딥러닝 기반 네트워크 침입탐지를 위한 데이터 전처리 방안 연구

정기문O
O한국과학기술정보연구원 슈퍼컴퓨팅클라우드센터
e-mail: kmjeong@kisti.re.krO

A Study of Data Preprocessing for Network Intrusion Detection based on Deep Learning

Kimoon JeongO
ODept. of HPC Cloud Center, Korean Institute of Science and Technology Information

● 요 약 ●

최근 딥러닝 기술이 발전함에 따라 이를 네트워크 침입탐지 분야에 적용하려는 연구가 활발히 이루어지고 있으며 이에 따라 대용량 네트워크 데이터에 대한 처리 방법이 주목받고 있다. 본 논문에서는 네트워크 데이터를 이미지화하는 전처리 방법을 제안한다. 네트워크 데이터를 세션단위로 처리하여 손실율을 줄이면서 딥러닝 알고리즘에 바로 적용할 수 있도록 정규화된 이미지로 변환하는 방법이다. 이를 통해 딥러닝 기술을 적용한 네트워크 정보보안 분야의 연구 활성화를 기대할 수 있다.

키워드: 네트워크 침입탐지(network intrusion detection), 딥러닝(deep learning), 전처리(preprocessing)

I. Introduction

해킹, 개인정보침해 등으로부터 정보통신 자산 및 개인정보를 보호하는 일은 필수적인 일이 되었다. 특히 정보보호를 위한 네트워크 침입탐지는 가장 기본적이면서 아직까지 완벽하게 해결하지 못하고 있는 분야이다. 현재까지 네트워크 침입탐지를 위해 가장 많이 사용되고 있는 기술은 시그니처 기반 탐지이지만 신규 침입을 탐지하지 못하는 한계가 있다. 이에 따라 통계적 기술, 기계학습 기술 등이 적용되어 왔고 최근에는 딥러닝 기술을 적용하는 침입탐지기술이 활발히 연구되고 있다[1].

침입탐지에 딥러닝 기술을 적용하기 위해서는 대용량의 네트워크 트래픽 데이터에 대한 처리가 필요하다. 이에 본 논문에서는 Convolutional Neural Network (CNN), Recurrent Neural Network(RNN), Long Short Term Memory Network(LSTM) 등 딥러닝 알고리즘에 손쉽게 적용할 수 있는 네트워크 트래픽에 대한 전처리 방법을 제안한다. 제안하는 방식은 손실없는 네트워크 데이터 분석을 지원하기 위하여 세션 기반으로 트래픽을 수집하고 이를 이미지화하는 전처리방식이다. 이를 통해 딥러닝을 이용한 네트워크 보안 연구에 기여할 것으로 기대된다.

II. Related works

정보보안 분야에서 딥러닝을 기반으로 제안된 연구는 악성코드 탐지 분야가 활발하다. 악성코드 분류를 위해 샘플 코드에 대한 이미지화 전처리 방식이 제안되고 있다. [2], [3]에서는 CNN기반의 악성코드 패밀리 분류기를 제안하고 있다. 네트워크 분야에서 딥러닝을 이용하는 기술은 [1]에서 찾아볼 수 있으나 데이터 전처리부분은 자세히 기술되지 않고 있다.

III. The Proposed Method

본 논문에서 제안하고 있는 네트워크 데이터 전처리방법은 Fig. 1.과 같다.

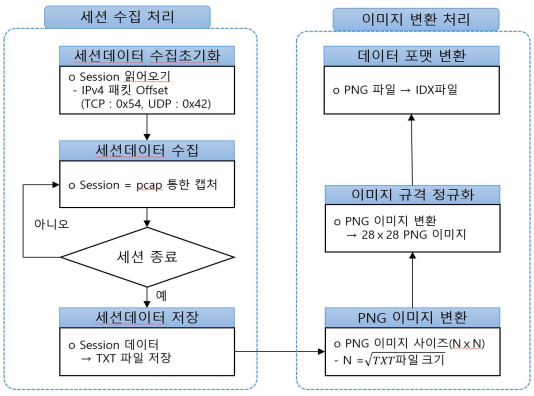

Fig. 1. Proposed Session to Image Process

전처리 단계는 크게 “세션수집처리” 부분과 “이미지 변환처리” 부분으로 구성된다. “세션수집처리”는 네트워크 트래픽을 pcap을 이용하여 세션 단위로 읽고 txt파일로 저장하는 단계이다. 이때 세션 종료 조건은 FIN flag나 RST flag를 수신하거나 마지막 패킷 수신이후 30초 이내에 추가적인 패킷이 없을 때까지 이다. 수집된 세션 데이터의 예는 Fig. 2.에서 볼 수 있다

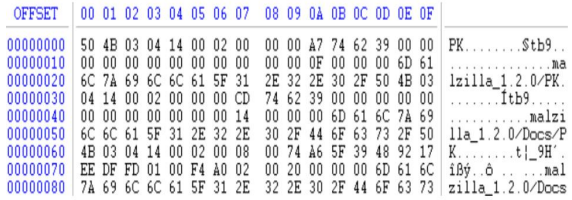

Fig. 2. Sample of Session data

“이미지변환처리”는 딥러닝에 적용하기 위한 데이터 전처리의 핵심 부분으로서 저장된 데이터를 이미지화한다. 먼저 수집된 세션별 txt파일을 이미지 파일인 PNG 형태로 변환한다. 이미지 사이즈를 N×N으로 맞추기 위하여 N값을 파일사이즈의 제곱근으로 하였다. 이에 따라 다양한 크기의 이미지가 생성되는데 이는 Fig. 3.에서 확인할 수 있다. 다양한 크기의 이미지는 딥러닝을 위한 계산 과정에서 사용할 수 없기에 정규화가 필요하다. 본 논문에서는 [4]를 참고하여 28×28 크기로 이미지를 정규화하였고 이는 Fig. 4.에서 확인할 수 있다. 이후 계산과정에서 바로 사용할 수 있도록 PNG 파일은 IDX 파일로 변경한다.

이와 같은 과정은 모든 수집된 세션 데이터에서 반복되고 이에 따라 수집하는 네트워크세션 데이터는 CNN, RNN, LSTM 등 다양한 딥러닝 알고리즘 등에 바로 적용해서 사용할 수 있다.

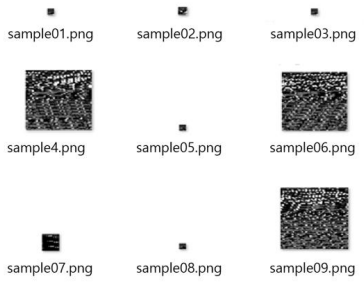

Fig. 3. Sample of initial PNG images

Fig. 4. Sample of nomaolized PNG file

IV. Conclusions

최근 정보보안분야에 인공지능을 적용하려는 연구가 증가함에 따라 보안 데이터 처리에 대한 관심이 증가하고 있다. 본 논문에서는 딥러닝에서 활용할 수 있도록 네트워크 세션 데이터를 전처리하는 방식을 제안하였다. 이를 이용해 네트워크 침입탐지 연구분야 활성화에 기여할 것으로 보인다.

REFERENCES

[1] A. E. Aminanto, and K. J. Kim, "Deep learning in intrusion detection system: An overview." International Research Conference on Engineering and Technology 2016. 2016.

[2] S. H. Seok, and H. W. Kim, “Visualized Malware Classification Based-on Convolutional Neural Network,” Journal of The Korea Institute of Information Security & Cryptology, Vol. 26, No. 1, pp.197-208, Feb. 2016.

[3] Daniel Gibert, “Convolutional Neural Networks for Malware Classification,” Master Thesis, Universitat de Barcelona, 2016

[4] THE MNIST DATABASE of handwritten digits, http://yann.lecun.com/exdb/mnist/