난독화된 악성코드 탐지를 위한 딥러닝 기반 다중 모델 앙상블 기법

한성훈, 한승우, 임찬미, 이광재*

*상명대학교

{201821266, 201821268, 201921267}@sangmyung.kr, *begleam@smu.ac.kr

A Deep Learning-based Multi-model Ensemble Method for Detecting Obfuscated Malware

Seong-Hun Han, Seung-Woo Han, Chan-Mi Lim, Kwangjae Lee*

*Sangmyung Univ.

요 약

최근 신종 악성코드는 2022년 기준으로 일 평균 약 27만 개 이상 제작되어 유포되는 실정이다. 하지만 유포되는 악성코드 대부분이 역공학 분석을 어렵게 하도록 난독화를 하므로 악성코드를 탐지하기 쉽지 않다. 따라서 본 논문에서는 난독화된 악성코드 탐지를 위해 동적분석 후, CNN+LSTM 앙상블 모델을 제안한다. 제안한 모델의 데이터셋은 Obfuscated-MalMem2022을 2차원 이미지로 변환하여 사용하였고, 제안한 모델을 통해 이진분류 및 다중분류를 시도하였다. 그 결과 이진 및 다중분류의 정확도는 각각 약 100%, 약 85%를 얻었다. 이는 제안된 방식이 난독화된 악성코드임에도 이진 및 다중분류에서 모두 우수함을 보여준다.

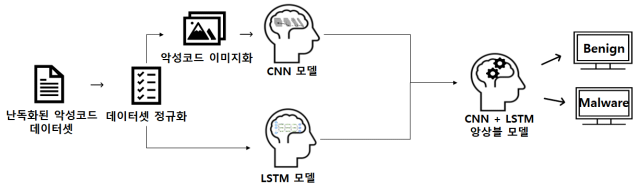
Ⅰ. 서 론

최근 많은 분야에서 소프트웨어가 활용되고 관련 산업 규모가 커지면서 악성코드로 악의적인 이득을 보고자 하는 시도도 늘고 있다. 신종 악성코드는 2022년 기준으로 일 평균 약 27만 개 이상 제작되어 유포되고 있다[1]. 또한 유포되는 신변종 악성코드들은 코드 삽입, 난독화, 문자열 암호화 등 다양한 악성코드 난독화 기술들을 가지고 있다. 악성코드 판별에 관한 선행 연구에서는 연산부호(opcode)의 빈도수를 기반으로 악성코드를 탐지하는 연구, N-gram를 활용한 정적분석 기반 악성코드를 탐지하는 연구 등이 제안되었고 악성코드 판별에 높은 정확도를 가졌다[2]. 하지만 대부분의 악성코드는 역공학 분석을 어렵게 하도록 난독화를 하므로 정적분석을 통해 악성코드를 탐지하기 쉽지 않다. 이에 엔트로피 및 문자열 등의 난독화 특징에 기반한 딥러닝 분석 방법이 제안되었고 더 나아가 2차원 배열로 특징을 재가공하는 기법을 적용하여 약 90%의 정확도를 가짐을 보여주었다[1].

본 논문에서는 난독화된 악성코드 탐지를 위해 정상 및 악성코드를 동적분석을 이용하여 특징을 추출하고 2차원 이미지 데이터셋으로 변환한다. 그리고 탐지 정확성을 높이기 위해서 2차원 데이터 분류에 강력한 딥러닝 알고리즘인 CNN과 LSTM을 사용하고, 두 알고리즘의 장점을 더욱 높일 수 있는 소프트 보팅 기법을 활용한 앙상블 모델을 제안한다.

Ⅱ. 본론

본 논문에서 제안하는 난독화된 악성코드 탐지 시스템의 개념도는 그림 1과 같다. 정상 코드와 악성코드는 실행하는 프로세스나 핸들링하는 프로그램들이 있는데, 여기서 악성코드는 정상 코드와 다르게 실행하고 핸들링하는 프로그램들의 개수가 다르다는 특징이 있다. 따라서 이러한 특징들을 이용하여 정상 코드와 악성코드를 분류하는 딥러닝을 제안한다. 제안하는 모델은 캐나다 사이버보안 연구소에서 제공하는 난독화된 악성코드 데이터셋인 Obfuscated-MalMem2022을 이용하여 학습을 진행한다[3]. 이때 딥러닝 알고리즘은 CNN과 LSTM 분류 모델을 설계한 뒤, 소프트 보팅 기법으로 앙상블하여 분류 모델을 설계한다. 마지막으로, 학습이 완료된 딥러닝 모델로 검사가 필요한 새로운 프로그램이 정상인지 악성인지 판별한다.



[그림 1] 제안하는 CNN+LSTM 앙상블 모델 구성도

제안하는 모델은 Obfuscated-MalMem2022 데이터셋을 사용한다. 이 데이터셋은 메모리에 상주한 데이터를 동적 분석 도구인 Volatility Memory Anal yzer(VolMemLyzer) 모듈을 사용하여 생성한다[4]. 데이터셋은 표 1에 제시된 필드 그룹으로 구성되어있다. 그리고 그 정보를 기반으로 3개의 카테고리와 15개의 악성코드 그룹으로 분류한다. 세부적으로 Ransomware 카테고리에는 Conti, MAZE, Pysa, Ako, Shade 악성코드 그룹, Spyware 카테고리에는 180Solutions, Coolwebsearch, Gator, Transponder, TIBS 악성코드 그룹, Trojan 카테고리에는 Reconyc, scar, Refroso, Emotet, Zeus 악성코드 그룹으로 구성된다. 본 논문에서 사용된 데이터셋은 데이터 편차가 크므로 특성 정보의 손실을 최소화하기 위해 16bit 픽셀 이미지를 사용하였다. 그리고 총 55개의 특성 정보를 가지므로 8*8 크기로 2차원 배열을 만든 후 부족한 공간에 0으로 패딩하였다. 이후 본 논문의 실험에 사용된 ResNet의 Input Size(224*224*3)로 이미지의 크기를 재조정 과정을 거쳐 데이터 전처리 단계를 완료한다.

[표 1] Obfuscated-MalMem2022 데이터셋의 필드 그룹

이름	설명
pslist	현재 실행 중인 프로세스 목록
dlllist	프로세스의 로드된 DLL 목록
handles	프로세스의 열린 handle 목록
ldrmodule	명령을 실행하여 로드되거나 인젝션 된 DLL 파일 목록
malfind	VAD 태그 및 페이지 권한과 같은 특성에 따라 사용자 모드 메모리에서 숨겨지거나 삽입된 코드/DLL 목록
callback	다양한 소스들로부터 설치된 callback 목록
psxview	메모리 덤프 파일을 생성한 윈도 시스템에서 은폐 기능으로 동작하는 프로세스 목록
svcscan	메모리 이미지에 등록된 서비스 목록

본 논문에서 제안하는 최종 모델인 CNN+LSTM 앙상블 모델을 생성하기 위해 먼저 생성되어야 하는 모델인 CNN과 LSTM을 설계한 뒤, 학습을 시켜 모델을 생성한다. 그 이후 두 모델을 가지고 소프트 보팅 기법을 이용하여 앙상블한 최종 모델을 생성한다. 먼저, CNN 모델은 imagenet에 미리 훈련된 ResNet50 모델을 사용하였다[5]. Dropout layer와 Dense layer는 따로 출력하기 위해서 레이어를 직접 설정했다. 사용된 모델의 Dropout layer는 0.25이고 출력 뉴런은 256이다. 그리고 ReLU 활성화 함수와 Adam optimizer를 사용하였다. 다음으로, LSTM 모델은 데이터셋 정규화를 거친 2차원 시퀀스 데이터를 3차원 배열로 변환하여 입력한다. 모델의 설정값은 Sequence model이고, units 수는 128, Dropout layer를 0.20이다. 그리고 sigmoid 활성화 함수와 Adam optimizer를 사용하였다. 마지막으로, 소프트 보팅은 각각 학습된 모델들이 각 카테고리일 가능성을 가지고 평균을 내어 가장 높은 확률이 나온 카테고리를 투표로 선정하는 알고리즘이다[6]. 제안한 모델은 미리 학습한 CNN과 LSTM 알고리즘을 사용하여 앙상블하기 때문에 소프트 보팅 기법을 선정하였다.

Ⅲ. 실험 결과

개발 및 실험에 사용한 환경은 다음과 같다. OS는 Windows 10 Pro 64bit를 사용하였고, CPU는 i9-12900K이며, GPU는 NVIDIA GeForce TRX 3090에 RAM은 64GB이다. 그리고 딥러닝 알고리즘 수행을 위해 Tensorflow, Python 3, Sklearn을 사용하였다.

먼저 정상 코드와 악성코드를 분류하는 이진분류 실험은 다음과 같이 수행되었다. 데이터의 수는 58,596개이며, 정상 코드 29,298개, 악성코드 29,298개로 구성되었다. 학습 데이터와 테스트 데이터의 비율은 7:3으로 통일하였으며, K겹 교차검증에서의 K값은 5로 설정하여, 5개의 데이터셋으로 분할 하여 진행하였다. 해당 모델 및 비교 모델의 성능평가는 표 2와 같다. 본 논문에서 제안하는 CNN+LSTM 앙상블 모델이 정확도(Accuracy) 99.99%를 기록하며, 모든 지표에서 만점에 가까운 성능을 보여주었다. 비교 모델들의 정확도 또한 GAN의 정확도가 약 97%인 것을 제외하면 모두 99%가 넘는 높은 정확도를 보여주었지만, CNN+LSTM 앙상블 모델의 정확도가 가장 높은 것을 확인할 수 있다.

다음으로 다중분류 실험은 4개의 분류 클래스를 지정하고 수행하였으며

[표 2] 제안 및 비교 모델의 이진분류 성능평가

Algorithm	Precision	Recall	F1-Score	Accuracy	AUC-Score
CNN	99.65	99.64	99.64	99.65	99.64
LSTM	99.76	99.83	99.80	99.80	99.80
GAN	94.54	99.74	97.07	96.99	96.99
Random Forest	99.98	99.98	99.98	99.98	99.98
CNN+LSTM Ensemble	100.0	99.98	99.99	99.99	99.99

클래스는 benign, ransomware, spyware, trojan이다. 실험 결과는 CNN과 GRU가 각각 72%와 73%의 정확도를 보이며 다른 모델들에 비해 낮은 정확도를 도출했다. 또한 LSTM은 82%의 정확도를 보였으며, Random Forest와 CNN+LSTM 앙상블은 각각 87%와 85%의 높은 정확도를 보이는 것을 확인할 수 있었다. 표 3은 CNN+LSTM 앙상블 모델의 다중분류 성능평가표이다. benign 클래스를 제외한 평균 정밀도(Precision), 재현율(Recall), F1-Score를 비교해보면 spyware와 trojan 클래스는 70% 이상의 정확도를 보여주었지만, ransomware 클래스의 분류에서는 60%대의 정확도를 보이는 것을 확인할 수 있었다.

[표 3] CNN+LSTM 앙상블 모델의 다중분류 성능평가

Class	Precision	Recall	F1-Score	support
benign	99.85	99.92	99.92	5860
ransomware	66.79	72.52	69.54	1958
spyware	74.21	73.80	74.00	2004
trojan	71.25	65.02	67.99	1898

Ⅳ. 결론

본 논문에서는 난독화된 악성코드를 탐지하기 위해 동적분석 데이터셋인 Obfuscated-MalMem2022을 사용하여, CNN+LSTM 앙상블 모델을 통해 이진분류 및 다중분류를 시도하였다. 이진분류에서는 CNN +LSTM 앙상블 모델이 정확도 약 100%를 기록하며, 다른 모델들과 비교했을 때 가장 좋은 성능을 보여주었다. 다중분류에서는 이진분류에서 가장 좋은 성능을 보인 CNN+LSTM 앙상블 모델은 약 85%의 정확도를 기록하였고, Random Forest가 약 87%로 기록하며 가장 좋은 성능을 기록하였다. 본 논문에서 사용한 데이터셋이 동적 분석을 한 뒤에 딥러닝을 통해 악성코드를 탐지하는 것이기 때문에 난독화된 악성코드임에도 이진분류 및 다중분류에서 모두 우수한 성능을 보여주었다. 하지만 오탐과 미탐을 최소화해야 하는 악성코드 탐지에 있어서 다중분류는 아직 어떠한 시스템에 적용할 수준은 아닌 것으로 판단되나, 이진분류에 있어서는 이번 실험 외의 난독화된 악성코드를 탐지하는데 유의미한 결과를 가져올 수 있을 것으로 기대한다.

참 고 문 헌

[1] S. B. Hwang *et al.*, "A Study on Two-dimensional Array-based Technology to Identify Obfuscatied Malware," *J. KIISE*, vol. 45, no. 8, pp. 769-777, 2018.

[2] S. M. Go *et al.*, "CNN-Based Malware Detection Using Opcode Frequency-Based Image," *J. Korea Inst. Inf. Secur. Cryptology*, vol. 32, no. 5, pp. 933-943, 2022.

[3] Hwang Y. C. and Mun H. J., "Detection Model based on Deeplearning through the Characteristics Image of Malware," *J. Convergence Inf. Technol.*, vol. 11, no. 11, pp. 137-142, 2021.

[4] T. Carrier *et al.*, "Detecting Obfuscated Malware using Memory Feature Engineering," in *ICISSP*, Feb. 2022, pp. 177-188.

[5] M. B. Hossain *et al.*, "Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images," *Infromatics in Medicine Unlocked*, vol. 30, 2022.

[6] B. U. Jeon, J. S. Kang, and K. Chung, "AutoML and CNN-based soft-voting ensemble classification model for road traffic emerging risk detection," *J. Convergence Inf. Technol.*, vol. 11, no. 7, pp. 14-20, 2021.