DGA 도메인 탐지를 위한 효과적인 방법 연구

강태우[1], 박순태[2], 엄익채[3]
[1]전남대학교 정보보안융합학과 석사과정
[2]한국인터넷진흥원
[3]전남대학교 정보보안융합학과 교수
taewoo.kang@kisa.or.kr, stpark12@kisa.or.kr, iceuom@chonnam.ac.kr

A Study on Effective Methods for DGA Domain Detection

Tae-Woo Kang[1], Soon-Tai Park[2], Ieck-chae Euom[3]
[1]Dept. of Information Security Convergence, Chonnam National University
[2]Korea Internet & Security Agency
[3]Dept. of Information Security Convergence, Chonnam National University

요 약

DGA(Domain Generation Algorithms)로 생성된 도메인을 탐지하기 위한 다양한 연구 결과가 선행되었다. 기존 연구 상에서는 딥러닝 모델인 LSTM을 이용한 DGA 도메인 탐지가 가장 효과적인 방법으로 대두되었다. 하지만 본 논문 실험 결과, TCN 모델을 이용한 탐지 결과가 LSTM 모델보다 우수한 탐지 정확도를 나타내는 것을 확인하였다. 또한, 탐지 모델을 대규모 도메인 처리가 필요한 현업에서 사용될 것을 고려하여, LSTM과 TCN 모델보다 빠른 결과를 도출할 수 있는 XGBoost 모델을 확인하였다. TCN과 XGBoost 모델을 활용하여 현업에서 DGA 도메인을 탐지하는데 효과적으로 사용될 수 있을 것이다.

1. 서론

DGA는 다량의 도메인을 주기적으로 생성하는 알고리즘으로 하루 수만 개의 도메인을 생성할 수 있으며, 도메인 주소를 시드(Seed)값과 다양한 정보를 이용해 무작위로 생성한다.[1] 이 때문에 기존 악성코드 분석을 통해 일일이 C&C 주소를 차단하는 것은 불가능에 가까워졌다. 또한 DGA는 Seed를 이용하여 특정 암호 알고리즘을 이용, 랜덤하게 도메인을 생성하는 특성으로 인해 공격자가 사용하는 시드 값을 알 수 없다면 무작위로 생성되는 도메인을 예측하는 것 또한 불가능하다.

공격자들은 위와 같은 DGA의 특성을 이용하여 다양한 방면으로 공격에 활용하고 있다. 2016년에 등장한 Locky 랜섬웨어는 공격대상의 PC나 서버에 침투한 이후 파일 암호화에 사용할 키(Key)값을 적용하기 위해 악성코드 내부에 저장된 IP 주소로 1차 접속을 시도한다[2]. 하지만 차단 등의 이유로 IP 접속에 실패할 경우, 2차로 DGA를 통해 생성한 도메인으로 접속을 시도한다. DGA 함수는 날짜 정보를 통해 생성하며 여기에 4Byte의 시드를 적용하여 가변적인 형태로 생성하기에 도메인 값을 지속적으로 변경하는 특징을 보이고 있다.

Locky 랜섬웨어와 같이 DGA를 이용할 경우 C&C 도메인이나 주소를 알기 어려운 문제가 있으며 이를 해결하기 위해 최근 딥러닝을 이용한 DGA 도메인 탐지 방법이 다양하게 연구되고 있다. 다양한 딥러닝 모델을 활용한 연구를 통해 DGA 탐지율을 높이고, 이를 악성코드 분석 및 C&C 차단, 피해방지 등 침해사고 대응 업무에 적용하여 악의적인 C&C 통신으로부터 내부 시스템을 보호할 필요가 있을 것으로 보인다.

2. DGA 종류

DGA의 종류는 크게 3가지로 구분이 가능하다.[3]

첫째는 PRNG(Pseudorandom number generator) DGA로 가장 일반적인 DGA 생성 방법이다. 이는 주로 시스템 날짜나 시간과 같은 무작위 Seed를 사용하여 도메인 시퀀스를 생성하며, 공격자와 악성코드가 도메인 시퀀스를 예측할 수 있다는 특징이 있다. 둘째는 문자 기반 DGA로 도메인을 생성할 때 특정 문자 패턴이나 규칙을 기반으로 생성한다. 예로, 특정 문자열이나 패턴이 XOR 연산과 같은 일정 규칙에 따라 변형되거나 조합되어 도메인이 생성된

다. 해당 DGA는 주로 검출하기가 쉬운 무작위 문자열로 구성되어 있다는 특징이 있다. 세번째는 사전 기반 DGA는 미리 정의된 단어 목록 또는 사전을 사용하여 도메인을 생성한다. 알려진 단어를 무작위로 결합하여 읽을 수 있는 도메인을 생성하는데, 합법적인 도메인으로 보이기 때문에 보안시스템에서 탐지가 어렵다. 마지막으로 고충돌 DGA는 정상 도메인으로 보이도록 설계되었으며 .com, .net, .org와 같은 최상위 도메인(TLD)과 결합되어 이용된다. 이 때문에 해당 DGA로 생성된 도메인은 이미 등록된 도메인과 충돌이 발생할 수 있다는 특징이 있다.

3. 기존 연구

3.1. CNN 기반 DGA 탐지

Bin Yu CNN과 RNN 모델을 이용하여 DGA를 탐지하는 방법에 대해 실험하였다.[4] 실험 결과, 해당 실험 환경에서 CNN/ RNN 모델은 97-98% 가량의 DGA 도메인을 탐지하는 비슷한 수준의 결과를 보여주었으며, Random Forest를 기반으로 한 탐지 결과보다 효율적인 결과를 나타내었다.

3.2. LSTM 모델 기반 DGA 탐지

*J. Woodbridge*는 DGA 도메인의 실시간 탐지를 위해 수동으로 추출한 특성을 사용하는 Random Forest 모델과 특성 추출이 필요없는 LSTM 모델의 성능을 비교하였다.[5] 실험 결과, LSTM 모델이 Random Forest 모델에 비해 더 높은 DGA 도메인 예측 결과를 나타내었으며, 특히 학습 데이터의 양이 적을 때에도 나은 성능을 보여주는 특성이 나타났다.

3.3 LSTM과 GRU 모델 비교를 통한 DGA 탐지

김현종 등은 DGA 도메인 탐지를 위해 LSTM 모델과 GRU 모델의 탐지 결과를 비교하였다.[6] 실험 결과, LSTM과 GRU 모델은 유사한 성능을 보였으나. LSTM에 비해 GRU 모델이 더 빠른 속도로 학습 및 실행될 수 있는 결과가 나타났다.

4. 제안하는 DGA 탐지방법

본 장에서는 TCN과 XGBoost 모델을 이용하여 기존 관련연구에서 사용되었던 LSTM, CNN 등의 딥러닝 모델보다 높은 탐지율, 빠른 속도로 탐지하는 방법을 소개한다. 먼저, DGA로 생성된 도메인과 정상 도메인을 <표 1>과 같이 데이터셋으로 구성한다. DGA 도메인은 Bambenek Consulting OSINT[7]에서 제공하는 DGA Domain Feed 데이터 582,045건으로 구성하며, 정상 도메인은 Alexa에서 수집한 정상 도메인 658,287건으로 구성하였다. 학습 세트와 테스트 세트는 8:2로 구성하였다.

<표 1> 데이터셋 구성

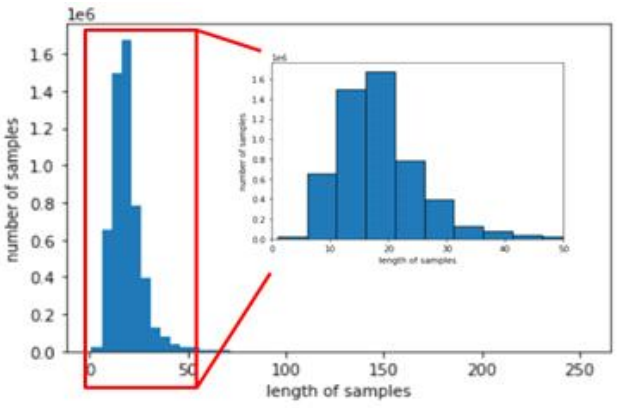
구분	학습 데이터	테스트 데이터	합계
DGA	465,429건	116,616건	582,045
정상	526,837건	131,450건	658,287
총계	992,266건	248,066건	1,240,332

4.1 실험환경 구성

실험을 위한 알고리즘은 TCN, XGBoost, DNN, LSTM, CNN으로 선정하였다. 실험 환경은 CPU intel Xeon Silver 4110 CPU @ 2.10GHz 16 Core-32 Thread * 2, 메모리 256BGB, GPU Nvidia Tesla T5 16GB로 구성하였다.

4.2 전처리

도메인에 대한 데이터 프로세싱 진행을 위해 Char2Idx dictionary를 생성한다. 이후 Char2Idx를 사용하여 <표 2>와 같이 char에 도메인을 매핑하여 정수로 변환한다. 도메인의 Max length는 253이며, 평균 length는 28로 지정하였다. 평균 length의 비율은 89%로, max length 28 이후 남는 길이는 0으로 padding하였다.



(그림 1) 수집(정상, DGA) 도메인의 평균 길이

<표 2> 전처리를 위한 정수 매핑 테이블

특수문자	value
-, ., _	0, 1, 12
숫자	Value
0 - 9	2 - 11
영어	value
a - z	13 - 38

내는 것을 확인하였다. 또한, XGBoost 모델 실험을 통해 해당 모델이 타 모델들보다 초당 처리 속도가 매우 우수한 것 또한 확인하였다. 따라서, 실제 현업에서 DGA 도메인을 탐지할 경우 정확한 탐지율 면에서는 TCN 모델을, 속도 면에서는 XGboost 모델을 적용하는 것이 효과적이다.

다만, XGBoost 모델의 탐지 성능이 타 모델에 비해 다소 떨어지는 것은 개선할 필요성이 있다. 따라서,

<표 3> 실험 결과

Algorithm	Label	Precision	Recall	F1-score	Accuracy	sample/sec
TCN	0	0.95	0.96	0.96	0.96	11,039
	1	0.96	0.95	0.95		
XGBoost	0	0.89	0.94	0.91	0.90	3,268,960
	1	0.93	0.86	0.89		
DNN	0	0.86	0.92	0.89	0.88	33,146
	1	0.90	0.83	0.86		
LSTM	0	0.91	0.86	0.88	0.88	5,291
	1	0.85	0.90	0.88		
CNN	0	0.91	0.95	0.93	0.92	19,295
	1	0.94	0.89	0.92		

4.3 실험결과

본 절에서는 DGA로 생성된 도메인 분류를 위한 모델의 실험 결과를 설명한다. 먼저, 모델별 DGA 실험 결과는 <표 3>과 같다. 여기서 Label 0은 정상 도메인에 대한 실험 결과이고, Label 1은 DGA 도메인에 대한 실험 결과를 뜻한다. 실험 결과, TCN 모델의 F1-Score는 0.96, Accuracy는 0.96으로 LSTM, XGBoost를 포함한 타 모델들보다 높은 성능을 나타내는 것을 확인하였다. 이는 TCN 모델들이 시퀀스 데이터에 대해 타 모델들보다 좋은 성능을 보이는 것으로 해석할 수 있다. 또한, XGBoost 모델의 경우, 다른 모델들과 비교하여 F1-Score, Accuracy는 다소 떨어지지만, 초당 DGA 도메인 처리량이 타 모델들과 비교하여 매우 우수한 것을 확인할 수 있다. 이는 처리속도가 중요한 통신사, 국가 DNS 등 대규모 네트워크 환경, 대규모 DNS를 운용하는 환경에 적용이 가능한 모델이 될 수 있을 것으로 생각된다.

5. 결론 및 향후연구

본 논문에서는 기존 연구[4],[5],[6]에서 사용한 CNN, LSTM 등 다양한 딥러닝 모델들 이외에 TCN 모델에 대한 탐지 성능 실험을 진행하였으며, 구축 환경에서 기존 모델들보다 높은 탐지율을 나타

향후 TCN과 XGBoost 모델 앙상블을 통해 본 논문에서 실험했던 TCN 모델보다 DGA 탐지율이 향상되는지 연구 예정이다.

참고문헌

[1] Stellarcyber, https://stellarcyber.ai/what-are-dgas/2024년 3월 16일 접속

[2] PIOLINK, https://www.piolink.com/kr/service/Security-Analysis.php?bbsCode=security&vType=view&idx=82/2024년 3월 16일 접속

[3] Akamai, https://www.akamai.com/glossary/what-are-dgas/2024년 3월 16일 접속

[4] Bin Yu, Jie Pan and Jiaming Hu, “Character Level based Detection of DGA Domain Names” In Int’l Joint Conf. on Neural Networks, pp.1-8, 2016

[5] J. Woodbridge, H.S. Anderson and A. Ahuja, D.Grand, “Predicting Domain Generation Algorithms with Long Short-Tern Momory Networks” arXiv:1611.00791, Nov. 2016

[6] 김현종, 길명선, 문양세, “DGA 도메인 탐지를 위한 LSTM과 GRU 모델 비교” 한국정보과학회, pp. 122-124, Jun. 2021

[7] Bambenek Consulting OSINT, https://osint.bambenekconsulting.com/feeds/ 2024년 3월 16일 접속