CNN 기반 침입 탐지 시스템의 입력 모형 최적화

백의준, 김보선, 박지태, 신창의*, 김명섭
고려대학교, *국방기술품질원
{pb1069, boseon12, pjj5846, tmskim}@korea.ac.kr, superego99@dtaq.re.kr*

Optimizing Input Shapes for CNN-Based Intrusion Detection Systems

Ui-Jun Baek, Boseon Kim, Jee-Tae Park, Chang-Yui Shin*, Myung-Sup Kim
Korea University, *Defense Agency for Technology and Quality

요 약

인터넷 사용자 및 환경이 급격하게 증가함에 따라 더욱 다양하고 복잡한 네트워크 악성 트래픽이 발생하고 있으며 이를 분류하고 탐지하는 기술은 대단히 중요하다. 침입 탐지 시스템은 악성 트래픽을 분류 및 탐지하는 시스템으로 기존에는 페이로드 기반, 통계적 특징 기반 등 전통적인 방법들을 활용하였으며 최근에는 딥러닝 기반의 악성 트래픽 탐지 기술에 관하여 많은 연구가 수행되고 있다. 우리는 패킷 원본 데이터가 인접한 픽셀과의 공간 정보를 가지고 있는 일반적인 이미지와는 달리 1차원적인 데이터라는 가정하에 정사각 모형뿐만 아니라 다양한 모형으로 변환하고 학습시킨 모델들의 분류 성능을 비교하였다. 우리는 침입 탐지 시스템에서 가장 많이 사용되었던 공용 데이터 셋을 통해 비교 실험을 수행했으며 결과적으로 정사각 모형보다 더 높은 분류 성능을 나타내는 입력 모형이 있음을 확인하였다.

Ⅰ. 서 론

인터넷 사용자 및 환경이 급격하게 증가하고 더욱 다양하고 복잡한 응용 및 악성 트래픽이 발생함에 따라 네트워크 관리 기술의 중요성이 대두되고 있다. 네트워크 관리 기술 분야에서 응용 트래픽 분류 기술은 네트워크 유지 및 관리, 비정상 행위 탐지, 사용자 서비스 품질 보장 등을 목적으로 하며 크게 일반 응용 트래픽 분류와 침입 탐지 시스템에서의 악성 트래픽 탐지로 분류된다. 악성 트래픽은 디도스 공격, 봇넷 통신 등을 통해 인터넷 망을 교란하거나 특정 서버 및 호스트에 피해를 끼치기 위해 발생시키는 트래픽을 의미하며 악성 트래픽 탐지는 이러한 트래픽 정보로부터 특정한 패턴을 추출하고 이를 분석하여 분류하는 기술이다. 기존에 호스트 기반 또는 페이로드 기반의 침입 탐지 시스템이 널리 활용되었으나 네트워크 환경의 변화와 트래픽 암호화 기술의 도입 등으로 인해 악성 트래픽을 정확하게 분류하는 것이 불가능한 실정이며 이는 악성 트래픽을 더 정확하고 강인하게 분류하는 방법이 필요함을 시사한다.

본 논문은 1장 서론에 이어 2장에서 관련 연구를 제시하고 3장에서 실험 방법을 설명한다. 4장에서는 실험 결과를 설명하고 5장에서 결론 및 향후 연구를 제시한다.

Ⅱ. 관련 연구

악성 트래픽 분류의 대표적인 방법은 응용 트래픽 분류와 마찬가지로 포트 기반 분류, 페이로드 기반 분류, 머신러닝 기반 분류로 나눌 수 있다. 그러나, 전통적인 분류 방법의 한계점으로 인해 최근 딥러닝 기반 분류 방법에 대하여 많은 연구가 수행되고 있으며 이 중 CNN(Convolutional Neural Network) 기반 분류 방법이 가장 활발히 사용되고 있다.

CNN은 딥러닝에서 주로 이미지 또는 영상 데이터를 처리할 때 사용하며 이미지로부터 공간적 특징을 추출하는 컨볼루션이라는 처리 작업이 들어가는 신경망의 일종이다. CNN의 학습 과정은 이미지 형태의 데이터를 입력받아 컨볼루션(Convolution) 연산을 통해 공간적 특징을 추출하고 이를 압축하는 풀링(Pooling)을 반복하여 최종적으로 이미지의 특징을 지닌 피쳐맵을 생성하고 이를 완전 연결 계층(FC; Fully Connected Layer)에 입력하여 분류 결과를 출력한다. 일반적으로, 트래픽 분류 분야에선 네트워크 패킷의 원자료를 이미지 형태로 변환하고 입력하여 분류 결과를 생성하였으나 인접한 픽셀 간 유의미한 관계정보가 담겨 있는 일반적인 이미지와는 달리 패킷은 인접한 데이터와의 관계정보가 담겨 있다고 보기 어렵다. [1][2]는 패킷에서 784바이트를 추출하고 이를 784*1로 변환한 이미지를 입력한 1차원 CNN 기반 모델과 28*28 이미지를 입력한 2차원 CNN 기반 모델의 분류 성능을 비교하였으며 1차원 CNN 모델이 분류 성능이 좋다고 보고하였다. 특히, [3]은 모형을 정사각 모형 및 선 모형뿐만 아니라 다양하게 변형된 모형을 입력한 분류 모델을 생성하여 각 입력 모형에 따른 분류 결과를 비교하였으며 그 결과 정사각 모형 또는 선 모형이 아닌 49*16 모형이 가장 분류 성능이 좋았다고 보고하였다.

Ⅲ. 실험

본 장은 악성 트래픽 데이터를 CNN 기반 모델에 입력할 때 최적화된 입력 모형이 무엇인지 결정하고 각 모형에 따른 모델의 분류 성능을 비교하기 위한 실험을 설계한다.

A. 데이터셋

데이터셋은 악성 트래픽 분류 연구에 자주 사용되는 “DARPA1998” [4]을 사용하였으며 이를 전처리하여 다양한 입력 모형으로 변환하였다. 입력 모형은 784바이트로 만들 수 있는 직사각형 모형의 모든 조합을 고려하였으며 총 14개의 모형을 생성하였다. 데이터셋의 요약 정보는 표 1과 같다.

본 논문은 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과 (2021RIS-004)이고 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구(NRF-2021S1A5C2A03097574)임

표 1. DARPA1998 데이터셋 요약

유형	플로우 개수 (비율)			
	학습		테스트	
Normal	49,777	(44.0%)	54,452	(73.2%)
DoS	36,561	(32.3%)	15,048	(20.2%)
R2L	9,141	(1.71%)	4,134	(5.56%)
R2R	104	(0.09%)	20	(0.02%)
Probe	24,677	(21.8%)	634	(0.85%)

B. 분류모델

분류 모델은 [3]에서 사용한 단일모형-다중입력 모델을 사용하였으며 각 입력 모형에 따라 모델의 하이퍼 파라미터를 최적화하기 위하여 Python 기반 딥러닝 라이브러리인 Keras의 Hyperband Tuner를 사용하여 각 모형에 적합한 모델 구조를 결정하였다. 모델 구조를 최적화하는 과정에선 컨볼루션 레이어와 풀링 레이어의 필터 크기 및 풀링 크기만 변경하였으며 컨볼루션 필터 개수는 16, 덴스 레이어의 유닛 개수는 12로 고정하였다. 또한, 데이터에 심각한 불균형 문제가 있으므로 각 클래스 개수에 따라 가중치를 부여하여 공정하게 학습이 되도록 하였다.

C. 평가지표

분류 모델의 성능을 평가하기 위하여 3가지의 평가지표를 설정하였으며 이는 수식 1과 같다. ACC는 분류 모델의 전체적인 분류 성능을 나타내며 DR은 악성 트래픽에 대한 탐지 성능을 나타낸다. FAR은 악성 트래픽을 정상 트래픽이라고 오탐한 비율을 나타낸다.

$$\begin{aligned} Accuracy(ACC) &= \frac{TP+TN}{TP+FP+FN+TN} \\ DetectionRate(DR) &= \frac{TP}{TP+FN} \\ FalseAlramRate(FAR) &= \frac{FP}{FP+TN} \end{aligned} \quad (1)$$

Ⅲ. 실험결과

본 장에서는 모형별 최적화된 모델 구조와 입력 모형 별 분류 정확도를 비교한다.

표 3은 입력 모형별 분류 모델의 최적화된 필터 크기와 풀링 크기를 나타낸다. 필터 크기 및 풀링 크기 모두 모형이 정사각형에 가까워질수록 인접한 데이터를 포함하여 연산을 수행하며 필터 크기는 대부분 1차원적 특성을 주로 학습하는 것을 확인할 수 있다. 풀링 크기의 경우에도 대부분의 분류 모델이 풀링 크기를 (1, 1)를 채택하고 있으며 이는 풀링을 수행하지 않는다는 것을 의미한다.

표 4는 입력 모형별 분류 성능을 나타낸다. 정확도 측면과 오탐률에서 가장 높은 성능을 보이는 모델은 (2, 392) 모형을 채택한 모델이며 탐지율(재현율) 측면에선 (14, 56) 모형을 사용한 모델이 가장 높은 성능을 나타내었다. 반면에 (28, 28) 정사각 모형은 3가지 지표 측면에서 타 모형보다 비교적 낮은 성능을 보이는 것을 확인할 수 있다.

Ⅲ. 결론

본 논문은 패킷 입력 모형에 따른 각 모형별 최적화된 하이퍼 파라미터와 각 대응되는 분류 모델의 성능을 비교하였다. 결과적으로 분류 모델의 성능은 정사각 모형이 아닌 다른 모형에서 일반적으로 더 좋은 성능을 보였으며 하이퍼 파라미터 최적화 결과를 통해 분류 모델이 2차원 공간적 특징이 아닌 1차원 공간적 특징에 초점을 맞춘다는 것을 확인했다. 본 실험 결과는 표현학습을 기반으로 악성 트래픽을 분류하는 모델에 성능향상을 가져다줄 것으로 기대되며 우리는 향후 연구로 최신 연구 모델들에 본 방법을 적용 및 실험하고 결과를 비교할 예정이다.

표 3. 입력 모형별 최적화된 모델 구조

입력 모형	필터 크기	풀링 크기
(784,1)	(5,1)	(1,1)
(392,2)	(3,1)	(1,1)
(196,4)	(4,1)	(1,1)
(112,7)	(2,1)	(1,1)
(98 ,8)	(2,1)	(1,2)
(56,14)	(2,2)	(1,2)
(49,16)	(2,2)	(1,2)
(28,28)	(2,3)	(2,2)
(16,49)	(1,2)	(1,2)
(14,56)	(2,1)	(1,1)
(8 ,98)	(1,7)	(1,1)
(7,112)	(1,7)	(1,1)
(4,196)	(1,7)	(1,1)
(2,392)	(1,8)	(1,1)

표 4. 입력 모형별 분류 성능

입력 모형	ACC	DR	FAR
(784,1)	99.02	96.59	0.037
(392,2)	99.19	96.34	0.040
(196,4)	98.95	96.90	0.053
(112,7)	98.43	96.01	0.057
(98 ,8)	98.89	97.01	0.055
(56,14)	99.11	97.14	0.038
(49,16)	99.08	96.81	0.044
(28,28)	98.63	96.30	0.083
(16,49)	99.17	97.10	0.060
(14,56)	99.34	**97.43**	0.052
(8 ,98)	98.95	96.18	0.095
(7,112)	98.71	96.99	0.060
(4,196)	99.02	97.04	0.046
(2,392)	**99.40**	97.24	**0.021**

참 고 문 헌

[1] L. Xu, X. Zhou, Y. Ren, and Y. Qin, “A Traffic Classification Method Based on Packet Transport Layer Payload by Ensemble Learning,” in 2019 IEEE Symposium on Computers and Communications (ISCC), Jun. 2019, pp. 1 - 6.

[2] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, “End-to-end encrypted traffic classification with one-dimensional convolution neural networks,” in 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Jul. 2017, pp. 43 - 48.

[3] Ui-Jun et al, “Multi-Shape CNN based Application Traffic Classification” in KNOM 2022 Conference, May 2022, pp. 73 - 76

[4] R. P. Lippmann et al., “Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation,” in Proceedings DARPA Information Survivability Conference and Exposition. DISCEX’00, Jan. 2000, vol. 2, pp. 12 - 26 vol.2.