GAN 기반의 악성코드 이미지 데이터 증강 분석

이원준[O], 강창훈[*], 강아름[**]
[O]배재대학교 사이버보안학과,
[**]배재대학교 사이버보안학과,
[*]배재대학교 신기술혁신융합대학사업단
e-mail: ayjmlo6040@gmail.com[O], chkang@pcu.ac.kr[*], armk@pcu.ac.kr[**]

Analysis of Malware Image Data Augmentation based on GAN

Won-Jun Lee[O], ChangHoon Kang[*], Ah Reum Kang[**]
[O]Dept. of Cyber Security, Pai Chai University,
[**]Dept. of Cyber Security, Pai Chai University,
[*]New Technology Innovation and Convergence University Project Group, Pai Chai University

● 요 약 ●

다양한 변종들의 존재와 잘 알려지지 않은 취약점을 이용한 공격은 악성코드 수집을 어렵게 하는 요인들이다. 부족한 악성코드 수를 보완하고자 생성 모델을 활용한 이미지 기반의 악성코드 데이터를 증강한 연구들도 존재하였다. 하지만 생성 모델이 실제 악성코드를 생성할 수 있는지에 대한 분석은 진행되지 않았다. 본 연구는 VGG-11 모델을 활용해 실제 악성코드와 생성된 악성코드 이미지의 이진 분류하였다. 실험 결과 VGG-11 모델은 99.9%의 정확도로 두 영상을 다르게 판단한다

키워드: 악성코드, GAN(Generative Adversarial Network), CNN(Convolutional Neural Network)

I. Introduction

SK 쉴더스에서 발표한 2023년 상반기 보안 트렌드 보고서에 따르면 침해 사고의 28%는 악성코드를 이용한 공격이다. 이를 위해 인공지능을 이용한 악성코드 탐지 연구가 많이 진행되었다.

악성코드는 수많은 변종이 존재하고 제로데이 공격의 경우 희소성으로 인해 악성코드 샘플 수집이 어렵다.

이 문제를 해결하고자 악성코드를 시각화하여 생성 모델을 통해 데이터를 증강한 연구가 존재한다[1]. 하지만 해당 연구는 실제 악성코드 이미지를 생성 여부를 분석하지 못하였다.

본 연구에서는 CNN 모델을 활용하여 생성 모델이 실제 악성코드 이미지 생성 여부를 분석하였다. 2장에서는 관련 연구를 소개하고, 3장에서는 실험 구성과 실험 결과를 설명하였다. 4장에서는 결론을 정리하였다.

II. Preliminaries

1. Malware Visualization

기존 연구에서는 악성코드 특징 정보를 활용한 연구가 진행되었다. 하지만 해당 연구들이 특징 정보에 의존적이다. 이 문제를 해결하고자 악성코드를 시각화하여 탐지하는 연구가 진행되었다[2]. Nataraj 외의 연구에서는 악성코드 시각화 방법으로 먼저 악성코드의 바이너리를 8비트 정수 벡터로 값으로 변환한다. 그 후, 파일 크기에 따라 width를 설정하여 행렬로 변환하여 Gray-Image로 시각화하였다.

2. Malware Data Augmentation

Yan과 Jiang의[1] 연구에서는 Convolutional 층을 추가한 DCGAN[3]을 25개 클래스를 갖는 Malimg 악성코드 데이터셋을 이용해 GAN을 학습시켰다. 학습된 GAN을 활용해 악성코드 이미지를 생성하고 훈련 데이터로 사용하였다. 그 후, CNN 모델을 활용하여 25개의 악성코드 패밀리 분류 정확도를 기존 모델 대비 6% 상승시켰다.

III. The Proposed Scheme

본 연구에서는 사용된 GAN은 SAGAN[4]이다. SAGAN은 Self-Attention 메커니즘을 활용하여 Long-term Dependencies 문제를 해결한 모델이다. SAGAN 모델을 Malimg 데이터셋을 이용하여

학습하였다. SAGAN은 총 150 Epoch 학습하였다. SAGAN을 이용해 Malimg 샘플 수와 같은 총 9339개의 이미지를 생성하고 총 18678개의 데이터셋을 구축하였다. Fig. 1은 실제 악성코드 이미지와 생성 이미지의 모습이다.

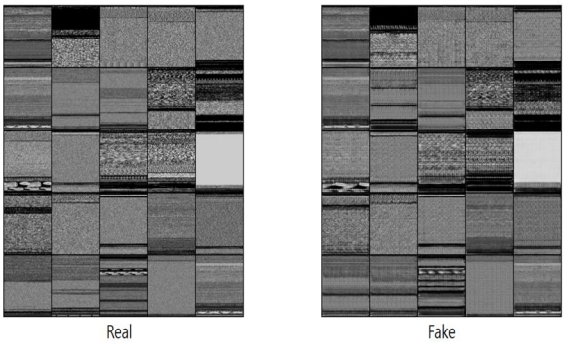

Fig. 1. Malware Image

실제 이미지와 생성 이미지를 분류하기 위해 VGG-11[5] 모델을 활용하였고, 마지막 Fully Connected Layer의 출력을 1로 설정하여 이진 분류를 진행하였다.

Table 1. Evaluation Metrics

	Validation	Test
Accuracy	0.9998	1.0000
Precision	1.0000	1.0000
Recall	0.9996	1.0000
F1 score	0.9998	1.0000

본 연구의 실험 환경은 Ubuntu 20.04 환경에서 Pytorch 프레임워크를 사용하였다. 훈련 데이터의 비율은 5:3:2이다. VGG-11 모델의 학습 횟수는 30 Epoch로 지정하였고, 검증 데이터의 정확도가 5회 동안 개선되지 않는다면 종료하게 하였다. Optimizer는 SGD를 사용하였다, 학습률은 0.001로 지정하였고, 10 Epoch마다 0.1씩 감소시켰다. Table 1에 검증용 데이터와 테스트 데이터의 성능지표를 정리하였다. VGG-11 모델은 실제 이미지와 생성 이미지를 쉽게 구별하는 것을 알 수 있다.

IV. Conclusions

본 연구에서는 CNN 모델을 통해 실제 악성코드 이미지와 GAN을 통한 생성된 악성코드 이미지를 분류하였다. 실험 결과 GAN을 이용해 악성코드 이미지 생성이 가능하지만, 이는 실제 악성코드와 다른 딥페이크 악성코드 이미지라는 것을 확인하였다.

ACKNOWLEDGEMENT

Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea.

REFERENCES

[1] L, Yan and L, Jiang, "Generative Adversarial Network for Improving Deep Learning Based Malware Classification," 2019 Winter Simulation Conference. IEEE, pp. 584-593, 2019.

[2] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunat, "Malware Images: Visualization and Automatic Classification," Proceedings of the 8th International Symposium on Visualization for Cyber Security, pp. 1-7, 2011.

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[4] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena, "Self-Attention Generative Adversarial Networks," International conference on machine learning. PMLR, Vol. 97, pp. 7354-7363, Jun 2019.

[5] K. Simonyan, and A. Zisserman "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.