시그니처 기반 필터링과 2D-CNN을 활용한 하이브리드 악성 트래픽 탐지 기법

김지민, 장윤성, 박재원, 백의준, 김명섭
고려대학교
{illiard1209, brave1094, 2018270614, pb1069, tmskim* }@korea.ac.kr

A Hybrid Malware Traffic Detection Method Using Signature-based Filtering and 2D-CNN

Ji-Min Kim, Yoon-Seong Jang, Jae-Won Park, Ui-Jun Baek, Myung-Sup Kim*
Korea Univ.

요 약

본 연구에서는 시그니처 기반 탐지와 딥러닝 기반 이상 탐지를 결합한 하이브리드 구조를 제안하였다. 딥러닝 모델은 정상 트래픽만을 학습한 2D-CNN 기반으로 구성되며, 라벨 스무딩과 Confidence 기반 판별 방식을 통해 Open-set 환경에서의 이상 트래픽 탐지를 수행한다. 테스트 단계에서는 시그니처 기반 탐지를 통해 사전에 필터링된 트래픽을 입력으로 사용하여, 알려지지 않은 이상 행위를 효과적으로 검출하도록 구성하였다. USTC-TFC2016 데이터셋으로 실험한 결과, Open-set이 포함된 테스트 데이터셋에서 99~100%의 높은 탐지율을 달성하여 제안된 구조의 효과성을 입증하였다.

Ⅰ. 서 론

현대의 네트워크 환경은 고도화된 공격 기법과 함께 지속적으로 진화하고 있으며, 이에 따라 악성 트래픽 탐지 시스템의 중요성이 더욱 커지고 있다. 특히 고속 네트워크에서는 알려진 공격뿐 아니라 이전에 관찰되지 않은 제로데이(Zero-day) 공격까지 탐지할 수 있는 능력이 요구된다. 이러한 보안 요구에 대응하여 다양한 탐지 기법들이 개발되었으며, 각 방식은 고유한 장단점을 가진다.

초기에는 시그니처 기반 탐지(Signature-based Detection)[1]가 주류를 이루며 고속 탐지와 낮은 오탐율을 달성했으나, 새로운 공격과 암호화된 변종에 취약한 한계가 있었다. 이를 보완하고자 통계적 특징 기반의 머신러닝 기법[2]이 도입되었으나, 수작업 특징 설계 및 일반화 성능의 한계가 존재했다. 최근에는 CNN[3], RNN 등 딥러닝 기반 탐지가 부상하며, 원시 트래픽 또는 이미지로 변환된 데이터를 모델에 직접 입력하여 자동화된 특징 학습과 변종 탐지에 강점을 보이고 있다.

본 연구는 이러한 흐름 속에서 시그니처 기반 탐지와 2D-CNN 기반 이상 탐지를 결합한 하이브리드 탐지 시스템을 제안한다. 라벨 스무딩과 Confidence 기반 판별 기법을 통해 Open-set 환경에서도 효과적인 이상 탐지가 가능하며, 이는 향후 미지 클래스 대응과 라벨 불확실성에 강인한 탐지 시스템 구축에 기여할 수 있다.

Ⅱ. 본론

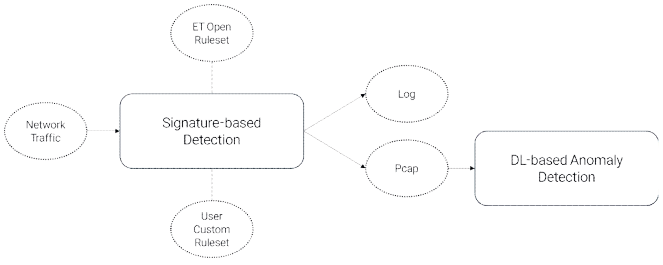

그림 1 하이브리드 악성 트래픽 탐지 모델 구조도

2.1 시그니처 기반 필터링 설계 및 구현

본 연구의 첫 번째 탐지 단계는 시그니처 기반의 필터링 엔진으로 ,사전에 정의된 룰셋을 이용하여 악성 트래픽을 실시간으로 탐지하고 제거한다. 이를 위해 본 연구에서는 Suricata를 탐지 엔진으로 사용하며, ET Open 룰셋을 활용하였다.

Suricata는 고속 패킷 분석 및 서명 기반 탐지 기능을 제공하는 오픈소스 IDS/IPS 시스템으로, 네트워크 트래픽을 실시간으로 처리하면서 룰셋에 기반한 탐지를 수행할 수 있다. 본 시스템에서는 Suricata가 PCAP 파일에 대해 탐지를 수행하며, 탐지된 트래픽은 로그를 통해 추출되어 이후 단계에서 제거 대상이 된다.

	Attack	Clean	Total	Ratio
Benign	-	-	-	
BitTorrent	0	7517	7517	0.00
FTP	2143	101037	103180	0.02
Facetime	0	6000	6000	0.00
Gmail	628	8629	9257	0.07
MySQL	0	86089	86089	0.00
Outlook	0	7524	7524	0.00
SMB	0	32661	32661	0.00
Skype	0	6321	6321	0.00
Weibo	0	24953	24953	0.00
WOW	0	7883	7883	0.00
Malware	-	-	-	-
Cridex	30025	38221	68246	0.44
Geodo	15311	32822	48133	0.32
Htbot	2045	6325	8370	0.24
Miuref	7628	13078	20706	0.37
Neris	1312	36569	37881	0.03
Nsis-ay	143	6061	6204	0.02
Shifu	104	9631	9735	0.01
Tinba	0	8503	8503	0.00
Virut	1887	35441	37328	0.05
Zeus	7404	10881	18285	0.40

표 1 Suricata로 필터링된 USTC-TFC2016 데이터셋 클래스 별 세션 수 및 비율

2.1.1 룰셋 구성

사용된 룰셋은 Emerging Threats에서 제공하는 오픈 시그니처 규칙인 ET Open 룰셋을 기반으로 하며, 다양한 알려진 악성 행위를 탐지할 수 있다.

2.1.2 트래픽 필터링 이상 방식

Suricata는 트래픽을 세션 단위 또는 패킷 단위로 분석하고, 조건에 맞는 트래픽에 대해 alert 로그를 생성한다. 본 연구에서는 Suricata의 eve.json 또는 fast.log 파일에서 탐지된 트래픽 정보를 기반으로, 탐지된 트래픽을 식별 후 필터링하였다. 탐지되지 않은 나머지 트래픽은 CNN탐지 단계로 전달된다.

2.2 CNN 기반 이상 탐지 구조

시그니처 기반 필터링을 통과한 트래픽은 알려진 공격으로 탐지되지 않은 세션으로 구성되어 있으며, 이 중에는 변종 악성 트래픽이나 제로데이 공격이 포함될 가능성이 존재한다. 본 연구의 두 번째 단계에서는 이러한 트래픽에 대해 2D Convolutional Neural Network(CNN) 기반의 이상 탐지 모델을 적용하여 정상 여부를 분류한다.

2.2.1 이미지 기반 트래픽 표현

딥러닝 모델의 입력으로 사용하기 위해, 트래픽 세션은 784 바이트(28x28)의 고정 크기 바이트 스트림으로 추출되어 흑백 PNG 이미지로 변환된다. 이는 패킷 페이로드의 내용을 시각적 패턴으로 표현하여, CNN이 공간적 특성을 학습할 수 있도록 한다. 본 연구에서는 모든 입력 데이터를 단일 채널 grayscale 이미지로 처리한다.

2.2.2 CNN모델 아키텍처

시그니처 기반 필터링을 통과한 트래픽은 알려지지 않은 공격이나 이상 징후를 포함할 수 있다. 이를 탐지하기 위해 본 연구는 Open-set 환경에 대응하는 2D CNN 기반 이상 탐지 모델을 적용한다. 학습 데이터는 정상 세션만으로 구성되며, 테스트 시에는 필터링된 트래픽을 입력으로 하여 모델의 Confidence 값을 기반으로 정상/비정상을 분류한다.

입력은 각 세션의 패킷 페이로드에서 최대 784바이트를 추출하여 28×28 크기의 grayscale 이미지로 변환한다. CNN 모델은 두 개의 합성곱 및 풀링 계층과 완전 연결층으로 구성되며, Label Smoothing 기법을 적용하여 학습 시 모델의 과신(overconfidence)을 억제한다. 이 구조는 미지 클래스에 대한 일반화 성능을 향상시키고, Confidence 기반 탐지를 가능하게 한다.

2.3 데이터 전처리 및 흐름

본 연구에서는 CNN 기반 이상 탐지를 위해 네트워크 트래픽을 이미지 형태로 변환하는 전처리 과정을 수행하였다. 먼저, 원본 PCAP 파일을 5-tuple 기반으로 세션 단위 분할한 뒤, 각 세션의 패킷 페이로드에서 이더넷 프레임만 제거한 바이트 스트림을 추출한다. 이후 앞부분의 최대 784 바이트를 28×28 크기의 grayscale 이미지로 변환하여 PNG로 저장하고, PyTorch 학습을 위해 Numpy 배열로 전환한다.

학습 데이터는 정상 트래픽만을 사용하며, 테스트 데이터는 시그니처 기반 필터링을 통과한 트래픽으로 구성된다. 이 구조는 CNN 모델이 정상 패턴만을 학습하도록 하여, 이후 Confidence 값 기반의 이상 탐지를 수행할 수 있도록 설계되었다.

2.4 실험 및 평가

Label Smoothing 미적용 시 일부 클래스에서 탐지율이 15~30% 수준으로 저조하게 나타났으며, 이는 모델이 학습되지 않은 이상 클래스에 대해 높은 confidence로 잘못 예측하는 경향을 보였기 때문이다. 반면, Label Smoothing을 적용한 모델은 모든 클래스에서 탐지율이 평균적으로 99% 이상으로 상승하였고, 특정 클래스에서는 100%의 탐지율을 기록하였다. 특히, 탐지율이 낮았던 Class 13 (Miuref) 의 경우 15.12% → 83.62%로 큰 개선이 있었으며, 전반적으로 Open-set 탐지 성능이 안정적으로 향상되었음을 확인할 수 있었다.

또한, 전체 테스트 데이터에 대해 Confidence 기반 이상 탐지를 수행한 결과, 이상으로 탐지된 샘플 중 대부분이 정상 클래스가 아닌 샘플에 해당하였으며, 탐지율(True Positive Rate)은 99.9%, 오탐율(False Positive Rate)은 0.1% 이하로 나타났다. 이러한 결과는 Label Smoothing이 모델의 일반화 성능을 높이고, Open-set 상황에서의 신뢰도 기반 탐지에 효과적으로 기여함을 시사한다.

Class	Detection Rate (Before)	Detection Rate (After)
Cridex	28.61	100
Geodo	24.12	99.72
Htbot	32.14	99.68
Miuref	15.12	83.62
Neris	24.74	99.87
Nsis-ay	34.78	99.09
Shifu	79.67	99.67
Tinba	21.15	99.38
Virut	19.83	99.96
Zeus	20.49	99.94

표 2 Label Smoothing 적용 전후 탐지율 차이

Ⅲ. 결론

본 연구는 시그니처 기반 탐지와 2D CNN 기반 이상 탐지를 결합한 하이브리드 악성 트래픽 탐지 구조를 제안하였다. 특히, CNN 모델에 Label Smoothing 기법을 적용함으로써 모델의 과신(overconfidence)을 완화하고, Open-set 환경에서의 이상 탐지 성능을 효과적으로 향상시킬 수 있음을 실험을 통해 확인하였다. 최대 confidence 기반 임계값 판별 방식을 통해 별도의 이상 클래스 학습 없이도 제로데이 및 변종 공격을 높은 정확도로 탐지할 수 있었으며, 향후에는 다양한 정규화 기법 및 데이터셋에 대한 일반화 가능성을 추가적으로 검증할 예정이다.

ACKNOWLEDGMENT

본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2023-00230661, 하이브리드 양자키분배 방법 및 망 관리 기술 표준개발)과 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원 (P0024177, 2023년 지역혁신클러스터육성)을 받아 수행된 연구임

참 고 문 헌

[1] Denning, D. E. (1986). An intrusion-detection model. *Proceedings of the 1986 IEEE Symposium on Security and Privacy*, 118 - 131. IEEE.

[2] Heberlein, L. T., Dias, G. V., Levitt, K. N., Mukherjee, B., Wood, J., & Wolber, D. (1990). A network security monitor. *Proceedings of the 1990 IEEE Symposium on Security and Privacy*, 296 - 304. IEEE.

[3] Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware traffic classification using convolutional neural network for representation learning. In *Proceedings of the 2017 International Conference on Information Networking (ICOIN)* (pp. 712 - 717). IEEE.