이미지 전처리와 앙상블 기법을 이용한 이미지 기반 악성코드 분류 시스템

김해수*, 김미희*
*한경대학교 컴퓨터응용수학부
e-mail:{ww232330, mhkim}@hknu.ac.kr

Image-based malware classification system using image preprocessing and ensemble techniques

Hae-Soo Kim*, Mi-hui Kim*
*School of Computer Engineering & Applied Mathematics, Hankyong National University

요 약

정보통신 기술이 발전함에 따라 악의적인 공격을 통해 보안문제를 발생시키고 있다. 또한 새로운 악성코드가 유포되어 기존의 시그니처 비교방식은 새롭게 발생하는 악성코드를 빠르게 분석 할 수 없다. 새로운 악성코드를 빠르게 분석하고 방어기법을 제안하기 위해 악성코드의 패밀리를 분류할 필요가 있다. 본 논문에서는 악성코드의 바이너리 파일을 이용해 시각화하고 CNN모델을 통해 분류한다. 또한 정확도를 높이기 위해 LBP, HOG를 통해 악성코드 이미지에서 중요한 특성을 찾고 데이터 클래스 불균형에서 오는 문제를 앙상블 모델을 통해 해결하는 시스템을 제안한다.

1. 서론

정보통신 기술의 발전이 많은 이점을 주었지만 이것을 이용한 악의적인 공격을 통해 보안 문제를 발생시키고 있다[1]. 그 중에서 악성코드는 다양한 공격 방식으로 동작하여 매번 새로운 방식으로 사람들에게 유포된다. 새로운 악성코드는 계속해서 발견되고 있고 그와 동시에 악성코드의 수가 빠르게 증가하면서 기존의 악성코드의 시그니처와 비교를 통한 분석방식으로는 모든 악성코드를 관리 할 수 없고 새로운 악성코드가 생길 때마다 분석하여 방어기법을 찾는 방식은 시간이 오래 걸리며 분석이 끝나기 전에 새로운 악성코드가 발견될 가능성 생긴다[2] 2021년 6월 WatchGuard technologies의 조사 결과에 따르면 기존 방식으로 탐지에 실패한 악성코드가 74%에 달한다고 한다[3].

새롭게 발생하고 있는 악성코드들을 대응하기 위해 악의적인 공격들에 대한 분석 후 해당하는 방어기법을 제공해야한다.

효과적인 분석을 위해 악성코드의 종류를 분류할 필요가 있으며 동일한 종류의 악성코드들은 동작에 유사성을 갖고 있어 기존에 분석된 악성코드를 이용해 새로운 악성코드의 유형을 알 수 있게 된다. 이를 위해 이미지 기반 분류 기법[4]이 제안되었으나 정확도 측면에서 부족한 점이 있다.

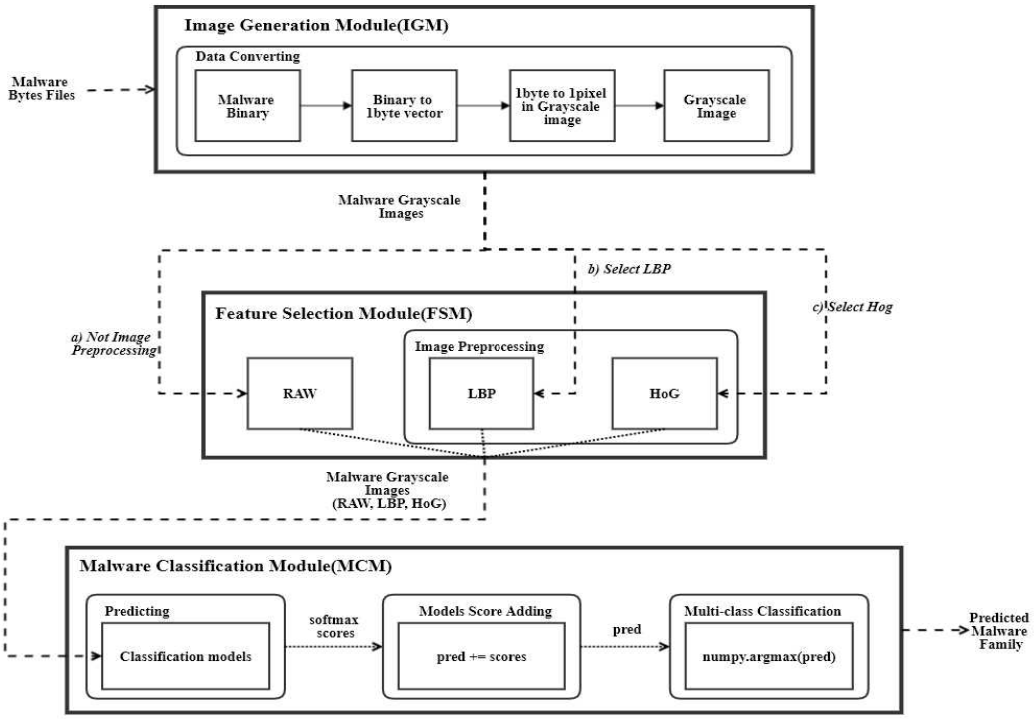
이에, 본 논문에서는 딥러닝을 통해 동작의 유사성에서 오는 이미지 기반 악성코드의 패턴을 이용하여 모델이 악성코드를 자동으로 분류하고 이미지 특징을 추출하는 기법과 앙상블 모델을 통해 정확도를 높이는 방법을 제안한다.

2. 관련연구

2.1 이미지 기반 악성코드 분류

딥러닝을 이용한 악성코드 탐지에 관한 연구는 여러 가지 방법으로 이루어지고 있다.[5] 이미지 기반 악성코드 탐지에서 [6]의 연구에서는 Linux 악성코드들을 64x64크기로 시각화하고 LBP, Median Filter와 같은 영상처리 기법과 Hard voting을 통해 악성코드를 탐지하는 시스템을 구성했으며 각각 영상처리 기법을 적용하지 않은 Original에서 98.77% LBP는 96.47% Median Filter에서는 98.57% 의 정확도를 도출 해냈고 이 세 가지 영상 처리 기법을 적용시킨 모델들로 Hard voting classification을 했을 때의 정확도는 98.87%로 상승한 것을 볼 수 있다.

본 논문에서는 영상처리 기법, 즉 LBP(Local



(그림 1) 제안 시스템

Binary Pattern), HOG(Histogram of Oriented Gredients)과 Soft Voting을 통해 악성코드 분류를 진행한다.

3. 제안 시스템

본 장에서는 본 논문에서 제안하는 이미지 기반 악성코드 분류 시스템을 설명한다. (그림 1)에서 보이는 것처럼 제안 시스템은 이미지 생성 모듈, 특성 선택 모듈, 악성코드 분류 모듈로 구성된다.

3.1 이미지 생성 모듈(Image Generation Module, IGM)

해당 IGM에서는 바이너리 형태로 이루어진 악성코드를 1Byte 단위로 나누어서 2차원 배열의 형태로 구성하고 해당 배열을 정해진 너비를 기준으로 회색조 이미지로 만든다. <표 1>은 이미지의 너비를 정하는 기준이다[7].

<표 1>파일 크기별 이미지 너비[7]

파일 크기 범위	이미지 너비
< 10kB	32
10kB ~ 20kB	64
20kB ~ 60kB	128
60kB ~ 100kB	256
100kB ~ 200kB	384
200kB ~ 500kB	512
500kB ~ 1000kB	768
1000kB <	1024

3.2 이미지 특성 선택 모듈(Feature Selection Module, FSM)

해당 FSM은 3.1장에서 생성된 이미지를 선택된 특성에 따라 전처리를 해주는 모듈이다. RAW는 생성된 이미지를 전처리를 하지 않고 LBP, HOG를 선택하면 이미지를 선택된 방식으로 이미지 특징추출 작업을 한다.

3.2.1 LBP

LBP는 모든 픽셀을 대상으로 주변 3x3 크기의 영역에서 중심 픽셀을 기준으로 상대적인 밝기의 크기를 2진수로 계산하는 알고리즘으로 동작하며[8] 본 논문에서는 동일한 종류의 악성코드 이미지의 유사성이 유사한 텍스쳐를 갖고 있을 것이라는 데에 사용된다.

3.2.2 HOG

HOG는 일정영역의 크기를 셀로 분할한 후 각 셀의 엣지(Edge) 정보를 특징으로 추출하는 알고리즘으로 동작하며[9] 본 논문에서는 동일한 종류의 악성코드 이미지의 유사성이 각 이미지에서 추출되는 엣지 정보가 유사할 것이라는 데에 사용된다.

3.3 악성코드 분류 모듈(Malware Classification Module, MCM)

해당 MCM은 3.2장에서 선택된 특성에 따라 처리된 이미지를 분류 모델들이 해당 악성코드의 패밀리를 예측한다.

3.3.1 CNN[10]

이미지 기반 악성코드 분류를 위해 기본적인 학습 모델로 CNN을 사용한다. CNN을 구성하고 있는 주요 계층(Layer)에는 Convolution Layer, Flatten Layer, Dense Layer가 있다. Convolution Layer는 이미지의 특징을 추출하는 역할을 하고 Flatten Layer는 2차원 이미지 데이터를 1차원 배열의 형태로 바꿔준다. 이후 Dense Layer를 통해 이미지를 분류한다. 본 논문에서는 4개의 Convolution Layer와 1개의 Flatten Layer 3개의 Dense Layer를 사용한다.

3.3.2 LSTM(Long short-term memory)[11]

앙상블 기법을 위해 추가한 모델로 악성코드의 데이터가 순차적으로 이루어져있는 시계열 데이터라는 점을 이용해서 시계열 데이터에 효과적인 LSTM 알고리즘을 사용한다.

3.3.3 앙상블 기법

딥러닝에서 조심해야하는 것 중에 하나가 데이터 클래스 불균형이다[12]. 데이터 클래스 불균형이란 데이터에서 각 클래스 간의 데이터의 개수차이가 큰 경우를 말한다[13]. 데이터 개수가 적은 클래스는 모델의 레이어들이 해당 클래스의 특징을 제대로 학습하지 못해 모델이 분류할 때 제대로 된 결과를 출력해내지 못할 것이다. 이러한 클래스 불균형을 해결하기 위해 두 개 이상의 모델이 도출해낸 결과 값들을 이용하여 최종적인 분류를 해내는 앙상블 기법이 있다.

본 논문에서 다수결 분류 방식의 앙상블 기법을 이용한다. 다수결 분류에는 모델이 가장 많이 선택한 클래스로 분류가 되는 Hard Voting이 있고 모델이 출력한 값들을 모두 더해서 가장 큰 값을 가진 클래스로 분류하는 Soft Voting이 있으며 본 논문에서는 Soft Voting 방법으로 분류한다.

4. 실험 결과

4.1 실험 환경

4.1.1 실험 데이터

본 논문에서 실험에 이용한 데이터는 Kaggle에서 제공되고 있는 Microsoft Malware Classification challenge(BIG 2015)[14] 데이터이다.

<표 2>에 표기된 샘플의 개수를 보면 패밀리 중에 가장 많은 데이터 수는 2942개이고 가장 적은 데이터 수는 42개이다. 실제 데이터는 클래스간의 데이터의 개수가 비슷하지 않을 것이기 때문에 이처럼 클래스 별 데이터 개수의 차이가 큰 데이터를 이용해서 제안 모델의 성능을 측정한다.

<표 2> 데이터 셋에 포함된 악성코드 패밀리[14]

Familly name	#Train Sample	Type
Rammit	1541	Worm
Lollipop	2478	Adware
Kelihos_ver3	2942	Backdoor
Vundo	475	Trojan
Simda	42	Backdoor
Tracur	751	TrojanDownloader
Kelihos_ver1	398	Backdoor
Obfuscator.ACY	1228	Any kind of obfuscated malware
Gatak	1013	Backdoor

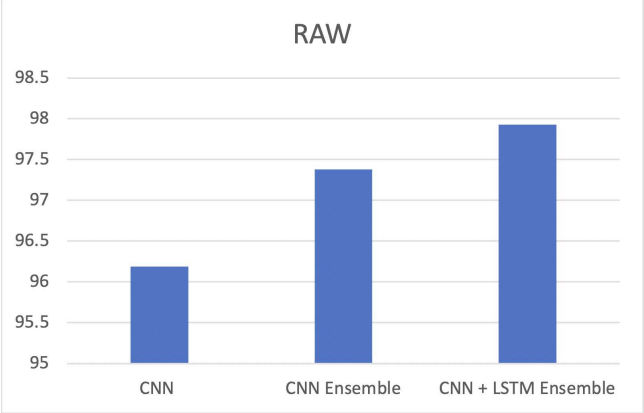
4.1.2 성능 평가 방법

본 논문에서는 기본 이미지(즉, RAW), 두가지 전처리 방법(즉, LBP, HOG)과 CNN모델에 앙상블 기법이 적용된 모델들을 통해 전처리와 앙상블이 정확도 개선에 얼마나 효과적인지 실험하여 평가한다.

평가 방법은 결과로 출력된 softmax값과 onehot encoding 방식으로 인출된 label값을 비교해서 boolean값을 인출하고 True를 1.0, False를 0.0으로 변환 후 모든 결과에 대해 평균을 구하고 해당 값을 정확도로 하여 모델의 성능을 측정한다. 각 전처리 방법 및 모델별 정확도를 비교한다.

4.2 실험 결과 분석

(그림2)에서 보이는 것처럼 데이터 클래스 불균형을 개선하기 위한 방법으로 앙상블기법을 채택하여 실험을 진행한 결과 정확도가 상승했다. LSTM을 앙상블 모델에 추가하였을 때 CNN 앙상블 모델에 비해 정확도가 각각 0.55% 증가 한 것을 볼 수 있다.



Accuracy(%)	CNN	CNN Ensemble	CNN + LSTM Ensemble
RAW	96.19	97.38	97.93

(그림 2)이미지 전처리 및 모델 별 결과

5. 결론

본 논문에서는 새로운 악성코드가 나타남에 따라 악성코드를 대처하는 데에 중요한 딥러닝을 이용해 악성코드를 분류할 때 분류의 정확도를 높이고 데이터를 수집 후 각 클래스간의 데이터 불균형이 생겼을 때 정확도를 높이는 방법을 제안하였다.

실험결과, 이미지의 픽셀에 시계열 정보가 남아있다면 일반적인 CNN 앙상블 모델보다 LSTM이 추가된 앙상블 모델이 정확도가 더 높다는 것을 확인했다.

향후 HOG, LBP에 대해서 악성코드 이미지에서 중요한 특성을 찾고 각 특성에 따른 앙상블 기법의 정확도 상승률을 확인한다. 또한 다른 데이터를 추가해 더 다양한 클래스를 가졌을 때 정확도의 변화에 대해 보이고자 한다.

6. Acknowledgement

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2018R1A2B6009620), 교신저자 김미희.

참고문헌

[1] “2021년 사이버 위협 전망”, KISA 한국 인터넷진흥원, January 26, 2021 https://krcert.or.kr/data/reportView.do?bulletin_writing_sequence=35878

[2] C. Beek, et al, 2021 McAfee Threats report https://www.mcafee.com/enterprise/en-us/lp/threats-reports/jun-2021.html

[3] Press Release, WatchGuard, **June 24, 2021** https://www.watchguard.com/wgrd-news/press-releases/new-watchguard-research-reveals-traditional-anti-malware-solutions-miss

[4] S. Yue, “Imbalanced Malware Images Classification: a CNN based Approach,”, http://arxiv.org/abs/1708.08042 2017

[5] M. Sahin, S. Bahtiyar, “A Survey on Malware Detection with Deep Learning,” 13th International Conference on Security of Information and Networks, 34, 1-6, 2020

[6] S. Kim, D. Kim, H. Lee, T. Lee, “A Study on Classification of CNN-based Linux Malware using Image Processing Techniques,” *Journal of the Korea Academia-Industrial cooperation Society,* 21, 9, 634-642, 2020

[7] L. Nataraj, S. Karthikeyan, G. Jacob, B. S. Manjunath. "Malware Images: Visualization and Automatic Classification," *Proceedings of the 8th International Symposium on Visualization for Cyber Security,* 4, 1-7, 2011.

[8] T. Ojala, M. Pietikäinen, D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition,* 29, 51-59, 1996

[9] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection,” International Conference on Computer Vision & Pattern Recognition (CVPR), San Diego, UnitedStates. 886–893, 2005

[10] K. O'Shea, R. Nash, “An Introduction to Convolutional Neural Networks,”, https://arxiv.org/abs/1511.08458, 2015

[11] S. Hochreiter, J. Schmidhuber, “LONG SHORT-TERM MEMORY,” Neural Computation, 9, 8, 1735-1780, 1997

[12] R. O’Brien, H. Ishwaran, “A random forests quantile classifier for class imbalanced data,” Pattern Recognition, 90, 232-249, 2019

[13] F. Provost, “Machine Learning from Imbalanced Data Sets 101,” AAAI Technical Report WS-00-05. 2000

[14] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, M. Ahmadi, “Microsoft malware classification challenge,” ArXiv e-prints “http://arxiv.org/abs/1802.10135” 2018