RNN/LSTM 기반 네트워크 침입 탐지 시스템에서 도메인 적응을 위한 LoRA 적용 위치에 따른 성능 비교 연구

A Comparative Study of LoRA Injection Points for Domain Adaptation in RNN/LSTM-based Network Intrusion Detection Systems.

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

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

요 약

복잡한 사이버 위협에 대응하기 위해 NIDS의 중요성이 증가하고 있다. 최근에는 시간에 따른 트래픽 패턴을 고려한 시계열 기반 모델이 주목받고 있으나, 학습 데이터 의존성이 높아 데이터셋 간 분포 차이로 인해 일반화 성능이 저하되는 문제가 발생한다. 이를 완화하기 위한 접근으로 도메인 적응이 널리 활용되고 있으며, LoRA가 하나의 방법으로 사용되고 있다. 본 논문에서는 LoRA를 NIDS에 적용하고, 입력 계층, 중간 계층, 완전 연결 계층에서의 적용 위치에 따른 성능 차이를 비교 분석하여 효과적인 적응 전략을 제시한다.

키워드 : Network Intrusion Detection System, Domain Adaptation, LoRA

1. 서론

최근 네트워크 공격은 단일 패킷이나 개별 플로우의 통계적 특징만으로 구분하기 어려우며, 시간에 따른 트래픽 패턴을 통해 점진적으로 드러나는 경우가 많다. 이에 따라 네트워크 침입 탐지 시스템(Network Intrusion Detection System, NIDS)은 네트워크 트래픽을 연속적인 시퀀스로 모델링하는 시계열 기반 접근으로 발전하고 있다. 그러나 시계열 기반 NIDS는 학습 데이터의 구조에 크게 의존하여, 서로 다른 환경의 트래픽에 대해서는 모델 성능이 급격히 저하되는 도메인 불일치 문제가 발생한다. 이러한 문제를 완화하기 위한 방법으로, 최근 다양한 분야에서는 파라미터 효율적인 적응 기법인 LoRA(Low-Rank Adaptation)를 활용하여 도메인 간 성능 저하를 줄이려는 연구가 진행되고 있다[1][2].

본 논문에서는 LoRA를 시계열 기반 NIDS 모델에 적용하여 도메인 적응 성능을 분석한다. 특히, LoRA를 모델의 입력 계층, 중간 계층, 완전 연결 계층에 각각 적용하여 적용 위치에 따른 성능 차이를 비교 분석한다. 이를 통해 LoRA의 적용 위치가 도메인 적응 성능에 미치는 영향을 평가하고, 효과적인 적용 전략을 제시한다.

2. 본론

본 장에서는 RNN 계열 NIDS 모델에 대해 LoRA의 적용 위치가 도메인 적응 성능에 미치는 영향을 실험적으로 분석한다.

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2026-RS-2023-00259967, 70%)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2026-25472714, 30%)

2.1 데이터셋

본 논문에서는 서로 다른 환경에서 수집된 2개의 데이터셋(KDD99[3], CICIDS2017[4])을 선정하여 실험을 진행하였다. KDD99 데이터셋은 1998년에 구축된 대표적인 침입 탐지 데이터셋으로, 총 42개의 feature와 4개의 주요 공격 범주(DoS, U2R, R2L, Probing), 그리고 22개의 세부 공격 클래스로 구성된다. CICIDS2017 데이터셋은 2017년에 수집된 데이터셋으로 총 79개의 feature와 14개의 공격 클래스를 포함하고 있으며, 보다 현실적인 네트워크 트래픽 환경을 반영하고 있다.

2.2 데이터셋 전처리

본 논문에서는 데이터 전처리 단계에서 음수 값을 0으로 치환하였다. 이후 Inf 값을 NaN으로 변환한 뒤, 결측값이 포함된 컬럼과 중복 데이터를 제거하였다. 다음으로 범주형 데이터는 LabelEncoder를 사용하여 수치형 데이터로 변환하였다. 라벨은 이진 분류 설정에 따라 정상 트래픽은 0, 나머지 공격 트래픽은 1로 매핑하였다. 마지막으로 수치형 특성에 대해 Min-Max Scaling을 적용하여 데이터의 범위를 [0, 1] 구간으로 정규화하였다.

2.3 실험 환경

본 논문에서는 먼저 각 데이터셋에 대해 RNN 및 LSTM 모델을 각각 학습하여 총 4개의 백본 모델을 구성하였다. 입력 데이터는 (batch_size, sequence_length, feature_dim) 형태의 3차원 시계열 텐서로 구성되며, RNN은 순차적 시간 의존성을 학습하고 LSTM은 게이트 메커니즘을 통해 장기 의존성 학습을 강화한다. 이후 각 모델의 입력 계층, 중간 계층, 완전 연결 계층에 LoRA를 적용하고, 학습에 사용되지 않은 다른 데이터셋에 대해 테스트를 수행함으로써 LoRA의 적용 위치에 따른 도메인 적응 성능에 대해 평가하였다.

표 1 : LoRA 위치에 따른 성능 비교
Table 1. Performance Comparison by LoRA Location

Model	Reference dataset	Test dataset	LoRA application location	Accuracy	Precision	Recall	F1-score
RNN	KDD99	CICIDS2017	None	0.7614	0.0970	0.0497	0.0657
			Input layer	0.8929	0.9485	0.3866	0.5493
			Hidden layer	0.8860	0.6985	0.5717	0.6288
			FC layer	0.8860	0.7028	0.5632	0.6253
	CICIDS2017	KDD99	None	0.6838	0.3271	0.2806	0.3021
			Input layer	0.9561	0.8680	0.9669	0.9148
			Hidden layer	0.9900	0.9894	0.9694	0.9793
			FC layer	0.9786	0.9925	0.9191	0.9544
LSTM	KDD99	CICIDS2017	None	0.8224	0.4610	0.3043	0.3666
			Input layer	0.8079	0.4565	0.7206	0.5589
			Hidden layer	0.8804	0.7376	0.4530	0.5613
			FC layer	0.8663	0.6421	0.4708	0.5433
	CICIDS2017	KDD99	None	0.6562	0.3050	0.3202	0.3124
			Input layer	0.9575	0.8698	0.9711	0.9176
			Hidden layer	0.9824	0.9777	0.9496	0.9634
			FC layer	0.9581	0.8825	0.9557	0.9176

이 과정에서 데이터셋마다 서로 다른 feature 구성을 가지고 있어 이로 인한 입력 차원 불일치 문제가 발생한다. 이를 해결하기 위해, 학습에 사용된 기준 데이터셋의 feature 목록을 기반으로 테스트 데이터셋의 feature를 재정렬하고, 누락된 feature는 0으로 채워 모든 실험에서 동일한 입력 구조를 가지도록 설정하였다. 또한 네트워크 침입 탐지 데이터셋에서 나타나는 클래스 불균형 문제를 완화하기 위해 학습 단계에서 정상 트래픽을 공격 트래픽의 개수에 맞추어 시퀀스 단위로 무작위 언더샘플링을 수행하였다.

2.4 실험 결과 및 성능 비교

본 논문에서는 해당 내용에 대한 성능을 평가하기 위해 분류 성능 평가에 일반적으로 사용되는 Accuracy, Precision, Recall, F1-score를 사용하였다.

표1을 기반으로 RNN과 LSTM 모델에 대해 LoRA의 적용 위치에 따른 성능을 분석한 결과, F1-score 기준에서는 대부분의 데이터셋 조합에서 hidden layer에 LoRA를 적용한 경우 가장 우수한 성능을 보이는 경향이 확인되었다. 그러나 NIDS 환경에서는 공격 탐지 성능을 나타내는 recall이 더욱 중요한 지표이다. 이에 따라 recall을 중심으로 추가적인 분석을 수행하였다. 그 결과, RNN 모델의 경우, hidden layer에 LoRA를 적용했을 때 가장 높은 recall을 보였으며, 이는 중간 계층의 feature 표현을 보정하는 것이 도메인 간 공격 패턴의 차이를 효과적으로 반영하는 데 유리함을 의미한다. 반면, LSTM 기반 모델에서는 input layer에 LoRA를 적용했을 때 가장 높은 recall을 보였다. 따라서, 이러한 결과는 LoRA의 최종 적용 위치가 모델 구조에 따라 상이한 영향을 미치며, 도메인 적응 전략 성능에 영향을 미치는 중요한 요소임을 보여준다. 특히, RNN 계열에서는 중간 표현 보정이, LSTM 계열에서는 입력 분포 정렬이 보다 효과적인 전략임을 확인할 수 있다.

3. 결론

본 논문에서는 KDD99, CICIDS2017 데이터셋을 활용하여 LoRA의 적용 위치에 따른 도메인 적응 성능을 실험적으로 분석하였다. 실험 결과, LoRA의 적용 위치에 따라 도메인 적응 성능에 유의미한 차이가 발생함을 확인하였으며, F1-score 기준에서는 대부분 hidden layer에서 가장 우수한 성능을 보였다. 한편, NIDS에서 중요한 지표인 recall 기준에서는 RNN은 hidden layer, LSTM은 input layer에서 가장 높은 성능을 보였다. 이는 LoRA의 적용 위치가 도메인 간 특징 차이를 효과적으로 보정하는 데 중요한 요소임을 시사한다. 따라서 본 연구는 시계열 기반 NIDS에서 LoRA를 적용할 때, 단순한 도입 여부뿐만 아니라 모델 구조를 고려한 적용 위치 선택이 중요함을 확인하였다. 향후 연구에서는 다양한 데이터셋과 모델 구조에 대해 LoRA 적용 가능성을 확장하고, 보다 일반화 가능한 도메인 적응 전략에 대한 연구를 진행할 예정이다.

참 고 문 헌

[1] Q. Hu, et al., “Vector-based LoRA expert fusion for efficient speech recognition,” SSRN, 2024. [Online]. Available: https://ssrn.com/abstract=5937761

[2] L. Page-Caccia, et al., “Multi-head adapter routing for cross-task generalization,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 56916 - 56931, 2023.

[3] University of California, Irvine, “KDD Cup 1999 Data,” 1999. [Online]. Available: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[4] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in Proc. Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), vol. 1, pp. 108 - 116, 2018.