패킷 구조 학습을 통한 트랜스포머 기반 응용 트래픽 분류

김주성, 장윤성, 김지민, 백의준, 김명섭
고려대학교
{jsung0514, brave1094, illiard1209, pb1069, tmskim}@korea.ac.kr

Transformer-Based Application Traffic Classification via Packet Structure Learning

Ju-Sung Kim, Yoon-Seong Jang, Jimin Kim, Ui-Jun Baek, Myung-Sup Kim
Korea University

요 약

본 논문은 입력 세션 내 패킷의 구조적 특성을 효과적으로 반영하도록 설계된 입력 표현 방식과 임베딩 방법, 그리고 사전학습 전략을 포함한 새로운 응용 트래픽 사전학습 방법을 제안합니다. 응용 트래픽 분류 태스크에 미세조정한 결과, 제안한 방법은 기존 방법과 비교하여 응용 타입 분류에서 약 7%, 응용 분류에서 약 5%의 정확도를 개선했습니다.

Ⅰ. 서 론

네트워크 트래픽의 정확한 분류는 네트워크 관리 및 보안 분야에 있어 매우 중요한 과제 중 하나이며 딥러닝을 활용한 다양한 연구가 수행되고 있습니다. 특히, 최근 트랜스포머 기반 사전학습 모델이 응용 트래픽 분류 뛰어난 성능을 보이며 최초 응용 트래픽 분류에 사전학습을 적용한 PERT[1] 이래 ET-BERT[2], Lens[3] 등 다양한 입력 표현 방식과 사전학습 전략들이 제안되었습니다. PERT[1]는 패킷 바이트 시퀀스를 정수화된 바이그램 배열로 변환하여 패킷 구조에 관한 특수토큰 없이 모델에 입력하였으며 사전학습 전략으로 MLM(Masked Language Model)을 채택하였습니다. ET-BERT[2]는 세션 내 버스트들의 바이트 시퀀스를 바이그램으로 나누고 이를 BPE(Byte-Pair Encoding)를 통해 토큰화하며 버스트 분할 지점을 나타내는 특수토큰 SEP과 함께 모델에 입력합니다. 또한, ET-BERT는 사전학습 전략으로 Same-Origin Burst Prediction과 Masked Burst Model를 제안하였습니다. ET-BERT 이후의 연구 또한 새로운 구조 표현과 사전학습 방법을 제안하였으나 데이터 입력 단계에서의 패킷 구조적 특성에 관한 고려는 여전히 충분하지 않습니다. 또한, 이들 연구는 학습 과정에서 패킷 내 구조적인 정보 또는 필드 문자열의 맥락을 학습한다고 주장하지만 이에 관한 근거는 미비합니다.

본 논문은 세션 내 패킷의 구조적 특성을 고려한 입력 표현 방식과 임베딩 방법, 그리고 사전학습 방법을 제안합니다. 제안한 방법은 패킷의 구조적 표현을 위해 계층을 표현하는 6개의 특수토큰과 5개의 추가적 임베딩과 각 계층의 구조 정보를 학습하기 위한 7개의 추가적인 사전학습 전략을 제안합니다. 제안한 방법을 기반으로 응용 트래픽 분류 태스크에 미세조정한 결과, 제안한 방법은 기존 방법과 비교하여 응용 타입 분류 정확도에서 약 7%, 응용 분류 정확도에서 약 5%의 성능 향상을 달성했습니다.

Ⅱ. 본론

본 장에서는 제안하는 방법의 전체적 구조는 6개 섹션으로 구성되며 이는 그림 1에 나타나 있습니다.

그림1-(a)는 패킷의 구조적 정보를 학습하기 위한 특수토큰에 대해 나타냅니다. 패킷 계층 구조는 세션, 패킷, 레이어, 필드로 구분되며 계층을 명시적으로 구분하기 위해 CLS(Classification)와 SEP(Seperator)을 확장한 새로운 6개의 특수토큰을 제안합니다. 각 특수토큰들은 계층의 시작 또는 종료를 알리며 토큰 간 계층 차이를 명시적으로 구분하는 역할을 합니다.

그림1-(b)는 각 필드에 대해 레이어 내 물리적 위치 및 각 계층 정보와 관련된 논리적 위치를 알려주기 위한 패킷 구조 벡터를 나타냅니다. 패킷 구조 벡터의 종류는 5개로 Positional은 토큰 순서 번호, Packet Offset은 해당 토큰이 속한 패킷의 순서 번호, Layer Offset은 해당 토큰이 속한 레이어의 순서 번호, Field Offset은 해당 토큰이 특수토큰이 아닌 필드인 경우 해당 필드가 레이어 내 시작하는 위치, Field Length는 해당 필드의 길이를 나타냅니다. 각 패킷 구조 벡터는 임베딩 레이어를 거쳐 528차원으로 확장되며 필드 임베딩 값과 합해져 트랜스포머 인코더에 입력됩니다.

그림1-(c)는 필드 벡터와 특수토큰 벡터에 대해 나타냅니다. 먼저, 필드 벡터는 필드를 구성하는 16진수 바이트 시퀀스를 정수형으로 변환한 것으로 자르거나 패딩하여 고정 크기 n으로 변환됩니다. 값의 범위는 0~255이고 패딩값은 256이며, 만약 n이 1,400이고 필드의 실제 길이가 4바이트라면 1396개의 값은 256으로 채워집니다. 일반 필드의 패딩값과 달리 CLS 토큰에 대한 패딩 값은 257, SEP 토큰에 대한 패딩 값은 258로 이는 일반 필드와 특수토큰의 표현을 명시적으로 구분합니다. 필드 벡터는 패킷 구조 벡터와 달리 이미 2차원 벡터이므로 n 크기의 바이트 시퀀스로 구성된 마지막 차원을 Convolutional Neural Network와 Multi-Head Attention으로 구성된 Custom Embedding Layer을 통해 528 크기의 차원으로 압축되고 모델에 입력됩니다. Custom Embedding Layer의 구조는 그림 2에 나타나 있습니다. Custom Embedding Layer의 입력은 1400 크기의 토큰 벡터이며 임베딩 레이어와 다양한 kernel_size로 구성된 다중 스케일 CNN을 거쳐 지역적 정보를 추출합니다. 지역적 정보는 두 갈래로 나뉘어 이후 Z_{local}과 Z_{global}로 변환됩니다. Z_{local}은 Top-k 샘플링을 추출된 다중 스케일 지역 정보 중 8개의 유용한 벡터만이 샘플링된 지역적 정보 잠재변수입니다. 이는 각 토큰의 바이트 시퀀스 내 일부 문자열 패턴을 유지하며 이후 학습 과정에서 다른 토큰과의 상관관계를 표현할 수 있습니다. Z_{global}는 바이트 시퀀스의 전체적인 맥락을 표현하며 이는

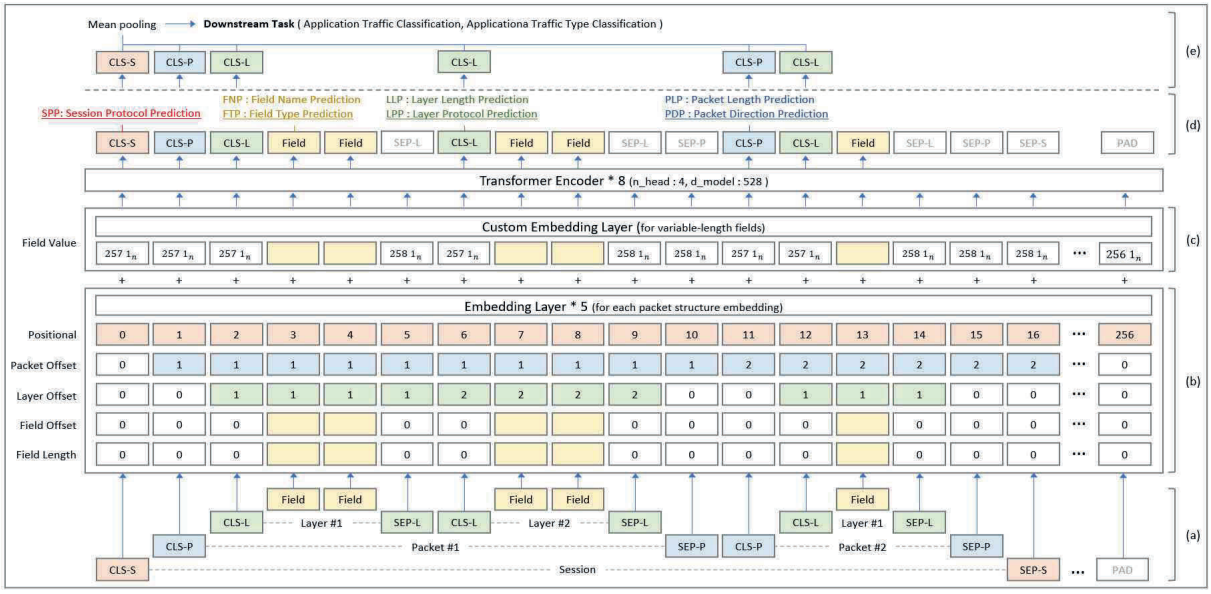

그림 1. 패킷의 구조적 특성을 고려한 입력 표현 방식과 임베딩 방법 그리고 사전학습 방법 개요

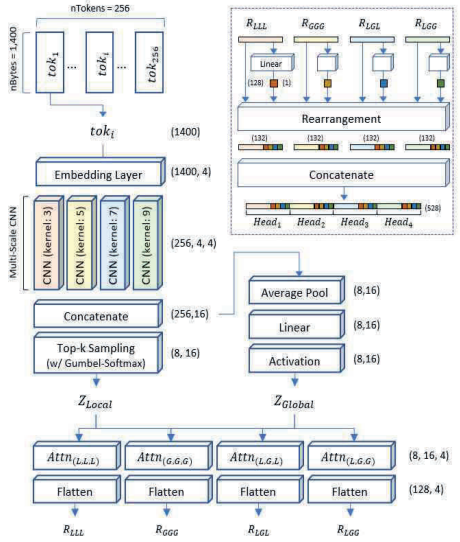

그림 2. 가변 길이 필드 처리를 위한 Custom Embedding Layer 구조

Z_{local}과의 교차 어텐션 연산을 통해 지역적 정보와 전역적 정보 간 관계 표현 R로 변환됩니다. 각 R은 선형 변환된 크기 1의 각 관계 표현의 전역 정보를 포함하게 되며 이는 모두 합쳐져 트랜스포머 인코더의 입력으로 사용됩니다. 또한, 합쳐진 4개의 관계 표현들은 트랜스포머 인코더에서 각 헤드에 대응합니다. 그림1-(d)는 패킷 구조 학습을 위한 사전학습 전략들을 나타냅니다. SPP는 CLS-S 토큰 위치 출력을 바탕으로 해당 세션이 TCP인지 UDP인지 분류합니다. PLP와 PDP는 CLS-P 토큰 위치 출력 바탕으로 해당 패킷의 바이트 수와 방향성을 분류합니다. LLP와 LPP는 CLS-L의 토큰 위치 출력을 바탕으로 해당 레이어의 바이트 수와 응용 프로토콜(TLS, HTTP 등)을 분류합니다. FNP와 FTP는 각 필드 토큰 위치의 출력을 바탕으로 해당 필드의 타입(String, Integer 등)을 예측합니다.

그림1-(e)는 사전학습 이후 두 개의 태스크에 대한 미세조정 과정을 나타내며 각 계층의 CLS 토큰 위치의 출력을 평균 풀링한 값을 바탕으로 미세조정을 수행합니다.

Ⅲ. 실험

제안하는 방법의 평가를 위해 공개 데이터셋인 ISCX-VPN 2016[4]를 사용하였으며 DNS 등의 응용과 관계없는 세션들을 제거하고 IP와 TCP

	분류 정확도		추론시간
	응용 타입 분류	응용 분류	
ET-BERT[2]	73.4%	63.9%	6.6s
Proposed	80.3%	69.1%	4.3s

표 1. 기존 연구와의 정확도 및 추론시간 비교

개의 세션을 사전학습에 사용하였으며 이를 다시 8:2로 분할 후 각각 미세조정 단계의 학습 및 테스트에 사용하였습니다. 실험 결과, 기존 ET-BERT 대비 응용 타입 분류에서 약 7%, 응용 분류에서 약 5%의 정확도를 개선했습니다. 또한, 필드 기반 토큰화를 통해 입력 시퀀스 길이를 기존 연구 대비 절반으로 줄여 추론 시간 또한 크게 개선하였습니다.

Ⅳ. 결론

본 논문은 세션 내 패킷의 구조적 특성을 고려한 입력 표현 방식과 임베딩 방법, 그리고 사전학습 방법을 제안하며 이는 최초입니다. 또한, 가변 길이의 필드를 임베딩하는 방법은 무작위 값으로 구성된 일부 필드에 대한 의존성과 연산 낭비를 줄이며 입력 토큰 시퀀스 길이를 단축하여 추론 시간을 개선하는 효과를 가집니다. 공개 데이터셋으로 평가한 결과, 제안하는 방법은 기존 연구 대비 크게 개선된 정확도를 보였으며 추론 시간 또한 개선합니다. 우리의 향후 연구는 패킷의 구조적인 정보를 학습할 수 있는 추가적인 사전학습 전략 연구입니다.

참 고 문 헌

[1] He, et al. "PERT: Payload encoding representation from transformer for encrypted traffic classification." 2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K). IEEE, 2020.

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

[3] Wang, et al. "Lens: A foundation model for network traffic." arXiv preprint arXiv:2402.03646 (2024).

[4] Gil, et al. "Characterization of encrypted and VPN traffic using time-related features." Proceedings of the 2nd international conference on information systems security and privacy (ICISSP 2016). Setúbal, Portugal: SciTePress, 2016.