클라우드 기반 AI 모델을 이용한 DDos 공격 탐지 및 완화

정 여 원
성신여자대학교

Detection and mitigation of DDos attacks using cloud-based AI models

Jeong Yeo won
Sungshin University

요 약

이 연구는 클라우드 환경에서 발생하는 DDoS 공격 탐지의 한계를 극복하기 위해 대규모 언어 모델(LLM)을 활용하여 보다 정교하고 신뢰성 있는 탐지 시스템을 제안한다. 주요 방법으로는 CICIDS2017 데이터셋을 사용하여 최신 LLM 모델(BERT, GPT, RoBERTa 등)의 성능을 비교하고, 상위 2개의 모델을 앙상블 기법으로 결합하여 탐지 성능을 극대화하였다. 또한 SMOTE, Borderline-SMOTE, ADASYN, SMOTEENN, SMOTETomek 등의 데이터 증강 기법을 활용하여 공격 패턴의 다양성을 학습하고, SHAP과 LIME을 도입하여 모델의 예측 과정을 설명 가능하게 하였다. 본 연구는 기존 머신러닝 및 딥러닝 기법의 한계를 극복하고, 클라우드 환경에서 신뢰할 수 있는 DDoS 탐지 체계를 구축하는 데 기여한다.

ABSTRACT

This study proposes a robust and reliable DDoS attack detection system in cloud environments using large language models (LLMs). Using the CICIDS2017 dataset, the performance of state-of-the-art LLMs (e.g., BERT, GPT, RoBERTa) was evaluated, and the top two models were combined into an ensemble to maximize detection accuracy. Additionally, data augmentation techniques such as SMOTE, Borderline-SMOTE, ADASYN, SMOTEENN, and SMOTETomek were applied to enrich attack patterns for learning. SHAP and LIME were incorporated to make the ensemble model's predictions interpretable. This research addresses the limitations of traditional machine learning and deep learning approaches, contributing to the development of a reliable DDoS detection system in cloud environments.

Keywords): DDoS detection, large language models, cloud environments, data augmentation, ensemble learning, interpretability, SHAP, LIME

I. 서 론

DDoS(Distributed Denial of Service) 공격은 대규모 네트워크의 가용성을 저하시켜 시스템에 큰 피해를 입히는 주요 사이버 공격 중 하나로, 특히 클라우드 기반 환경에서는 공격의 빈도와 강도가 높아지고 있다. 기존의 DDoS 탐지 방법은 주로 머신러닝 및 딥러닝 기반의 기술을 사용하였으나, 네트워크 트래픽의 복잡한 패턴을 완전히 반영하지 못하는 한계를 보였다.

본 연구는 클라우드 환경에서 대규모 언어 모델(LLM: Large Language Model)을 활용하여 DDoS 공격을 보다 정교하게 탐지하고, 이를 완화하기 위한 시스템을 제안한다. 특히 여러 LLM 모델을 비교 분석

하여 상위 성능을 보이는 2개의 모델을 앙상블 방식으로 결합한 후, 데이터 증강 기법을 적용하여 앙상블 모델의 성능을 추가적으로 향상시킨다. 마지막으로, 설명 가능한 AI 기법(SHAP, LIME)을 도입하여 앙상블 모델의 예측 과정을 설명하고, 모델의 해석 가능성을 높인다.

본 연구는 기존 연구와 달리 LLM 모델을 통한 DDoS 탐지 성능 향상과 모델의 해석 가능성을 동시에 고려한 연구로, 클라우드 환경에서 DDoS 탐지 시스템을 보다 신뢰성 있고 효율적으로 구축하는 것을 목표로 한다.

II. 연구 동기

클라우드 환경이 확산됨에 따라, 대규모 네트워크 인프라가 DDoS 공격의 주요 표적이 되고 있다. 특히 클라우드 서비스는 다수의 사용자와 트래픽을 처리해야 하기 때문에, 가용성에 치명적인 영향을 미치는 DDoS 공격에 매우 취약하다. 기존의 머신러닝 및 딥러닝 기반 탐지 방법들은 일정 수준의 탐지 성능을 보여왔으나, 다양한 공격 패턴을 완전히 반영하지 못하고 정확도가 떨어지거나, 오탐률이 높은 문제가 있다.

대규모 언어 모델(LLM)은 자연어 처리(NLP) 분야에서 탁월한 성능을 입증한 모델로, 문맥을 이해하고 복잡한 데이터를 처리하는 능력이 뛰어나다. 이러한 특성을 네트워크 트래픽 분석에 적용함으로써, 기존 탐지 모델의 한계를 극복하고 보다 정교한 DDoS 탐지를 기대할 수 있다.

본 연구는 LLM 모델을 여러 개 학습시켜 그 성능을 비교하고, 상위 2개의 모델을 결합하여 앙상블 모델을 구성함으로써 탐지 성능을 극대화하고자 한다. 또한, 데이터 증강을 통해 더 다양한 공격 패턴을 학습할 수 있도록 하여 모델의 일반화 성능을 높이는 것을 목표로 한다. 마지막으로, 설명 가능한 AI 기법을 도입하여 모델의 예측 과정을 투명하게 설명하고, 실시간 클라우드 환경에서 신뢰성 있는 DDoS 탐지 시스템을 구축하는 데 기여하고자 한다.

III. 관련 연구

3.1 소단원 작성 기법

DDoS 공격 탐지에 관한 기존 연구는 주로 머신러닝 및 딥러닝 기법을 활용하여 네트워크 트래픽 데이터를 분석하는 방식으로 진행되었다. 대표적으로 SVM(Support Vector Machine), Random Forest, KNN(K-Nearest Neighbors)과 같은 전통적인 머신러닝 모델들이 많이 사용되었다. Li et al. (2020)의 연구에서는 랜덤 포레스트 알고리즘을 사용하여 DDoS 공격을 탐지했으나, 대규모 네트워크 트래픽의 복잡성을 처리하는 데 한계가 있었다. Kumar et al. (2019)은 SVM을 활용하여 DDoS 공격 탐지를 시도했으나, 데이터의 비선형적 특성을 반영하지 못해 탐지율이 낮아지는 문제가 있었다.

딥러닝 기반의 연구로는, Tang et al. (2021)이 LSTM(Long Short-Term Memory) 모델을 사용하여 시계열 데이터를 기반으로 DDoS 공격을 탐지한 연구가 있다. 이 연구는 DDoS 공격 탐지에 있어서 기존의 머신러닝 모델보다 높은 성능을 보였지만, 네트워크 트래픽의 복잡한 문맥적 정보를 충분히 학습하지 못했다.

특히, Sharafaldin et al. (2018)의 연구에서는 CICIDS2017 데이터셋을 사용하여 DDoS 공격을 탐지하는 다양한 모델(SVM, KNN, Random Forest)을 비교 분석하였다. 이 데이터셋은 실제 네트워크 환경에서 발생할 수 있는 다양한 공격 패턴을 포함하고 있어 DDoS 탐지 연구에서 자주 사용된다. 그러나 해당 연구에서도 전통적인 모델들이 대규모 데이터에서의 성능 저하와 오탐률 증가 문제를 겪었다.

이와 대비하여, 대규모 언어 모델(LLM)은 자연어 처리에서 문맥을 이해하는 능력을 기반으로 텍스트뿐만 아니라 복잡한 패턴을 포함한 네트워크 트래픽 분석에서도 유망한 접근법으로 주목받고 있다. BERT, GPT, RoBERTa와 같은 모델들은 NLP 분야에서 높은 성능을 보였으며, 이를 DDoS 탐지에 적용하여 기존 모델들의 한계를 극복할 수 있는 가능성이 제시되고 있다.

본 연구는 CICIDS2017 데이터셋을 사용하여 여러 LLM 모델(BERT, GPT, RoBERTa등)의 성능을 비교 분석하고, 상위 2개의 모델을 앙상블하여 성능을 극대화하는 것을 목표로 한다. 또한, 데이터 증강 기법을 적용하여 공격 패턴을 더욱 다양화하고, 설명 가능한 AI 기법(SHAP, LIME)을 사용하여 모델의 예측 과정을 설명함으로써 연구의 신뢰성을 높인다.

IV. 연구 설계 및 실험

4.1 연구 방법 및 설계

본 연구는 클라우드 환경에서 DDoS 공격을 효과적으로 탐지하기 위해 여러 대규모 언어 모델(LLM)을 학습하고, 성능을 비교한 후, 상위 모델들을 앙상블하여 성능을 극대화하는 방식을 채택하였다. 전체적인 연구 절차는 다음과 같이 구성된다.

4.1.1 데이터 준비 및 전처리

본 연구에서는 CICIDS2017 데이터셋을 사용하여 DDoS 공격 탐지 모델을 학습시킨다. 이 데이터셋은 다양한 형태의 네트워크 공격과 정상 트래픽을 포함하고 있어 모델 학습에 적합하다. 네트워크 트래픽 데이터를 자연어 처리 모델이 학습할 수 있는 형식으로 변환하기 위해 각 트래픽 세션을 텍스트 형식으로 변환하고, 이를 토크나이징(Tokenization)하여 모델의 입력으로 사용할 수 있도록 전처리한다.

4.1.2 모델 학습 및 성능 비교

여러 LLM 모델을 비교하기 위해 BERT, RoBERTa, GPT-2, XLNet, T5 이상 5개의 최신 대규모 언어 모델을 학습시킨다. 각 모델에 대해서 다음과 같은 절차를 수행한다.

4.1.2.1 모델 학습

각 LLM 모델을 CICIDS2017 데이터셋에 학습시킨다. 학습 과정에서는 모델의 하이퍼파라미터를 조정하고, 성능이 최적화될 수 있도록 한다.

4.1.2.2 성능 평가

각 모델의 성능을 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수등의 지표로 평가한다. 각 지표를 바탕으로 DDoS 공격 탐지에서의 성능을 비교한다.

4.1.3 상위 2개 모델 앙상블

성능 평가 결과를 바탕으로 상위 2개의 모델을 선정하여, 이들을 앙상블 모델로 결합한다. 스태킹 앙상블, 확률평균 앙상블, 랜덤포레스트 앙상블, XGBOOST 앙상블 의 4개 앙상블 기법을 사용하여 최적의 결과가 나온 앙상블 모델을 저장한다.

4.1.4 데이터 증강 적용

DDoS 공격 탐지 성능을 더욱 향상시키기 위해 데이터 증강(Data Augmentation) 기법을 적용한다. 증강 기법으로는 SMOTE(Synthetic Minority Over-sampling Technique) 데이터 증강, Borderline-SMOTE, ADASYN(Adaptive Synthetic Sampling), SMOTEENN(SMOTE + Edited Nearest Neighbor), SMOTETomek, Random Under-sampling, 그리고 NearMiss기법이 사용된다. 이러한 데이터 증강 기법은 기존 데이터의 불균형 문제를 해결하고, 다양한 공격 패턴을 모델이 학습할 수 있도록 돕는다. 증강된 데이터는 모델 학습 과정에서 더욱 균형 잡힌 데이터셋을 제공하며, 이후 이 데이터를 사용하여 앙상블 모델의 성능을 재평가한다.

4.1.5 모델 해석 가능성 확보

증강된 데이터를 사용한 앙상블 모델이 탐지한 결과를 해석하기 위해 설명 가능한 AI 기법을 도입한다. 구체적으로, SHAP(Shapley Additive Explanations)와 LIME(Local Interpretable Model-agnostic Explanations)을 사용하여 모델의 예측 과정에서 어떤 특징이 중요한 역할을 했는지 시각적으로 설명한다.

SHAP을 통해, 각 트래픽 세션의 특징이 DDoS 공격 탐지에 어떻게 기여했는지 평가한다.

LIME을 사용하여 개별 데이터 포인트에 대한 모델의 예측을 로컬 수준에서 해석하고, DDoS 공격 탐지의 주요 패턴을 시각화한다.

4.2 연구 결과

4.2.1 여러 LLM모델 결과

Table 1. LLM 모델의 결과

Model	eval_loss	eval_runtime	eval_sample_per_sec	eval_steps_per_seco	epoch

			ond	nd	
BERT	0.0393202379345893 86	4024.8686	140.663	8.792	0.0176631010753295 94
RoBERTa	0.0673397555947303 8	4115.18	137.576	8.599	0.0105978606451977 57
gpt-2	0.0702500045299953	4648.2064	121.799	7.613	0.0176631010753295 94
XLNet	0.2144566774368286	6372.2992	88.845	5.553	0.0070652404301318 38
T5	0.0610836111009120 94	1341.8428	421.919	26.37	0.0105978606451977 57

4.2.2 앙상블 모델 결과

Table 2. 앙상블 모델에 대한 결과

Ensemble	Precision	Recall	F1-score	Confustoin Matrix
Stacking	0.9751572760251641	0.9592989113249311	0.9671630914970168	[[90552, 541] [901, 21236]]
Probability Averaging	0.9849974530946922	0.9850498720301546	0.9849489367598133	[[45201 1, 2277] [

				6187, 105674]]
Random Forest	0.9918206517336 99	0.9918396184756 69	0.9918243051180 21	[[90740, 353] [571, 21566]]
XGboost	0.9895644258762287	0.9895963967146516	0.9895690978668543	[[90656, 437] [741, 21396]]

4.2.3 데이터 증강 후 앙상블 모델 결과 변화

Table 3. 데이터 증강후의 모델 결과

Technique	Precision	Recall	F1-score	Confustion Matrix	Accuracy
SMOTE	0.6831993168037969	0.7508522476375519	0.7101328724812705	[[83186, 7875] [20336, 1833]]	0.7508522476375519
Borderline-SMOTE	0.6837290339874998	0.7506491212576173	0.7103740443247817	[[83125, 7936] [20298, 1871]]	0.7506491212576173
ADASYN	0.6835659891898225	0.7527333745473814	0.7108511102904942	[[83444, 7617] [20381, 1788]]	0.7527333745473814
SMOTEE E	0.6862232754	0.77084694	0.71619861931	[[86019,5042]	0.7708469486885

NN	913581	86885101	9471	[20905, 1264]]	101
SMOTETomek	0.68327415517868828	0.7556919544290382	0.7114648187972993	[[83899,7162] [20501, 1668]]	0.7556919544290382
RandomUnder-sampling	0.68511819550624848	0.7353793164355736	0.7064617587973262	[[80757, 10304] [19659, 2510]]	0.7353793164355736
NearMiss	0.68499099556282877	0.5055197385851806	0.5562208313163571	[[46366, 44695] [11295, 10874]]	0.5055197385851806

4.3 모델 설명력

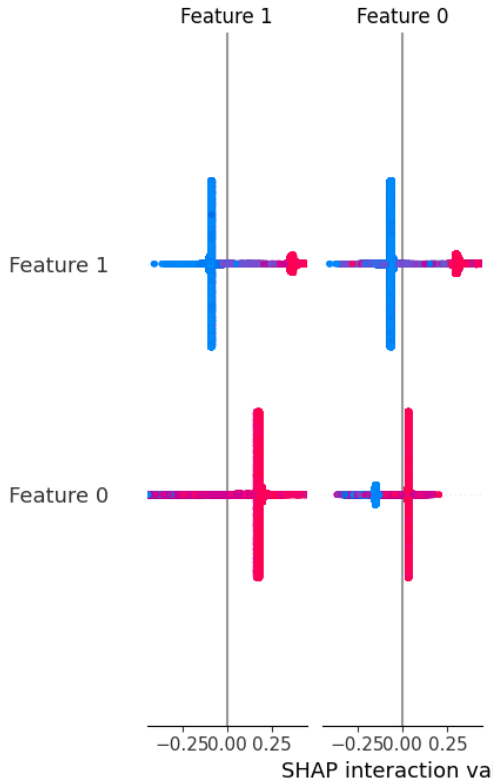

Fig.1. SHAP 결과

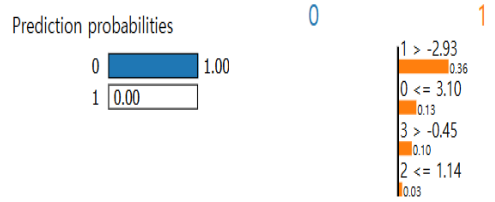

Fig.2. LIME 결과 -1

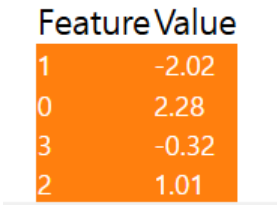

Fig.3. LIME 결과 -2

V. 결 론

본 연구는 대규모 언어 모델과 데이터 증강 기법을 결합하여 클라우드 환경에서 발생하는

DDoS 공격 탐지 성능을 효과적으로 향상시켰다. 특히, CICIDS2017 데이터셋과 최신 LLM 모델을 기반으로 한 탐지 성능 분석 및 앙상블 기법의 적용은 기존 탐지 기법의 한계를 극복하는 데 기여하였다. 데이터 증강을 통해 공격 패턴의 다양성을 학습하고, 설명 가능한 AI 기법을 통해 모델의 해석 가능성을 제공함으로써 신뢰성을 강화하였다. 이 연구는 클라우드 기반 시스템에서 DDoS 탐지를 위한 실질적이고 신뢰할 수 있는 프레임워크를 제시하며, 향후 대규모 네트워크 환경에서의 확장 가능성을 시사한다.

References

[1] F. Li, Y. Yang, and Y. Luo, "Random forest algorithm for DDoS attack detection in cloud computing," *International Journal of Digital Content Technology and its Applications*, vol. 14, no. 1, pp. 56-63, 2020.

[2] R. Kumar, M. Singh, and A. Singh, "Support vector machine-based DDoS attack detection and analysis," *International Journal of Network Security*, vol. 18, no. 2, pp. 163-170, 2019.

[3] S. Tang, J. Zhang, and Y. Wu, "LSTM-based DDoS attack detection in cloud computing environments," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 10, no. 3, pp. 112-130, 2021.

[4] A. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108-116, 2018.

〈저자소개〉

정여원 (Jeong-Yeo-won)
성신여대 20221426 AI학과