Journal of Convergence for Information Technology
Vol. 11. No. 11, pp. 137-142, 2021
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2021.11.11.137

악성코드의 특성 이미지화를 통한 딥러닝 기반의 탐지 모델

황윤철[1], 문형진[2*]
한남대학교 탈메이지 교양교육대학 조교수[1], 성결대학교 정보통신공학과 조교수[2]

Detection Model based on Deeplearning through the Characteristics Image of Malware

Yoon-Cheol Hwang[1], Hyung-Jin Mun[2*]
[1]Assistant Professor, Department of Talmage Liberal Arts College, Hannam University
[2]Assistant Professor, Department of Information & Communication Engineering, Sungkyul University

요 약 인터넷의 발달로 많은 편리와 이익을 얻었지만 반대로 지능화되는 악성코드로 인하여 사용자의 경제적, 사회적 피해를 주고 있다. 이를 탐지하고 방어하기 위해 대부분 시그니처 기반의 탐지나 방어 프로그램을 사용하지만 지능화된 악성코드의 변종을 막기에는 매우 어렵다. 따라서 본 논문에서는 쏟아져 나오는 지능화된 악성코드를 탐지하고 방어할 수 있는 모델을 제안한다. 제안 모델은 악성코드의 특성을 이미지화하여 딥러닝을 이용한 학습을 통해 만들어지며 새롭게 탐지된 악성코드와 악성코드 변종들은 이미지화를 수행한 다음 만들어진 모델에 적용하여 탐지한다. 제안된 모델을 사용하면 기존에 탐지되었던 악성코드와 더불어 유사한 변종도 대부분 탐지됨을 알 수 있다.

주제어 : 악성코드, 지능화, 악성코드 변종, 딥러닝, 탐지 모델

Abstract Although the internet has gained many conveniences and benefits, it is causing economic and social damage to users due to intelligent malware. Most of the signature-based anti-virus programs are used to detect and defend this, but it is insufficient to prevent malware variants becoming more intelligent. Therefore, we proposes a model that detects and defends the intelligent malware that is pouring out in the paper. The proposed model learns by imaging the characteristics of malware based on deeplearning, and detects newly detected malware variants using the learned model. It was shown that the proposed model detects not only the existing malware but also most of the variants that transform the existing malware.

Key Words : Malware, Intelligence, Malware Variants, Deeplearning, Detection Model

1. 서론

현재 우리 사회는 ICT의 급격한 발달로 세상에 존재하는 사람과 사람이, 사물과 사물이 연결되는 초연결 시대로 무수히 많은 정보들이 정보통신기기를 통하여 생성되고 처리되고 있다. 이렇게 생성되고 처리되는 데이터를 활용하는 데 있어 악성코드의 공격과 같은 사이버 보안 문제가 빈번하게 발생되어 사용자들에게 경제적 피해뿐만 아니라 사회적 문제를 야기하고 있다[1].

악성코드는 사용자의 의사와는 관계없이 악의적인 목적을 가지고 사용자들에게 해를 끼치기 위해 제작된 모든 실행 가능 소프트웨어이다. 최근 10년간 이러한 악의적인 목적을 가진 악성코드가 크게 증가하고 있는 추세이다. 악성코드를 활용한 공격을 탐지하고 방어하

*Corresponding Author : Hyung-Jin Mun(jinmun@gmail.com)
Received August 9, 2021
Revised September 4, 2021
Accepted November 20, 2021
Published November 28, 2021

기 위해서 현재 대부분 시그니처와 휴리스틱 기반의 바이러스 방지 프로그램을 사용하거나 주기적인 백업과 같은 기본적인 방법을 사용하고 있다. 하지만 완벽하게 악성코드를 활용한 지능적인 공격을 막는 데는 역부족인 상태이다. 이를 보다 정확하게 탐지하고 실시간으로 방어하는 방법이 절실히 요구되고 있다.

또한, 최근에 발생되는 사이버 위협들은 인공지능을 이용하여 더욱 복잡하고 다양하게 지능형 공격 방식으로 진화하고 있다. 해커와 같은 공격자들은 기존 사이버 위협을 더 다양하게 변형하여 신속하고, 은밀하고, 정확하게 진행하기 위해 인공지능을 활용하고 있다. 공격자들은 빠른 시간 안에 탐지되는 것을 피하기 위해 쉽고, 빠르게 유사한 변종 악성코드를 자동 생성한다. 뿐만 아니라 기존의 시스템을 우회하거나 취약점을 분석하는 해킹 자동화 도구를 통해 악성코드를 생성하여 공격하거나, 인공지능으로 진위를 가릴 수 없는 이메일이나 사이트, 동영상을 생성하여 피싱을 유도하는 기발한 악성코드를 이용하여 공격을 하고 있다[2].

그래서 본 논문에서는 빅데이터와 인공지능의 딥러닝 기법을 이용하여 APT(Advanced Persistent Threst)[3], 랜섬웨어 기법을 활용한 지능적 공격 방식을 포함한 악성코드의 특성을 학습시켜 악성코드의 패턴을 인식하고, 인식된 패턴을 이미지화 하는 방식의 딥러닝 기반 악성 코드 탐지 모델을 제안한다. 탐지 기법들의 주요 특징은 악성코드의 특성 정보를 인공지능의 딥러닝을 이용하여 학습시켜 이미지화하는 것이다[4]. 악성코드의 변종들은 주요 특성이 같기 때문에 이미지화하면 비슷한 이미지가 형성된다. 따라서 형성된 이미지를 활용하면 같은 종류의 새로운 방식의 악성코드를 탐지하고 방어하는 데 효과적이다[5].

본 논문의 구성은 다음과 같다. 2장에서는 딥러닝과 관련된 악성코드 추출 방법에 대해 살펴보고 3장에서는 딥러닝 이용한 악성코드 탐지 모델을 제안한다. 4장에서는 모델의 성능을 평가하고 끝으로 결론을 맺는다.

2. 관련연구

딥러닝이란 데이터로부터 기계가 학습하는 머신러닝의 기술의 일종으로 뉴런과 시냅스의 신경 네트워크 구조를 본떠서 심층 신경망 알고리즘을 모델링한 것으로 많은 데이터 속에서 패턴을 찾아내어 컴퓨터가 데이터를 스스로 분류하는 기술의 하나이다.

2.1 합성곱 신경망

합성곱 신경망인 CNN(Convolutional Neural Network)는 이미지 처리 분야에 많이 사용되는 딥러닝 알고리즘이다. CNN은 여러 개의 층으로 이루어져 있으며 입력 데이터의 특징 추출이 자동적으로 이루어지는 특징을 가진다. CNN을 통해 특징 추출을 활용하여 악성코드를 학습하고, 학습으로 패턴을 분류하고 정확한 탐지가 가능하다면 다양한 종류의 악성코드 감염으로부터 시간과 노력, 비용을 감소시킬 수 있다. 이를 통해 증가하는 변형된 사이버 보안 공격에 쉽게 대응할 수 있다[6].

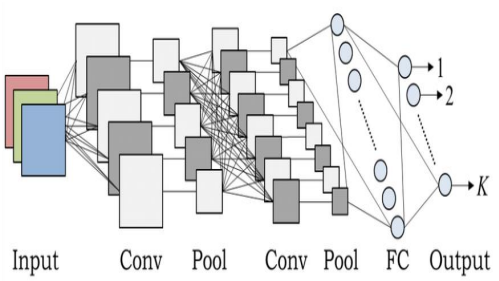

Fig. 1. CNN structure

심층 신경망(DNN)의 한 종류인 CNN은 Fig. 1과 같이 완전연결계층(Fully Connected Layer), 합성곱계층(Convolution Layer), 폴링계층(Pooling Layer)으로 구성되며. 영상 이미지와 같은 n차원 입력 데이터를 사용시 현재 계층의 뉴런을 이전 계층의 문제를 해결하기 위하여 CNN이 사용되는데 CNN에서 각 뉴런은 입력 뉴런의 일부분에만 연결된다. 완전연결계층에서는 입력 데이터와 가중치를 대응시키고, 합성곱계층에서는 필터를 통해 새로운 값을 갖게 하는 합성곱 연산이 이루어지며, 폴링계층에서는 대상 영역의 평균을 구해준다[7].

2.2 악성코드 탐지

딥러닝을 이용한 악성 코드 탐지는 크게 특징 데이터 추출하고 딥러닝을 이용해 학습을 진행하여 악성코드 탐지 모델을 생성하는 단계와 생성된 탐지 모델을 이용하여 악성코드를 탐지하는 탐지 단계로 이루어진다. 첫 번째 단계인 특징 데이터를 추출하는 과정에서는 특징을 추출하기 위해 악성코드를 실행하여 API를 추출하거나 정적분석을 통하여 opcode와 같은 어셈블

리 코드를 특징으로 추출하여 사용하기도 한다[8,9,10]. 다른 방법으로는 악성코드의 특징 데이터를 이미지 파일로 간주하여 이미지로 추출하기도 한다. 그리고 탐지 모델을 생성하기 위해 학습 과정에서는 추출한 여러 악성코드의 특징 데이터를 입력으로 딥러닝 모델에 적용하여 탐지 목적에 맞는 최적의 모델을 생성한다. 악성코드 탐지 단계에서는 딥러닝 모델을 이용하여 생성된 모델에 탐지 대상이 되는 악성코드를 입력하면 악성코드 존재를 신속하고 정확하게 판단할 수 있다[11].

본 논문에서 사용할 탐지 방법은 악성코드로부터 특징을 추출하여 악성코드를 이미지화 하는 것이다. 이 기법은 같은 부류에 속하는 악성코드의 이미지들은 유사하고 다른 부류에 속하는 악성코드의 이미지들과 구별된다는 것이다. 악성코드를 이미지로 시각화하는 것은 바이너리의 다른 세션들이 쉽게 구별될 수 있다는 장점이 있고, 악성코드 제작자들은 변종을 만들 때 기존 악성코드의 일부분만을 고쳐 변종을 생성함으로 전체적인 구조는 유지되기 때문에 작은 변화를 탐지하는데 유용하다. 결과적으로 변종에 속하는 악성코드의 이미지를 이 방법을 사용하면 같은 부류의 악성코드 이미지로 쉽게 분류하여 탐지할 수 있다.

3. 딥러닝 기반의 악성코드 탐지 모델

Fig. 2는 제안한 딥러닝 기반의 악성코드 탐지 모델을 가시화한 그림이다. 딥러닝을 이용한 악성 코드 탐지 모델은 크게 악성코드로부터 특징 데이터를 추출하는 단계와 추출한 데이터를 입력으로 딥러닝 모델에 적용하여 트레이닝을 하는 두 단계로 이루어진다. 그리고 침투 악성코드가 발견되면, 이를 이미지화시켜 딥러닝을 통해 학습된 탐지 모델을 이용하여 악성코드 여부를 판별한다.

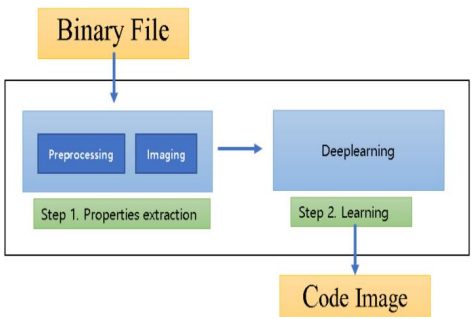

Fig. 2. Deeplearning-based malware detection model

제안 모델에서 악성코드의 특징을 추출하는 단계는 먼저 전처리를 걸쳐 악성코드를 이미지화하는 과정이 Fig. 2와 같이 진행된다. 전처리 과정에서는 악성코드 파일을 8bit 정수의 벡터로 읽어 2차원 배열로 변환한다. 그런 다음 이미지 과정을 거쳐 변환된 2차원 배열을 이미지로 시각화한다. 악성코드를 이미지로 시각화하는 것의 장점은 바이너리의 다른 섹션들이 쉽게 구별될 수 있고 변종에 속하는 악성코드 이미지는 같은 종류의 악성코드 이미지와 유사하고 다른 종류에 속하는 악성코드의 이미지와는 쉽게 구별된다.

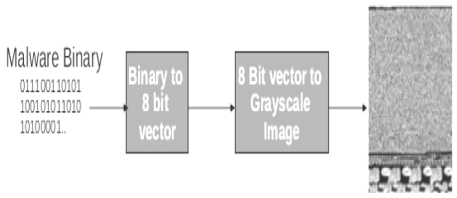

Fig. 3. Visualize malware as an image

Fig. 3은 일반적인 Dontove의 이미지 시각화 단계를 나타낸 것이다. 악성코드를 입력받아 8bit 바이너리를 생성하고 이것을 부호 없는 2차원 그레이스케일 배열을 생성한 후 생성된 배열을 출력하면 이미지가 시각화 된다[12].

악성코드의 특징이 이미지로 생성되면 두 번째 단계인 트레이닝이 진행된다. 트레이닝에서는 악성코드 특징을 추출하여 만든 이미지가 입력되면 딥러닝을 이용해 모델에서 사용할 가중치 값들과 연산을 통한 학습을 진행한 후 탐지에 적합한 최적의 모델을 생성한다.

딥러닝 학습으로 생성된 모델을 이용해 새롭게 탐지된 악성코드를 판별하는 과정은 Fig. 4와 같다.

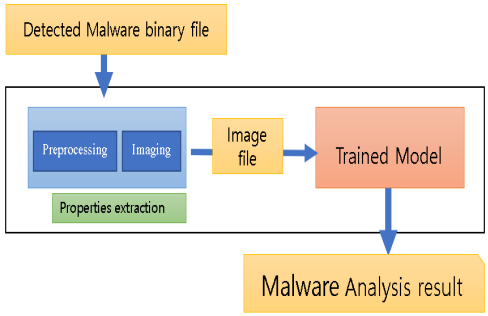

Fig. 4. Detected malware identification process

Fig. 4는 새로운 악성코드를 탐지하는 과정이다. 학습을 통해 생성한 모델에 새로 탐지된 악성코드를 입력하면 입력된 악성코드의 특징을 추출하여 이미지를 만들어 학습된 모델에 입력되면 학습을 통해 생성한 모델에서 가중치 값들과 연산을 수행한 후 악성코드 여부를 판별하여 결과값으로 출력해 준다.

4. 탐지 모델 평가

딥러닝을 이용하여 제안하는 기법들의 성능은 단순히 분류 정확도(accuracy)만 가지고 평가하기는 어렵다. 따라서 모델의 성능을 측정하기 위해서는 오차 행렬(confusion matrix)를 이용하여 모델의 예측값과 실제값 사이의 관계를 세분화하여 성능 지표를 정의한다. 성능 지표는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1 스코어(F1-score) 등이 있다[13].

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

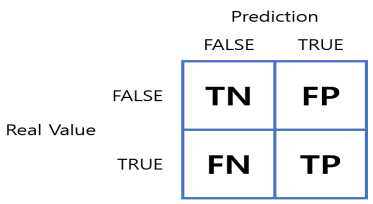

Fig. 5 Components of the confusion matrix

Fig. 5에서 TN는 예측값을 Negative로 예측하고 실제 값 역시 Negative인 것을 표현하고 FP는 예측값을 Positive로 예측하고, 실제 값은 Negative인 것 표현한다. FN은 예측값을 Negative로 예측하고, 실제 값은 Positive인 것을, TP는 예측값을 Positive로 예측하고, 실제 값 역시 Positive인 것을 표현한다.

오차 행렬을 이용하여 정확도는 식(1)과 같이 True Positive 샘플과 True Negative 샘플의 합을 전체 데이터 샘플의 개수로 나누어서 산출한다. 정밀도는 식(2)과 같이 True Positive 샘플을 True Positive 샘플과 False Positiv 샘플을 합한 값으로 나누어서 구한다. 스팸메일 여부 판단할 경우 정밀도 수치가 필요하다. 재현율은 식(3)과 같이 True Positive 샘플을 True Positive 샘플과 False Negative 샘플을 합한 값으로 나누어서 산출한다. 암 판단 모델이나 금융사기 판단 모델과 같이 Positive 데이터를 Negative 로 잘못 판정하면 큰 영향을 미치는 분야에서 재현율이 중요한 지표로 사용한다. 악성코드를 판정하는 모델에서는 재현율이 중요한 지표이다. F1 스코어는 정밀도와 재현율을 결합한 지표로서 정밀도와 재현율의 조화평균으로 구한다.

악성코드 분류에 사용된 데이터는 Microsoft Malware Classification Challenge(Big 2015) 대회에 사용된 데이터를 사용한다. 데이터 수는 총 10,868개이고 9종류의 악성코드로 구성된 데이터 셋이다[14]. Table 1과 같이 각 데이터는 악성코드 실행파일의 16진수 바이트 정보와 IDA(Interactive Disassembler)로 분석한 결과이다.

Table 1. Microsoft Malware Classification Challenge DataSet

sequence	Type	size
1	Ramit	1,541
2	Lollipop	2,478
3	Kelihos_ver3	2,942
4	Vundo	475
5	Simda	42
6	Tracur	751
7	Kelihos_ver1	398
8	Obfuscator.ACY	1,228
9	Gatak	1,013
Total		10,868

제안 모델의 평가는 데이터 챌린지 악성코드 데이터 셋 중 1850개를 사용해 검증하여 Table 2와 같은 결과를 얻었다.

Table 2. Test result of Model

		Prediction Result	
		Normal (564)	Malware (1286)
Real Result	Normal (550)	TN 466	FP 84
	Malware (1300)	FN 98	TP 1202

제안 모델의 평가 결과는 정확도는 90.2%, 정밀도는 93.5%, 악성코드 탐지 모델에서 중요한 지표로 사용하는 재현율은 92.5%로 나타났다.

5. 결론

현대는 기술 변혁시대에 따라 정보통신 기술의 발달과 IoT(Internet of Things)의 활성화로 네트워크를 통한 정보 교환이 무수히 많이 이루어지고 있는 상황이고 이를 위협하는 다양한 종류의 악성코드도 많이 등장하고 있다. 그리고 새롭게 등장하는 대부분의 악성코드는 기존 악성코드를 일부 수정한 변종들이고 기존 악성코드와 많은 유사성을 가지고 있다. 그리고 이런 변종의 악성코드는 기존의 탐지 방법으로 탐지하고 방어하는데 한계점이 많이 존재한다. 따라서 본 논문에서는 악성코드들의 유사성을 기반으로 악성코드를 이미지화하여 분류하고 딥러닝을 이용하여 학습시키는 모델을 제안하였다. 그리고 새로게 탐지된 악성코드를 학습된 탐지 모델에서 분석하면 기존 악성코드를 변형하여 나타난 변종도 탐지할 수 있음을 보여주었다. 최근에는 인공지능을 이용하여 악성코드 탐지 기술을 회피하는 변종 악성코드도 제작하는 단계에 이르렀다[15].

본 연구는 비정형 데이터를 가지고 딥러닝을 이용해 학습할 경우 신뢰도를 높이기 위해서는 충분한 데이터 셋이 필요하다. 하지만 본 연구는 Microsoft에서 제공하는 제한된 데이터 셋으로 학습을 했기 때문에 신뢰도 등 원하는 수치 도출에 미흡함이 있다.

향후 연구로는 많은 데이터 셋을 수집하여 악성코드로부터 더 정확한 특징들을 추출하고 분류하여 다양한 변종 악성코드도 정확하고 신속하게 탐지할 수 있는 딥러닝 기법에 대한 연구가 필요하다.

REFERENCES

[1] McAfee Labs Threats Report. (accessed January 6, 2021), McAfee Labs (Online) https://www.mcafee.com/enterprise/en-us/treat-center/macfee-labs/report.html

[2] D. Ucci, L. Aniello & R. Baldoni.(2019). Survey of machine learning techniques for malware analysis, *Computers & Security, 81,* 123-147. DOI : 10.1016/j.cose.2018.11.001

[3] H. J. Mun, S. H. Choi & Y. C. Hwang. (2016). Effective Countermeasure to APT Attacks using Big Data. *Journal of Convergence for Information Technology, 6(1),* 17-23. DOI : 10.22156/CS4SMB.2016.6.1.017

[4] C. Chen, S. Wang, D. Wen, G. Lai & M. Sun. (2019). Applying Convolutional Neural Network for Malware Detection. *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, Morioka, Japan, pp. 1-5.

[5] L. Nataraj, S. Karthikeyan, G. Jacob, & B. Manjunath. (2011). Malware Images: Visualization and Automatic Classification. *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, pp.1-7.

[6] J. Kim, S. Hong & H. Kim. (2019). A Style GAN Image Detection Model Based on Convolutional Neural Network, *Journal of Korea Multimedia Society, 22(12),* 1447-1456.

[7] A. Hidakay & T. Kurita. (2017). Consecutive Dimensionality Reduction by Canonical Correlation Analysis for Visualization of Convolutional Neural Networks, *In Proceedings of the ISCIE international symposium on stochastic systems theory and its applications* (Vol. 2017, pp. 160-167).

[8] D. Moon, S. B. Pan & I. Kim.(2016). Host-based intrusion detection system for secure human-centric computing, *Journal of Supercomputing.* 72(7), 2520-2536.

[9] W. Huang & J. W. Stokes. (2016). MtNet: A Multi-Task Neural Network for Dynamic Malware Classification, *International Conference on Detection of Intrusions and Malware & Vulnerability Assessment* (pp. 399-418).

[10] W. Xu, Y. Qi & D. Evans. (2016). Automatically Evading Classifiers, *In Proceedings of the 2016 network and distributed systems symposium* (Vol. 10).

[11] D. Gibert. (2016). *Convolutional Neural Networks for Malware Classification,* Master Thesis, University Rovira i Virgili, Tarragona, Spain

[12] G. E. Dahl, J. W. Stokes, L. Deng & D. Yu. (2013). Large-Scale Malware Classification using Random Projections and Neural Networks, In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3422-3426).

[13] S. Visa, B. Ramsay, A. L. Ralescu & E. van der Knaap. (2011). Confusion Matrix-based Feature

Selection, *Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference*, Cincinnati, Ohio, USA, April 16-17.

[14] Microsoft. (2015), Microsoft Malware Classification Challenge(BIG 2015). (accessed July 28, 2021). Kaggle (Online). https://www.kaggle.com/c/malware-classification

[15] Y. M. Cho. & H. Y. Kwon.(2020). Machine Learning Based Malware Detection Using API Call Time Interval, *Korea Institute of Information Security and Cryptology, 30(1),* 51-58. DOI : 10.13089/JKIISC.2020.30.1.51.

황 윤 철(Yooncheol Hwang) [정회원]

· 2008년 2월 : 충북대학교 전자계산학과(이학박사)
· 2019년 3월 ~ 2021년 2월 : 가천대학교 소프트웨어 중심대학 사업단 소프트웨어교육센터 초빙교수
· 2021년 3월 ~ 현재 : 한남대학교 탈메이지 교양교육대학 조교수
· 관심분야 : 네트워크 및 웹 보안, IDS, ITS, Fusion IT Technology(AI)
· E-Mail : dolpin98@nate.com

문형진(Hyung-Jin Mun) [종신회원]

· 2008년 2월 : 충북대학교 전자계산학과(이학박사)
· 2017년 3월 ~ 현재 : 성결대학교 정보통신공학과 조교수
· 관심분야 : 정보보안, 네트워크 보안, 빅데이터분석
· E-Mail : jinmun@gmail.com