렇게 도출된 엣지 동작 가중치를 그래프 어텐션 네트워크(GAT)에 통합함으로써, 다단계로 이루어지는 복합적인 다중 클래스 침입 탐지 분류 성능을 대폭 향상시킨다.

반면, 네트워크 페이로드의 순차적 특성에 집중하여 자연어 처리 기법을 도입한 모델들도 존재한다. SecurityBERT [3]는 프라이버시 보존 고정 길이 인코딩(PPFLE) 기법을 사용하여 트래픽 데이터를 텍스트 형태로 변환한 후 분석한다. 이 모델은 1,100만 개의 파라미터로 구성된 경량화된 15계층 BERT 아키텍처를 사용하여, 구조적 분석과는 다른 언어적 문맥 분석 접근을 통해 다중 분류 태스크를 효과적으로 수행한다.

2.3 위협 예측 및 능동적 대응 (Threat Prediction and Proactive Response)

최근에는 공격의 발생을 사전에 예측하거나 분석가가 즉각적으로 대응할 수 있도록 해석 가능한 완화 전략을 생성하는 방향으로 NIDS의 역할이 진화하고 있다. 능동적 대응 태스크에서는 방어 결정에 대한 신뢰성 확보와 실효성 있는 대응책 마련이 핵심이다.

위협 예측 분야에서 Diaf 등 [2]은 장단기 메모리(LSTM)의 시계열 분석 능력과 대형 언어 모델의 문맥 이해 장점을 융합하여 사전 예방적 침입 예측 시스템을 구현하였다. 또한, AI 방어 시스템의 불투명성이라는 고질적 한계를 극복하기 위해 Ali와 Kostakos [1]는 HuntGPT를 제안했다. 이 프레임워크는 이상 탐지 엔진에 GPT-3.5 모델을 결합하고 SHAP 및 LIME과 같은 설명 가능한 AI(XAI) 기법을 도입하여 탐지 결과에 대한 투명하고 논리적인 해석을 제공함으로써 보안 담당자의 신뢰를 제고한다.

한발 더 나아가, 위협 완화 전략(Mitigation)을 직접적으로 생성하는 태스크를 다룬 ShieldGPT 프레임워크 [6]는 시사하는 바가 크다. 이 시스템은 GPT-4를 활용하여 공격 감지 알림을 넘어서는 상세한 분석 보고서를 도출하며, 특히 역할 기반 프롬프트(Role-based prompts)를 적용하여 Cisco 라우터나 Snort IPS 등 특정 보안 하드웨어 환경에 즉시 적용할 수 있는 실용적이고 구체적인 대응 지침을 동적으로 생성해 낸다.

2.4 기술적 한계 및 운영상 과제 (Technical Limitations and Challenges)

혁신적인 기법들의 등장에도 불구하고, 이를 실제 운영 환경에 배포하기 위해서는 여러 한계점들을 극복해야 한다. 첫째, 대부분의 인공지능 기반 NIDS는 고품질의 라벨링된 데이터셋에 절대적으로 의존하므로 훈련 데이터에 포함되지 않은 새로운 형태의 프로토콜이나 공격 시나리오 발생 시 탐지 성능이 급격히 저하되는 일반화 한계를 겪는다. 둘째, LLM이나 멀티 홉 어텐션(Multi-hop attention)을 활용하는 GNN 모델들은 학습 및 추론 과정에서 막대한 컴퓨팅 자원을 요구한다. 이로 인해 자원 제약이 극심한 대규모 동적 네트워크 환경에서는 실시간 위험 탐지를 보장하기 어렵다.

3. 결론

본 논문에서는 차세대 NIDS에 적용된 최신 인공지능 기술의 연구 동향을 이상 탐지, 다중 클래스 공격 분류, 위협 예측 및 능동적 대응이라는 세 가지 핵심 보안 태스크를 기준으로 분류하고, 실제 운영 환경에서의 기술적 한계점을 진단하였다. 대형 언어 모델(LLM)의 문맥 분석과 그래프 신경망(GNN)의 관계 추론 능력을 융합한 지능형 방어 체계가 현장에 온전히 안착하기 위해서는 여러 당면 과제를 해결해야 한다. 가장 시급한 현안은 모델 압축 및 양자화(Quantization) 기법을 통해 막대한 컴퓨팅 자원 없이도 엣지 디바이스에서 실시간 위험 식별이 가능한 경량 아키텍처를 구축하는 일이다. 이와 더불어, 진화하는 제로데이(Zero-day) 침해와 적대적 공격(Adversarial attacks)으로부터 모델 스스로를 보호할 수 있도록 지속적 학습(Continuous learning) 메커니즘과 입력 데이터 무결성 검증 절차가 요구된다. 마지막으로, 소프트웨어 의존성이 급증하는 시점인 만큼, 런타임 환경 내 서드파티 모듈의 취약점이나 인공지능 모델 자체를 겨냥한 공급망 공격(Supply Chain Attack)을 원천 차단하는 강건한 보안 인프라 연구가 병행되어야 한다.

참 고 문 헌

[1] Ali, T., and P. Kostakos. "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (llms)." arXiv preprint arXiv:2309.16021 (2023).

[2] Diaf, A., et al. "Beyond detection: Leveraging large language models for cyber attack prediction in IoT networks." 2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSSIoT). IEEE, 2024.

[3] Ferrag, M. A., et al. "Revolutionizing cyber threat detection with large language models: A privacy-preserving Bert-based lightweight model for iot/iiot devices." IEEE Access (2024).

[4] Nguyen, H., and R. Kashef. "TS-IDS: Traffic-aware self-supervised learning for IoT network intrusion detection." Knowledge-Based Systems 279 (2023): 110966.

[5] Wang, Y., et al. "BS-GAT behavior similarity based graph attention network for network intrusion detection." arXiv preprint arXiv:2304.07226 (2023).

[6] Wang, T., et al. "ShieldGPT: An LLM-based framework for DDoS mitigation." Proceedings of the 8th Asia-Pacific workshop on networking. 2024.

[7] Wang, X., et al. "Advanced network intrusion detection with tabtransformer." Journal of Theory and Practice of Engineering Science 4.03 (2024): 191-198.

[8] Yang, Y., et al. "An anomaly detection model training method based on LLM knowledge distillation." 2024 international conference on networking and network applications (NaNA). IEEE, 2024.