합성곱 신경망(Convolution Neural Network)를 이용한 악성코드 탐지 방안 연구

최신형
고려대학교 컴퓨터정보통신대학원
e-mail: shchoi7@korea.ac.kr

Convolution Neural Network for Malware Detection

Sin-Hyung Choi
Graduate School of Computer & Information Technology, Korea University

요 약

새롭게 변형되는 대규모 악성코드들을 신속하게 탐지하기 위하여 인공지능 딥러닝을 이용한 악성코드 탐지 기법을 제안한다. 대용량의 고차원 악성코드를 저차원의 이미지로 변환하고, 딥러닝 합성곱신경망(Convolution Neural Network)을 통해 이미지의 악성코드 패턴을 학습하고 분류하였다. 본 논문에서는 악성코드 분류 모델의 성능을 검증하기 위하여 악성코드 종류별 분류 실험과 악성코드와 정상코드 분류 실험을 실시하였고 각각 97.6%, 87%의 정확도로 악성코드를 구별해 내었다. 본 논문에서 제안한 악성코드 탐지 모델은 차원 축소를 통해 10,868개(200GB)의 대규모 데이터에 대하여 10분 이내의 학습시간이 소요되어 새로운 악성코드 학습 및 대용량 악성코드 탐지를 신속하게 처리 가능함을 보였다.

1. 서론

최근 발생하는 인터넷상의 공격은 APT(Advanced Persistent Threat) 및 랜섬웨어(Ransomware) 기법을 활용한 지능형 공격 방식으로 공격 여부를 은닉하기 위해 다양하게 변형되어 기존의 시그니처 기반 Rule-set 방식으로는 악성코드를 탐지하기 어려워 빅데이터와 인공지능을 활용한 새로운 접근 방법이 필요하다.

악성코드 탐지를 위해서 버퍼오버플로우, 널 포인터 참조 등 널리 알려진 취약점 유형 정의 규칙(Rule)을 적용하는 정적 분석 도구(static analysis tool)나 다양한 입력값과 실행경로를 테스트하는 퍼징(fuzzing), 기호실행(Symbolic Excution) 방식이 있으나 탐지 대상이 제한적이고 높은 오탐율과 시간 소요 비용 등의 문제점으로 인해 불특정 다수의 웹사이트, 이메일 등을 통해 유포되는 대규모의 파일들에서 악성코드를 탐지하는데는 적용하기 어렵다. 이에 최근 이미지 인식, 텍스트 분류, 기계번역 등 다양한 패턴 분류 및 생성에 활용되고 있는 인공지능 딥러닝을 악성코드 탐지 및 취약점 분석에 적용하려는 노력들이 활발해졌다.

본 논문에서는 악성코드 바이너리 파일을 학습하여 악성코드 패턴을 인식하고 검출하는 딥러닝 기반의 악성코드 탐지 시스템을 제안한다. 특히 다양하게 변형된 악성코드들이 내포된 대규모 파일들에서 신속하고 효율적으로 악성코드를 탐지하기 위하여, 대용량 고차원의 바이너리 코드를 저차원의 이미지로 축소(24x24) 변환하고, Convolution Neural Network(CNN) 인공신경망을 이용하여 이미지 악성코드 패턴을 학습하고 분류하는 방법을 제안한다.

2. 관련 연구

기존의 악성코드 탐지 및 취약점 분석 방법은 시그니처 기반의 Rule-set 방식으로 바이너리 파일을 해싱하여 악성코드 데이터베이스에서 동일한 해싱값이 존재하는지 검색하는 방식이다. 이러한 시그니처 기반 탐지 방식은 악성코드의 일부분이 변형된 신종 파일은 해싱값이 달라져서 탐지할 수 없는 문제점이 있다.

이런 문제점을 해결하기 위하여 인공신경망이 악성코드를 스스로 학습하여 분류하는 기계학습 기반의 악성코드 탐지 방법의 연구가 진행되었다.[1] 특히 심층신경망을 이용하여 악성코드 특징을 자동으로 추출하는 딥러닝 방식의 연구가 활발하게 진행되고 있다.[2]

최근의 딥러닝 연구에서는 이미지 분류에 뛰어난 성능을 보이는 CNN 심층신경망 모델을 텍스트 분류에 적용하여 좋은 결과를 보이고 있다.[3] 이에 악성코드를 이미지화하여 CNN 심층신경망을 통해 분류하는 연구가 진행되었다.

선행 연구 모델은 입력으로 사용될 악성코드의 특징들을 시각화하기 위하여 바이너리 파일의 주소를 제외하고 행과 열을 재배열하고 바이트 값을 정수로 환산하여 1/10 크기로

이미지를 재조정 한 후 이를 CNN을 이용하여 분류하여 91.7%의 정확도로 악성코드 종류를 구별하였다.[4]

본 논문에서는 선행 연구 모델과 달리 대규모 악성코드 파일의 신속한 검출에 초점을 맞춰 계산 오버헤드가 발생하는 바이너리 입력 파일의 이미지 변환 작업을 생략하고 일괄적으로 저차원(24x24) 이미지로 축소 한 후 정규화하였고, 제안하는 악성코드 분류 모델이 악성코드 종류 분류뿐 아니라 악성코드와 정상코드를 분류할 수 있는지 binary prediction task 실험을 수행하였다.

3. 악성코드 탐지 기법 제안

3.1 악성코드 데이터 셋

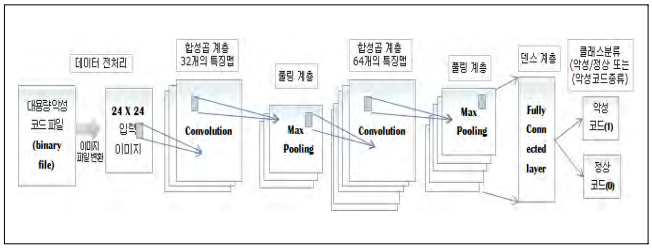
본 논문에서는 악성코드 종류 분류와 악성코드와 정상코드 분류 두 종류의 실험을 위하여 9종류의 악성코드 레벨을 분류한 Microsoft malware classification 데이터셋과 악성코드와 정상코드를 분류한 정보보호학회 데이터 챌린지 악성코드 데이터셋을 이용하였다.

Microsoft malware classification 데이터셋은 9종류의 악성코드들로 분류된 총 10,868개의 바이너리 파일들로 각 파일은 32비트의 주소와 16바이트의 동일한 크기의 워드 행(row)로 이루어져 있고, 각 파일들의 크기는 15kb~140MB로 파일 길이가 서로 매우 다르며 대용량 파일들을 포함하여 총 200GB 용량이다.

정보보호학회 데이터 챌린지 악성코드 데이터셋은 총 7,500개의 다양한 크기(1~34,000KByte)의 바이너리 파일들로 악성코드 파일(class 1) 5,250개와 정상파일(class 0) 2,250개로 구성되어 있고, 파일크기는 별도로 가공되지 않아 열과 행의 길이가 모두 서로 다르고 총 6GB 용량이다.

3.2 악성코드 탐지 모델 구조

본 논문에서 제안하는 악성코드 탐지 모델 구조는 그림 1과 같이 대용량 고차원의 바이너리 코드를 저차원의 이미지로 축소(24x24) 변환하는 데이터 전처리 단계와 Convolution Neural Network(CNN) 신경망을 이용하여 이미지 악성코드 패턴을 학습하고 분류하는 두 단계로 이루어져 있다. 데이터 전처리 단계에서는 새로운 악성코드를 신속하게 학습하여 빠르게 탐지할 수 있도록 고차원의 악성코드를 저차원(24x24)의 RGB 이미지로 축소 변환한 후 이미지를 정규화한다. 악성코드 분류 학습 단계에서는 이미지 분류에 좋은 성능을 보이는 합성곱 신경망(CNN)을 두 계층으로 구성한다. 첫 번째 합성곱 단계에서 32개의 특징을 학습하고 ReLU 활성화를 거쳐 최댓값 풀링을 진행한다. 두 번째 합성곱 단계에서는 64개의 특징을 학습하고, 마지막 단계에서는 평탄화 후 512개의 신경망을 갖는 덴스(Dense) 네트워크를 거쳐 소프트맥스(Softmax) 분류기를 통해 악성코드를 분류한다. 악성코드 분류 클래스는 악성코드 여부 판단 시 클래스를 악성코드(1), 정상코드(0) 2개로 나누었고, 악성코드 종류 별 분류 시에는 클래스를 9개로 분류하였다.



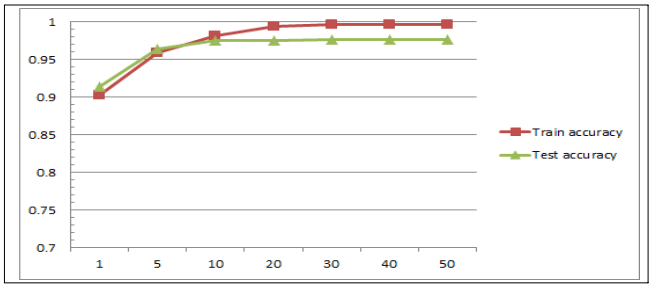
(그림 1) 악성코드 분류모델 구조

4. 실험 결과

실험은 동일 모델을 이용하여 9가지 종류의 악성코드 종류(Class) 분류와 악성코드/정상코드 바이너리 분류 두 가지 방법으로 진행하였다.

4.1 악성코드 클래스 분류 실험 결과

악성코드 Class 분류는 Microsoft malware classification 데이터셋으로 8,151개의 학습 데이터와 2,717개의 검증 데이터를 사용하였다. 9종류의 악성코드 클래스 분류 실험 결과 50회 반복에서 Train Accuracy 99.7%, Test Accuracy 97.6%의 탐지율을 보였다.

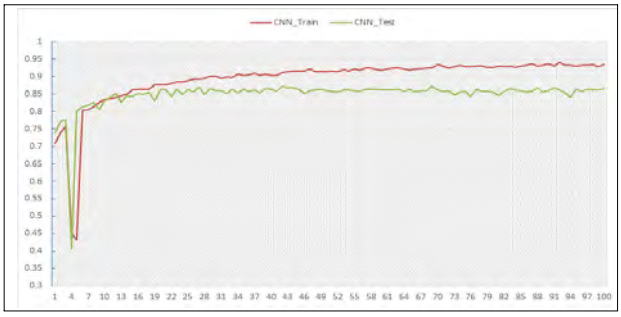


(그림 2) 실험 반복회수 별 모델의 학습 정확도 검증

4.2 악성코드와 정상코드 바이너리 분류 실험 결과

악성코드와 정상코드 분류 실험은 정보보호학회 데이터 챌린지 악성코드 데이터셋 총 7,500개의 파일 중에 학습 데이터로 5,623개, 검증 데이터로 1,875개를 사용하였고, 데이터 크기가 너무 작아 이미지 변환(24x24)이 어려운 2byte 이하 파일은 데이터셋에서 제외하였다.

바이너리 분류 실험 결과 100회 반복에서 Train Accuracy 94%, Test Accuracy 87% 탐지율을 보였다.



(그림 2) 실험 반복회수 별 모델의 학습 정확도 검증

1,875개 검증 데이터로 모델을 평가 한 결과 정확도(accuracy)는 87%, 재현율(recall)은 90%로 나타났다.

$$\begin{aligned} \text{Accuracy} &= (TP + TN) / (TP + TN + FP + FN) = 0.871467(87\%) \\ \text{Recall} &= tp / (tp + fn) = 1186 / 1319 = 0.899166(90\%) \end{aligned} \quad (1)$$

<표 1> 악성코드 탐지 모델 정확도

카테고리		실제 결과	
		Malware (1319개)	Benign (556개)
실험 결과	Malware (1294개)	TP 1186	FP 108
	Benign (581개)	FN 133	TN 448

5. 결론 및 향후 연구 방향

본 논문에서 제안하는 CNN을 이용한 악성코드 탐지 모델은 약 1만여개의 바이너리 파일을 24x24 크기의 이미지로 변환하여 데이터를 학습하는데 5분 이내의 시간이 소요되어, 새로운 악성코드 학습 및 대용량 악성코드 탐지를 신속하게 처리 가능함을 보였다.

악성코드 클래스 분류는 97.6%의 높은 탐지율을 보였고, 악성코드와 정상코드 분류에서는 87%로 좀더 낮은 성능을 보였다. 악성코드와 정상코드 비율을 일정하게 조정하여 데이터를 추가 확보하고, 파라미터 등을 조정한다면 성능 향상이 가능할 것으로 예상되어 본 모델을 활용하여 인터넷을 통해 유입되는 대량 데이터의 악성코드 탐지 및 악성코드 종류 분류를 통한 신속한 대응이 가능할 것이다.

추후 연구에서는 딥러닝을 통한 소스코드 바이너리 파일의 취약점 탐지 및 자동 패치 방법을 연구하여 소프트웨어 보안을 향상하고자 한다.

참고문헌

[1] Gustavo Grieco, Guillermo Luis Grinblat, Lucas Uzal, Sanjay Rawat, Josselin Feist, Laurent Mounier, "Toward large-scale vulnerability discovery using Machine Learning", 2015.

[2] Aragorn Tseng, YunChun Chen, "Deep Learning for Ransomware Detection", IEICE-IA2016-46

[3] 김혜정, 윤은주, "악성코드로부터 빅데이터를 보호하기 위한 이미지 기반의 인공지능 딥러닝 기법", IEIE Vol.54, No2. 2017.

[4] Kim, Yoon. “Convolutional neural networks for sentence classification”, arXiv preprintarXiv:1408.5882. 2014.