어텐션 시각화를 통한 암호화 트래픽 분석

김주성, 장윤성, 박지태, 최유진, 백의준, 김명섭
고려대학교
{jsung0514, brave1094, pjj5846, yujin0706, pb1069, tsmkim}@korea.ac.kr

Encrypted traffic analysis using attention visualization

Ju-Sung Kim, Yoon-Seong Jang, Jee-Tae Park, Yu-Jin Choi, Ui-Jun Baek, Myung-Sup Kim
Korea Univ.

요 약

현대 네트워크에서의 데이터 보호를 위해 대부분의 통신이 암호화되지만, 이는 악성 행위 탐지를 어렵게 만들고, 응용 프로그램의 다양화는 트래픽 분석을 복잡하게 한다. 딥러닝, 특히 트랜스포머 모델을 활용한 응용 트래픽 분석은 높은 정확도로 암호화된 트래픽을 분류하지만, 모델의 복잡성으로 인한 해석 어려움이 있다. 이를 극복하기 위해 어텐션 메커니즘의 시각화를 제안하여 딥러닝 기반 트래픽 분류 모델의 해석성을 향상시키는 방법을 제안하였다.

Ⅰ. 서 론

현대의 네트워크 환경에서 데이터 보호는 최우선 과제 중 하나이고, 특히 금융 거래 및 개인 데이터 보호를 위해 대부분의 데이터 통신이 암호화되고 있다. 이는 데이터의 기밀성과 무결성을 보호하는 중요한 조치이지만, 동시에 악성 행위자들이 암호화된 트래픽을 이용하여 탐지를 회피하려는 시도도 증가하고 있다. 더불어, 인터넷 기술의 발전과 스마트 디바이스의 보급은 다양한 응용 프로그램의 등장을 촉진시켜 응용 트래픽의 양과 복잡성이 계속해서 증가하고 있다. 이러한 응용 프로그램들은 각기 다른 데이터 전송 패턴과 네트워크 행위를 보여주며, 사용자의 디지털 경험을 풍부하게 만든다. 하지만, 이는 네트워크 관리자와 보안 분석가에게는 트래픽을 정확하게 분석하고 이해해야 하는 도전과제를 제시한다.

페이로드나 통계 특징 기반의 전통적인 응용 트래픽 분석 방법은 암호화 데이터 확인 불가, 스케일링 문제, 정밀도의 한계 등의 이유로 위와 같은 요구사항에 대해 만족하지 못하고 있다. 이러한 한계를 극복하기 위해 딥러닝과 같은 고급 기계학습 기법이 도입되고 있다. 딥러닝 기반 분석 방법은 암호화 응용 트래픽의 복잡한 시퀀스 데이터의 패턴을 학습하여 분류하는 방식이다. 특히, 트랜스포머 모델을 적용하여 어텐션 매커니즘과 딥러닝을 통한 학습으로 암호화 트래픽의 복잡한 패턴을 인식하여 높은 정확도로 분류하였다[1]. 그러나 딥러닝 모델은 내부 매커니즘이 복잡하여 모델의 예측이나 분류 기준을 사람이 이해하기 어렵다는 한계가 있어 실시간 네트워크 환경에서의 적용에 제한적이다. 이에 본 논문은 어텐션 메커니즘을 시각화하여 딥러닝 기반 트래픽 분류 모델의 단점 중 하나인 해석의 어려움을 극복할 수 있는 방법을 제안한다.

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(No.RS-2023-00230661, 하이브리드 양자키분배 방법 및 망 관리 기술 표준개발), 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(00235509, ICT융합 공공 서비스●인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)

Ⅱ. 본론

본 장에서는 어텐션 메커니즘을 시각화하여 암호화 트래픽을 분석하는 방법을 제안하고, 본 논문에서 제안하는 방법론의 한계점을 언급한 후 향후 연구 방향을 제시한다.

어텐션 메커니즘은 딥러닝, 특히 자연어 처리에서 중요한 입력 시퀀스의 부분에 집중하도록 모델을 돕는 기술이다. 이 메커니즘은 특정 태스크에 있어 중요한 정보에 주의를 더 많이 기울이고, 덜 중요한 정보는 상대적으로 덜 고려함으로써 모델의 성능을 향상시킨다. 어텐션 스코어는 입력 시퀀스의 각 요소가 현재 처리 중인 요소와 얼마나 관련이 있는지를 수치적으로 나타낸다. 예를 들어, 트랜스포머 모델에서는 쿼리(query), 키(key), 값(value)라는 개념을 사용하여 어텐션 스코어를 계산한다[2]. 여기서 쿼리와 각 키 사이의 적합도를 점수로 평가하고, 이를 통해 값에 가중평균을 적용한다. 하지만 트랜스포머 레이어의 특성상, 정보가 레이어에 따라 계속 혼합되는 것을 감안하여 어텐션 롤아웃 방법을 통해 신뢰성을 확보하였다[3]. 해당 방법에서 사용하는 용어와 공식은 다음과 같다.

- Q, K, V: 쿼리, 키, 값 행렬
- d_k : K의 차원, l_i : i번째 레이어
- $A: raw-attention$, $\tilde{A}$: $attention\ rollout$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}) & \text{if } i > j \\ A(l_i) & \text{if } i = j \end{cases} \quad (2)$$

이렇게 계산된 어텐션 스코어는 모델이 입력의 어떤 부분을 주목하고 있는지를 보여주며, 이 과정을 통해 모델은 주어진 문제에 대해 더 효과적인 정보 처리가 가능해진다. 그러나, 어텐션 스코어가 제공하는 정보는 종종 많은 계산과 맥락 해석을 필요로 하며, 이는 특히 다층 트랜스포머 아키텍처에서 복잡성을 증가시킨다. 본 논문에서는 이 부분을 CLS 토큰이 가지고 있는 어텐션 관련 정보를 시각화하여 해석하는 방법을 제안한다.

CLS(Classification) 토큰은 트랜스포머 기반의 언어 모델인 BERT에서 사용되는 특수한 토큰이다. 주로 입력 시퀀스의 시작 부분에 배치되며, 모델이 전체 입력에 대해 어떤 결론을 내려야할 때 중요한 역할을 한다. 특히, CLS 토큰은 Self-Attention 메커니즘을 통해 다른 토큰들과의 어텐션 스코어를 계산하고 이를 통해 전체 시퀀스에서 중요한 부분에 더 많은 더 많은 어텐션을 할당한다. 이러한 특성을 활용하여 CLS 토큰이 갖는 어텐션 스코어를 시각화한다.

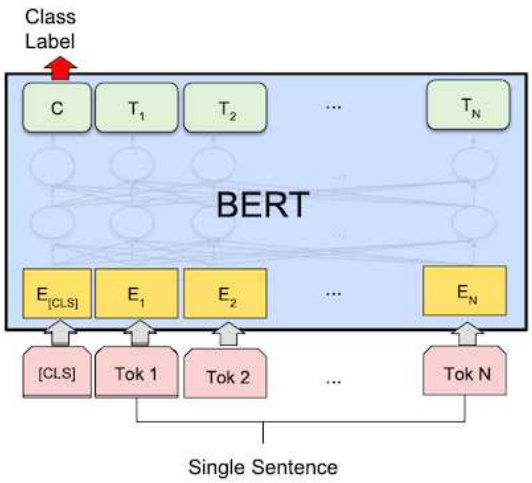

그림 1. BERT 모델 입력 시퀀스 및 토큰 구조

USTC-TFC 2016 데이터셋[4]을 대상으로 ET-BERT 모델을 사용하여 실험을 진행하였다. 그림2는 실험의 결과로 나온 어텐션 스코어를 시각화한 히트맵 이미지 중 하나이다.

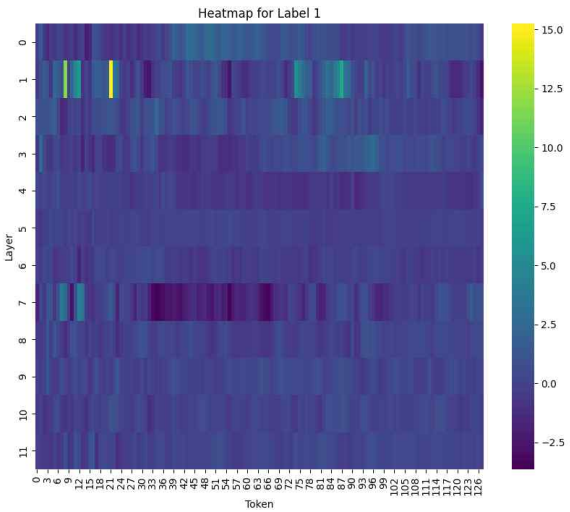

그림 2. 특정 label에 대한 히트맵 이미지

x축은 입력 시퀀스의 각 토큰을 나타내며 총 128개의 토큰으로 구성되어 있음을 의미한다. y축은 트랜스포머 모델의 각 어텐션 레이어를 나타내고 총 12개의 레이어는 각각 독립적인 어텐션 계산을 수행한다. 노란색으로 표시된 일부 토큰 위치에서 특히 높은 어텐션 스코어가 보이는 것을 관찰할 수 있고 이를 통해 모델이 해당 토큰을 매우 중요하게 생각하고 있고 다른 토큰들과의 관계를 강조하고 있음을 확인할 수 있다.

Ⅲ. 결론

현대의 네트워크 환경에서 응용 트래픽 분류는 중요한 과제 중 하나이며, 특히 금융 거래 및 개인 데이터 보호를 위한 데이터 통신의 대부분이 암호화되면서 더욱 복잡해지고 있다. 이러한 응용 트래픽은 다양한 응용 프로그램들이 제공하는 각기 다른 데이터 전송 패턴과 네트워크 행위를 통해 사용자의 디지털 경험을 풍부하게 만들지만, 네트워크 관리자와 보안 분석가에게는 이를 정확하게 분석하고 이해해야 하는 도전과제를 제시한다. 딥러닝과 트랜스포머 모델을 이용한 분석 방법은 높은 정확도로 이러한 복잡한 패턴을 인식할 수 있으나, 모델의 복잡한 내부 매커니즘은 이해와 실시간 적용에 제한을 가져왔다.

이에 본 연구에서는 어텐션 롤아웃기법을 통해 트랜스포머 모델의 어텐션 스코어를 계산하고 히트맵으로 시각화함으로써 모델의 결정 과정을 투명하게 만들었다. 이 기법은 각 입력 토큰이 최종 분류 결정에 미치는 기여도를 추적하여, 응용 트래픽 분석의 투명성과 이해도를 제공하며, 네트워크 보안 분야에서 딥러닝의 적용을 향상시켜 실시간 환경에서의 트래픽 모니터링과 위협 탐지를 더욱 효과적으로 수행할 수 있도록 한다.

참 고 문 헌

[1] Lin, Xinjie, et al. "Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification." Proceedings of the ACM Web Conference 2022. 2022.

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[3] Abnar, Samira, and Willem Zuidema. "Quantifying attention flow in transformers." arXiv preprint arXiv:2005.00928 (2020).

[4] Wang, Wei, et al. "Malware traffic classification using convolutional neural network for representation learning." 2017 International conference on information networking (ICOIN). IEEE, 2017.