J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

윈도우 PE 포맷 바이너리 데이터를 활용한 Bidirectional LSTM 기반 경량 악성코드 탐지모델

Bidirectional LSTM based light-weighted malware detection model using Windows PE format binary data

박 광 연[1] 이 수 진[2*]
Kwang-Yun PARK Soo-Jin LEE

요 약

軍 PC의 99%는 윈도우 운영체제를 사용하고 있어 안전한 국방사이버공간을 유지하기 위해서는 윈도우 기반 악성코드의 탐지 및 대응이 상당히 중요하다. 본 연구에서는 윈도우 PE(Portable Executable) 포맷의 악성코드를 탐지할 수 있는 모델을 제안한다. 탐지모델을 구축함에 있어서는 탐지의 정확도보다는 급증하는 악성코드에 효율적으로 대처하기 위한 탐지모델의 신속한 업데이트에 중점을 두었다. 이에 학습 속도를 향상시키기 위해 복잡한 전처리 과정 없이 최소한의 시퀀스 데이터만으로도 악성코드 탐지가 가능한 Bidirectional LSTM(Long Short Term Memory) 네트워크를 기반으로 탐지모델을 설계하였다. 실험은 EMBER2018 데이터셋을 활용하여 진행하였으며, 3가지의 시퀀스 데이터(Byte-Entropy Histogram, Byte Histogram, String Distribution)로 구성된 특성 집합을 모델에 학습시킨 결과 90.79%의 Accuracy를 달성하였다. 한편, 학습 소요시간은 기존 탐지모델 대비 1/4로 단축되어 급증하는 신종 악성코드에 대응하기 위한 탐지모델의 신속한 업데이트가 가능함을 확인하였다.

☞ 주제어 : Bidirectional LSTM, 윈도우 PE 포맷 악성코드 탐지, EMBER2018

ABSTRACT

Since 99% of PCs operating in the defense domain use the Windows operating system, detection and response of Window-based malware is very important to keep the defense cyberspace safe. This paper proposes a model capable of detecting malware in a Windows PE (Portable Executable) format. The detection model was designed with an emphasis on rapid update of the training model to efficiently cope with rapidly increasing malware rather than the detection accuracy. Therefore, in order to improve the training speed, the detection model was designed based on a Bidirectional LSTM (Long Short Term Memory) network that can detect malware with minimal sequence data without complicated pre-processing. The experiment was conducted using the EMBER2018 dataset, As a result of training the model with feature sets consisting of three type of sequence data(Byte-Entropy Histogram, Byte Histogram, and String Distribution), accuracy of 90.79% was achieved. Meanwhile, it was confirmed that the training time was shortened to 1/4 compared to the existing detection model, enabling rapid update of the detection model to respond to new types of malware on the surge.

☞ keyword : Bidirectional LSTM, Windows PE malware, Detection, EMBER2018

1. 서 론

최근 국내 · 외를 막론하고 표적형 랜섬웨어 및 사회기반시설을 겨냥한 사이버공격이 급증하고 있다. 그리고 코로나로 인해 전환되고 있는 업무 환경에서 필수 도구가 된 원격 접속 VPN 인프라 및 스마트폰을 겨냥한 사이버 위협 역시 증가하고 있다[1].

국방영역에 대한 사이버위협도 해마다 증가하고 있다. 軍을 대상으로 하는 사이버공격 시도는 지난 2016년 3,150건에서 2019년에는 9,121건으로 3배 이상 증가하는 추세를 보이고 있다. 특히 2016년에는 국방망 해킹사건으로 인해 대규모의 군사기밀 유출까지 경험하였다. 최근 들어서는 4차 산업혁명 핵심기술을 기반으로 스마트국방을 구현하기 위해 IoT장비 및 스마트 기기들을 전력체계들에 적용시키는 사례가 늘고 있어 軍에 대한 사이버위

1 Cyber Warfare(Integrated course), Korea National Defense Univ., Chungcheongnam-do, 33021, Korea.
2 Dept. of National Defense Science, Korea National Defnese Univ., Chungcheongnam-do, 33021, Korea.
* Corresponding author (cyberkma@gmail.com)
[Received 8 October 2021, Reviewed 2 November 2021, Accepted 7 December 2021]

협의 영향력이 점차 증가하고 있다.

지난 2017년 한국국방연구원(KIDA)에서 軍에서 사용중인 PC의 MS 윈도우 제품 의존도 실태를 확인한 결과 우리 軍의 PC 99%가 윈도우 운영체제를 설치 및 운용하고 있음을 확인하였다[2]. 이 조사결과는 국방 사이버공간을 안전하게 보호하기 위해서는 윈도우 운영체제에 대한 사이버공격 대비책을 구비하는 것이 무엇보다도 중요함을 의미한다. 이에 본 연구에서는 윈도우 실행 프로그램의 표준인 PE 포맷의 악성코드를 효율적으로 탐지할 수 있는 방안을 제시하고자 한다.

한편 2021년 4월 맥아피(Mcafee)에서 발표한 보고서에 의하면, 2020년에 발견된 윈도우 운영체제 기반의 신종 악성코드가 약 3억 2천만건(일일 평균 87.6만건)에 달한다[3]. 이렇듯 급증하는 신종 악성코드에 능동적으로 대응하기 위해서는 탐지규칙을 신속하게 업데이트 할 수 있어야 한다. 만약 인공지능을 기반으로 악성코드를 탐지하고자 시도한다면 탐지모델이 새롭게 출현하는 악성코드를 적시에 학습해야만 한다. 그러나 인공지능의 특성상 학습하지 않은 데이터는 처리할 수 없으므로 탐지모델이 신종 악성코드를 학습하기 이전에는 이를 탐지해 낼 수 없다는 한계가 존재한다. 따라서 이러한 한계를 극복하기 위해서는 신종 악성코드를 빠르게 학습할 수 있는 탐지모델을 개발해야 한다.

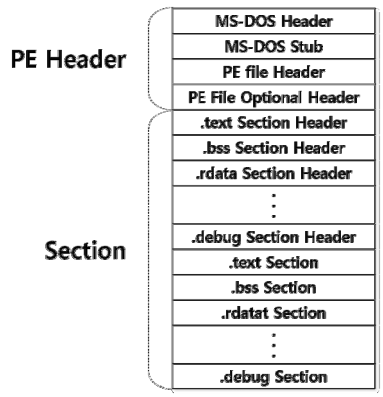
이에 본 연구에서는, 신속한 탐지모델을 개발하기 위해 간단한 전처리와 최소한의 특성을 사용하여 학습하는 방안을 최우선으로 고려하였다. 먼저, 윈도우 PE 포맷의 실행파일로 부터 추가적인 구문 분석 없이 원시 특성을 추출한다. 원시 특성을 추출하기 위해 H. Anderson등이 공개 프로젝트인 "youarespecial"를 통해 제안했던 PEFeatureExtractor를 활용하여 시계열 데이터로 구성된 3가지의 특성 집합(Byte-Entropy Histogram, Byte Histogram, String Distribution)을 추출하였다[4]. 그리고 시계열 데이터 분석을 위해 이전 정보를 현재의 문제해결에 활용하는 RNN 모델을 적용하였다. 그러나 RNN 모델의 경우 시계열 데이터의 시퀀스(sequence) 길이가 길어지면 기울기 소실(gradient vanishing)이 발생하는 한계로 인해 제한적인 학습 데이터만 선택이 가능하다. 따라서 이러한 문제를 보완한 LSTM을 적용하여 시퀀스 길이가 긴 데이터도 활용할 수 있도록 탐지모델을 설계하였다.

본 논문의 구성은 다음과 같다. 2장에서는 윈도우 PE 포맷과 Bidirectional LSTM에 대해 간략하게 살펴본다. 3장에서는 연구에 활용한 EMBER2018 데이터셋에 대해 설명하고, 4장에서 제안하는 탐지모델의 작동개념을 기술한다. 5장에서는 EMBER2018 데이터셋을 대상으로 진행한 실험결과를 분석하고, 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

2.1 윈도우 PE 포맷

윈도우 운영체제에서의 exe, dll 등의 실행파일 또는 이미지 파일들은 PE 포맷을 따르고 있다. PE 포맷은 헤더와 섹션으로 구분되며 구조는 (그림 1)과 같다[5].



(그림 1) 윈도우 PE 포맷 구조
(Figure 1) Structure of Windows PE Format

PE 포맷으로 작성된 실행파일에는 소스코드 외에도 전역/정적변수 정보, Import/Export API 정보, 타임스탬프 값, 메모리 로드 위치 및 크기 등 다양한 정보들이 포함되어 있다. 따라서 윈도우 악성코드를 분석하기 위해서는 PE 포맷에 대한 이해가 필수적이다.

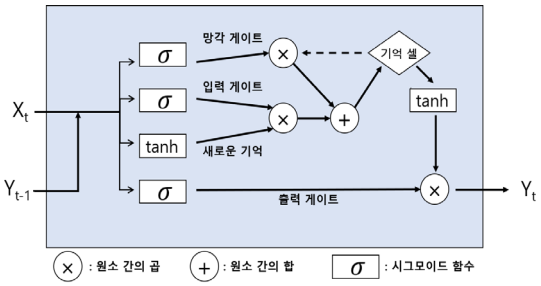
2.2. Bidirectional LSTM 네트워크

LSTM 네트워크는 Simple-RNN 네트워크에서 긴 시퀀스를 가지는 데이터를 학습할 경우 기울기가 발산하거나(blow up) 사라지는(vanish) 문제를 해결하기 위해 고안되었다[6].

LSTM에서 Simple-RNN의 한계를 극복하기 위한 핵심 개념은 (그림 2)에서 보는 바와 같이 게이트(gate)와 셀(cell)의 구조를 도입했다는 점이다. 각 게이트는 기억 셀 주변의 데이터 흐름을 제어하는 역할을 담당한다. 3개의 게이트는 각기 다른 파라미터를 활용하여 입력값을 받으

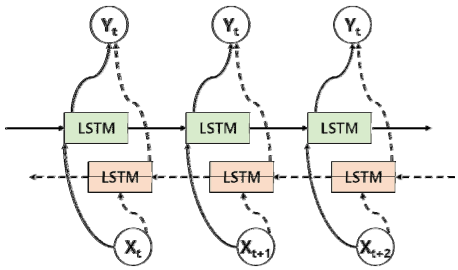
며, 하이퍼볼릭 탄젠트(tanh) 활성화 함수를 거치는 경로까지 포함하여 총 4개의 파라미터셋을 학습한다[7].

이와 같은 LSTM 네트워크는 많은 시퀀스를 지닌 데이터의 패턴을 학습할 수 있어 언어 번역 및 시계열 데이터 분석과 같은 분야에 주로 활용됨은 물론, 네트워크 트래픽 분석, 행동 탐지, 정적 파일 분석과 같은 문제에도 적용할 수 있다. 프로그램 코드 역시 언어와 마찬가지로 순서가 중요하기 때문에 시계열 데이터로 볼 수 있으며 LSTM 네트워크를 통해 학습시킬 수 있다[8].



(그림 2) LSTM Cell 구조[7]
(Figure 2) Structure of LSTM Cell[7]

그러나 LSTM 네트워크는 한 방향으로만 셀 정보를 전달하기 때문에 입력데이터를 입체적으로 해석하는 것이 제한된다. 이러한 문제점을 보완하기 위해 등장한 Bidirectional LSTM 네트워크는 (그림 3)과 같이 정방향과 역방향 모두 2번에 걸쳐 데이터를 학습하여 LSTM 네트워크에 비해 더욱 많은 패턴에 대한 학습이 가능하다[9].



(그림 3) Bidirectional LSTM 네트워크
(Figure 3) Bidirectional LSTM Network

2.3. PE 포맷 바이너리 분석을 통한 특성 추출

악성코드를 탐지하기 위한 시도의 일환으로 파일 시그니처 기반 탐지 방법이 주로 사용되었으나, 바이러스의 샘플 분석에 투입되는 인력과 시간이 과다하여 신속한 대응이 제한되는 한계가 존재하였다.

그러나 기계학습 및 DNN(Deep Neural Network)기법이 사이버 보안 분야에 적용되면서부터는 시그니처를 분석하고 이를 데이터베이스에 등록시키는 학습 과정이 자동화되어 보안인력에 의존했던 기존 방식보다 더욱 효율적인 대처가 가능하게 되었다.

기계학습 및 DNN을 활용하기 위해서는 충분한 양의 악성코드와 양성 프로그램의 샘플들이 필요하며, 악성 프로그램 탐지모델에게 학습을 시키기 위해서는 이러한 샘플들로부터 적절한 특성을 추출하는 것이 무엇보다 중요하다. 특히 LSTM에서 활용하기 위해서는 의미있는 시퀀스를 지닌 특성을 선정해야 한다.

이윤석(2020)은 FUSE 형식 파일의 동적분석을 통해 시스템 호출 정보를 추출하여 LSTM 모델로 악성코드를 분류하는 기법을 제안하였다. 해당 기법은 기계학습 기법(HMM:47.5%, SVM:87.4%)들과 비교했을 때 보다 향상된 92.2%의 Accuracy를 달성하였다[10].

Joshua Saxe 등(2015)은 PE 포맷 파일에서 1,024바이트 단위로 엔트로피(entropy)값과 블록 해쉬값을 추출하여 각각 X, Y축 16개로 매핑(mapping)하는 2차원 평면을 구성하였다. 이를 2-Layer DNN을 활용하여 이진분류를 시도하였으며, 95%의 Accuracy를 달성하였다[11].

M. Ahmadi 등(2016)은 2015 MS Malware Challenge 데이터셋에서 10,000바이트 단위로 엔트로피값을 산출하여 악성코드 패밀리를 분류하였으며, 99.77%의 Accuracy를 달성하였다[12].

3. EMBER2018 Dataset

3.1. 구성

본 연구에서는 윈도우 PE 포맷 악성코드 분석을 위해 EMBER2018 데이터셋을 사용한다. 해당 데이터셋은 100만개의 윈도우 PE 포맷 파일 데이터를 포함하고 있으며, 이는 다시 800,000개의 학습용 데이터셋과 200,000개의 평가용(test) 데이터셋으로 구분된다. 본 연구에서는 정상파일(Benign) 또는 악성코드(Malware)로 구분되지 못한 미분류(Unlabeled) 데이터 200,000개를 제외한 600,000개의 학습용 데이터셋을 이용하였다[13].

3.2. 특성 집합

특성 집합(feature set)은 같은 범주(Domain)의 정보를 갖고 있는 특성의 집합이며 하나의 특성 집합은 여러 개의 특성 집합들로 나누어 질 수 있다. EMBER2018 데이터셋은 8개의 특성 집합을 포함하고 있으며, 각 특성 집합이 의미하는 바는 (표 1)에서 보는 바와 같다[13].

(표 1) EMBER2018 데이터셋의 특성 집합
(Table 1) Feature Sets of EMBER2018 Dataset

특성 집합	포함 정보	특성 수
General	파일크기, Import/Export 함수 개수, 리소스 정보 등	10
Header	타임스탬프, 플랫폼 정보, 이미지 버전, 링커 버전 등	62
Imports	Import 함수 및 소스 라이브러리	1,280
Exports	Export 함수 및 소스 라이브러리	128
Section	섹션 이름, 크기, 섹션 특징 문자열 목록 등	255
Byte Histogram	바이트값의 발생빈도 값	256
Byte-entropy Histogram	바이트 단위 엔트로피 산출 값	256
Strings	문자열 수 및 평균 길이, 출력이 가능한 문자열 히스토그램, 출력 가능한 문자열 엔트로피 값 등	104
총 계		2,351

본 연구에서는 위의 특성 집합들 중 LSTM으로 분석이 용이한 시퀀스 형태를 가지는 특성 집합만을 선택하여 학습에 활용한다. 이에, 학습에 활용한 Byte-Entropy Histogram, Byte Histogram과 Strings 특성 집합에 포함된 String Distribution 특성 집합에 대해 설명한다.

3.2.1. Byte-Entropy Histogram

바이트 값 X와 엔트로피 값 H에 대하여 $p(H, X)$의 결합 분포 값을 산출한 데이터이다. 값을 산출하기 위해서는 고정 길이의 슬라이딩 윈도우(sliding window)내 바이트 값의 발생빈도에 대한 엔트로피 H 값을 계산해야 한다. EMBER2018 데이터셋에서는 2,048바이트 단위로 슬라이딩 윈도우 크기를 지정하였으며, 1,024바이트 단위로 스텝을 이동한다. 이를 통해 한 개의 파일당 16 × 16 크기의 엔트로피 배열이 생성된다[14].

3.2.2. Byte Histogram

바이트 값의 발생 빈도를 나타내기 위해 파일 크기를 토대로 정규화하여 256바이트 단위로 히스토그램에 저장한다. 일반적으로 파일의 바이트 분포는 악성코드 여부를 판단하는 중요한 지표가 된다. 그 이유는 압축 또는 난독화된 프로그램 파일의 경우 파일의 정규분포 값이 높게 관측되기 때문이다[13].

3.2.3. String Distribution

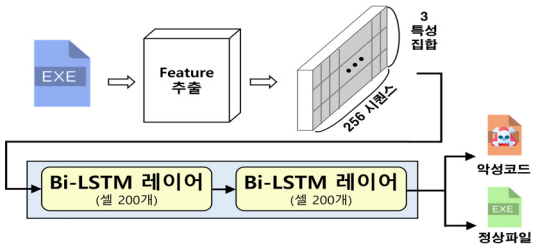
0x20(공백)과 0x7f(del) 사이의 ASCII값 또는 'C:\', 'HKEY_', 'http://'와 같은 특수한 단어를 포함한 문자열 중에서 최소한 5개이상 연속된 것을 토대로 분포값을 산출한다. 이는 앞서 언급한 Byte Histogram값과는 별도의 정보를 제공한다.[14] EMBER2018 데이터셋에서는 Strings 특성 집합 중에서 String Distribution을 별도의 96개의 시퀀스로 구분하고 있다.

4. Bidirectional LSTM 기반 악성코드 탐지모델

4.1. 탐지모델 작동개념

윈도우 PE 포맷 실행파일에서 Byte Histogram, Byte-Entropy Histogram, String Distribution의 3가지 특성 집합을 추출한다. EMBER2018 데이터셋의 경우 Hyrum Anderson이 제안한 PEFeatureExtractor를[4] 이용하여 원시 특성과 PE 포맷 테이블의 각 데이터들을 구문 분석(Parsing)한 특성을 각각 추출하였다.

이후 추출된 특성 집합들을 LSTM 네트워크에 입력할 수 있도록 3차원 배열로 데이터를 가공하고 사전에 학습된 Bidirectional LSTM 모델을 통해 악성코드 여부를 평가한다. 이러한 과정을 순서대로 표현한 모델 작동 개념은 (그림 4)와 같다.

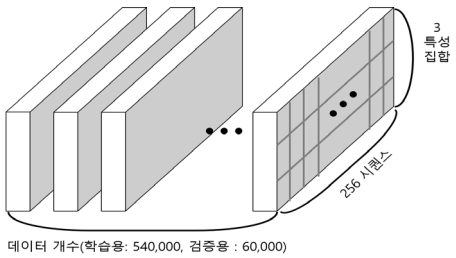


(그림 4) Bidirectional LSTM 기반 악성코드 탐지모델
(Figure 4) Bidirectional LSTM based Malware Detection Model

4.2. 데이터 전처리

LIEF(Library to Instrument Executable Formats) 라이브러리를 사용하여 윈도우 PE 포맷 실행파일의 바이너리 값을 불러온 이후 JSON(JavaScript Object Notation) 형식으로 변환한다. 이후 PEFeatureExtractor는 JSON 형식으로 변환된 실행파일 정보를 전달받아 3가지 원시 특성 집합(Byte-Entropy Histogram, Byte Histogram, Printable String Distribution)의 클래스를 생성하고 시퀀스 데이터를 Numpy 1차원 배열로 구성하여 저장한다.

Bidirectioinal LSTM 네트워크는 Keras로 구현하였다. 이 네트워크에서는 3차원 배열을 입력받기 때문에, 추출한 원시 특성 집합을 (그림 5)와 같은 배열로 변환해야 한다.



(그림 5) LSTM 네트워크 입력 자료 구조
(Figure 5) Structure of Input Data for LSTM Network

이때, Byte Histogram 및 Byte-Entropy Histogram값의 시퀀스 길이(256개)와 String Distribution의 시퀀스 길이(96개)가 다르기 때문에, String Distribution에 해당하는 행의 나머지 160개의 값을 '0'으로 추가하는 제로 패딩(zero padding)을 수행한다.

5. 실험 및 결과

5.1 실험 방법

본 연구에서는 Keras 라이브러리를 활용하여 한 개의 레이어당 200개의 셀을 지닌 Bidirectional LSTM 레이어 2개를 연결하여 탐지모델을 구성하였다. 한 개의 셀은 256개의 시퀀스로 이루어진 특성 집합 3개를 입력받는다. 해당 모델의 출력층은 이진 분류를 위해 시그모이드(sigmoid) 활성 함수를 사용하며, 최적화 함수는 RMSprop, 오차 함수는 Binary crossentropy를 사용하였다.

모델 학습을 위해서는 학습용 데이터셋 800,000개 중 미분류 데이터 200,000개를 제외한 600,000개를 사용하였으며, 검증용(validation) 데이터셋은 학습용 데이터셋의 10%(60,000개)를 추출하여 구성하였다. 학습 시 Batch 크기는 256으로, 학습률은 0.0005로 설정하였다. 실험에 사용된 시스템은 Nvidia GTX 1660(메모리 6GB)가 장착된 환경에서 구성하였고, 학습시간은 5시간 30분이 소요되었다.

5.2 실험 결과

3가지 특성 집합을 달리 사용하며 실험을 수행하였으며, 그 결과는 (표 2)와 같다. 모든 특성 집합들을 사용하였을 때 가장 우수한 90.79%의 Accuracy를 달성하였다.

(표 2) 특성 집합별 탐지모델 성능 비교
(Table 2) Comparison of Detection Model Performance by Feature Set

특성 집합	Accuracy	F1-score
①	85.50%	0.8557
②	88.70%	0.8853
③	88.56%	0.8861
①, ②	88.35%	0.8852
②, ③	89.32%	0.8919
①, ③	89.39%	0.8949
①, ②, ③	90.79%	0.9071

①:Byte-Entropy Histogram, ②:Byte Histogram, ③:String Distribution

EMBER2018 데이터셋을 활용했던 기존 연구들을 살펴보면, 먼저 H. S. Anderson 등[14]이 기계학습 모델인 LightGBM을 이용하였다. 악성코드를 정상파일로 잘못 판별하는 오탐율(False Positive Rate : FPR)이 0.1%미만이 되도록 평가결과에서 악성코드 여부를 결정하는 기준 임계값을 조정한 경우 92.99%의 Accuracy를 달성하였다. 한편, 오탐율이 1%미만이 되도록 임계값을 조정한 경우에는 98.2%의 Accuracy를 달성하였다. S. Parmanik 등[15]은 CNN과 Feed Forward Network을 이용하여 학습시킨 결과 각각 0.95와 0.97의 F1-score 및 Precision을 달성하였다.

위의 실험결과를 기준으로 기존 연구들과 비교했을 경우에는 제안하는 Bidirectional LSTM 기반 모델의 성능이 다소 떨어지는 것으로 보일 수도 있다. 그러나 기존 연구들에서 활용된 모델들은 모든 특성 집합(8개, 7.52GB)을 사용하였으며, 기계학습 기반 탐지모델의 경우에는 GPU (Titan X) 2개를 병렬연결한 시스템을 이용하여 23시간

동안 학습시켰다. 그에 비해 제안모델은 3가지 특성 집합(3.43GB)만 사용하였으며, 상대적으로 적은 자원(GTX 1660)을 가진 시스템상에서도 5시간 30분만에 학습을 완료하였다.

(표 3) 기존 연구와의 성능 비교
(Table 3) Comparison of Performance with Existing Studies

모 델		탐지 성능	학습 시간
LightGBM [14]	0.1% FPR	92.99%(Accuracy)	23시간
	1% FPR	98.2%(Accuracy)	
CNN[15]		0.95 (F1-score, precision)	.
Feed Forward Network[15]		0.97 (F1-score, precision)	
Proposed Model (Bi-LSTM)		90.79%(Accuracy) 0.9071(F1-score)	5시간 30분

6. 결 론

본 연구에서는 윈도우 PE 포맷 형식의 악성코드를 검출하기 위해 시퀀스 데이터를 활용하여 Bidirectional LSTM 기반 탐지모델을 제안하였다. 해당 모델의 학습에는 EMBER2018 데이터셋을 사용하였으며, 기존 연구들은 기계학습, CNN, Feed Forward Network 등을 사용하여 이진분류를 시도했던 것과는 달리 시퀀스 데이터를 Bidirectional LSTM 네트워크에 학습시켜 악성코드와 정상파일을 분류하였다.

(표 3)에서 보는 바와 같이 기존 연구에서 제시된 탐지모델과의 분류 성능을 비교해본 결과 90.79%의 Accuracy를 보여 기존 모델보다 탐지 정확도는 다소 낮게 나타났다. 그러나 기존 모델이 8개의 특성 집합 전체를 사용한 것에 비해 제안모델은 3개의 특성 집합만을 사용하였음에도 비교적 우수한 성능을 달성하였다.

그리고 학습시간은 기존 연구 대비 획기적으로 감소하였다. 서론에서 전술한 바와 같이 2020년 기준 윈도우 운영체제 기반의 신종 악성코드가 일평균 87.6만건이 발생한다는 점을 고려하면[3], 기존 연구에서 제시한 기계학습 기반 탐지모델의 학습시간(23시간)으로는 실시간으로 발생하는 신종 악성코드를 제시간에 학습하는 것이 불가능하다. 반면, 제안모델은 기존 모델에 비해 최소한의 특성 집합을 활용하기 때문에 학습시간이 기존 모델 대비 약 1/4로 단축되었다. 이러한 결과는 신종 악성코드에 대응하기 위해 탐지모델의 업데이트 주기를 단축시킬 수 있음을 시사한다.

향후에는 본 연구에서 활용한 3가지 특성 집합 이외에 추가적인 특성 집합을 활용하여 학습시간은 증가시키지 않으면서 Accuracy를 높일 수 있는 방안에 대해 계속 연구를 진행할 예정이다. 또한, 저전력 소모와 휴대성이 요구되는 무기체계에도 적용이 가능하도록 모델을 보다 경량화시킬 수 있는 방안을 연구하고자 한다.

참고문헌(Reference)

[1] KISA, “2021 First Half KISA Cyber Security Issue Report”, 2021. https://www.krcert.or.kr/filedownload.do?attach_file_seq=3431&attach_file_id=EpF3431.pdf

[2] Shim Seung-bae, “Military introduction direction and tasks of Open O.S”, 33rd SPRi Forum, 2017. https://spri.kr/download/21770

[3] Mcafee, "Mcafee ATR Threats Report 4.21", 2021. https://www.mcafee.com/enterprise/en-us/lp/threats-reports/apr-2021.html

[4] Clarence Chio, David Freeman, "Machine Learning and Security", pp. 175, O’Reilly Media, Inc., 2018.

[5] R. Kath, “The Portable Executable File Format from Top to Bottom”, MSDN Libary, Microsoft Corporation, 1993. http://www.csn.ul.ie/~caolan/pub/winresdump/winresdump/doc/pefile2.html

[6] S. Hocheriter, J. Schmidhuber, “Long short-term memory”, Nneurl computation 9, no. 8, pp. 1735-1780, 1997. http://dx.doi.org/10.1162/neco.1997.9.8.1735

[7] Yukinaga Azuma, “Introduction to core deep learning”, onlybook, 2020.

[8] J. Saxe, H. Sanders, “Malware Data Science”, Youngjin, 2020.

[9] Mike Schuster, Kuldip K. Paliwal, “Bidirectional Recurrent Neural Networks”, IEEE Transactions on signal processing, Vol. 45, No. 11, 1997. http://dx.doi.org/10.1109/78.650093

[9] Yunseok Rhee, “Malicious Code Detection Method Using LSTM Learning on the File Access Behavior”, The Journal of Korean Institute of Information Technology, Vol.18, no. 2, pp.25-32, 2020.

http://dx.doi.org/10.14801/jkiit.2020.18.2.25

[10] J. Saxe, K. Berlin, “Deep neural network based malware detection using two dimensional binary program features”, 2015 10th International Conference on Malicious and Unwanted Software(MALWARE), pp. 11-20, 2015.
http://dx.doi.org/10.1109/MALWARE.2015.7413680

[11] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, G. Giacinto, “Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification”, In Proceedings of the 6th ACM CODASPY ‘18, pp. 183-194, 2016.
http://dx.doi.org/10.1145/2857705.2857713

[12] Y. Oyama, T. Miyashita, H. Kokubo, “Identifying Useful Features for Malware Detection in the Ember Dataset”, 2019 7th International Symposium on Computing and Networking Workshops(CANDARW). IEEE, pp. 360-366, 2019.
http://dx.doi.org/10.1109/CANDARW.2019.00069

[13] H. S. Anderson, P. Roth, “EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models”, arXiv preprint arXiv:1804.04637, 2018.
https://arxiv.org/abs/1804.04637v2

[14] S. Parmanik, H. Teja, “EMBER – Analysis of Malware Dataset Using Convolutional Neural Networks”, 2019 3rd International Conference on Inventive Systems and Control(ICISC), pp. 286-291, 2019.
http://dx.doi.org/9/ICISC44355.2019.9036424

◐ 저 자 소 개 ◑

박 광 연(Kwang-yun Park)
2010년 육군사관학교 컴퓨터공학과(공학사)
2020년~현재 국방대학교 사이버전 협동과정(공학석사)
관심분야 : 국방보안정책, 사이버보안, 딥러닝
E-mail : edig1097@gmail.com

이 수 진(Soo-jin Lee)
1992년 육군사관학교 전산학과(공학사)
1996년 연세대학교 대학원 컴퓨터과학과(공학석사)
2006년 한국과학기술원 전산학과(공학박사)
2006년~현재 국방대학교 국방과학학과 교수
관심분야 : 국방보안정책, 사이버안보, 인공지능
E-mail : cyberkma@gmail.com