바이트 윈도우 분석을 통한 네트워크 트래픽 페이로드의 구조적 특성 분류

김주성, 남승우, 유경민, 장윤성, 김지민, 김명섭*

고려대학교, *고려대학교 세종캠퍼스

{jsung0514, nam131119, rudals2710, brave1094, illiard1209, tmskim}@korea.ac.kr

Classifying Structural Characteristics of Network Traffic Payloads via Byte Window Analysis

Ju-Sung Kim, Seung-Woo Nam, Gyeong-Min Yu, Yoon-Seong Jang, Ji-Min Kim

Myung-Sup Kim*

Korea Univ.

요 약

암호화 트래픽의 증가로 기존 네트워크 트래픽 분석 방법의 한계가 드러나면서 딥러닝 기반 분석 방법이 널리 활용되고 있다. 그러나 기존 방법은 구조적 구간과 랜덤 구간을 구분하지 않고 학습하여 노이즈 과적합 및 일반화 성능 저하 문제가 발생한다. 본 논문에서는 값 생성 메커니즘을 기준으로 네트워크 페이로드를 structured, opaque, mixed로 정의하고, 32B 윈도우와 256B 컨텍스트를 입력으로 CNNEncoder, 명시적 특징, LSTM을 결합한 분류 모델을 제안한다. 실험 결과 제안 모델은 XGBoost 및 CNN 대비 높은 정확도를 달성하였으며, 다양한 데이터셋에서도 일관된 성능을 보여 특정 프로토콜에 의존하지 않는 범용적 분류 가능성을 확인하였다.

Ⅰ. 서 론

네트워크 트래픽 분석은 네트워크를 통해 송수신되는 데이터를 관찰하고 해석하는 기술로, 보안 위협 탐지, 네트워크 성능 모니터링, 이상 행위 식별 등 다양한 목적으로 활용된다. 전통적인 분석 방법은 패킷의 페이로드를 직접 열람하는 DPI(Deep Packet Inspection) 방식에 의존해왔다. 그러나 TLS, QUIC 등 암호화 프로토콜의 확산으로 네트워크 트래픽의 암호화 비중이 크게 증가하면서 페이로드를 직접 분석하는 기존 방법의 한계가 드러났으며, 이를 대체하기 위한 딥러닝 기반 트래픽 분석 방법이 널리 연구되고 활용되고 있다.

그러나 기존 딥러닝 기반 분석 방법은 트래픽 페이로드 내의 구조적 구간과 랜덤한 구간을 구분하지 않고 학습한다. 암호화 또는 압축된 구간의 바이트는 본질적으로 랜덤에 가까운 분포를 가지며, 이러한 구간에서 의미 있는 패턴을 학습하려는 시도는 노이즈 과적합과 일반화 성능 저하로 이어진다. 나아가 구조적 구간은 필드 파싱, 시그니처 매칭 등 내용 기반 분석이 가능하지만, 랜덤 구간은 내용 자체에 의미가 없으므로 분석 대상에서 제외하거나 별도로 처리해야 한다. 이를 구분하지 않으면 분석 자원이 낭비되고 결과의 신뢰성이 저하된다. 따라서 어떤 분석을 수행하든 구간의 구조적 특성을 먼저 판별하는 과정이 선행되어야 한다.

이에 본 논문은 네트워크 트래픽 페이로드를 의미론적 해석 가능성의 관점에서 재정의하고, raw byte sequence를 일정 길이의 윈도우로 분할하여 각 윈도우를 Structured, Opaque, Mixed로 분류하는 문제를 제안한다. 이를 위해 로컬 바이트 패턴을 학습하는 CNN 인코더, 주변 문맥을 반영하는 컨텍스트 인코더, 그리고 FFT 및 엔트로피 기반 수작업 피처를 결합한 모델을 설계하였다.

본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며 (RS-2025-16067207, 양자 네트워크 라우팅 및 제어/관리 핵심 기술과 표준 개발), 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원(TIPA)의 창업성장기술개발사업(TIPS)사업의 지원을 받아 수행된 연구임(RS-2025-25466990, 5G/6G 네트워크 Cross-domain Observability Engineering Orchestrator 기술 및 표준 개발).

Ⅱ. 본론

클래스	포함 필드	필드 생성 기준	예시
Structured	Deterministic	프로토콜 상태, 의미	tcp.flags
Opaque	Random	난수 생성기, 랜덤 알고리즘	tls.random
	Encrypted	암호화	tls.app_data
Mixed	두 종류 클래스 혼재	윈도우-필드 경계 불일치	-

표 1. 클래스 구분 기준

본 논문에서는 네트워크 페이로드를 고정 크기의 윈도우로 분할, 각 윈도우 내 필드 구성에 따라 세 가지 클래스로 정의하고 이는 표 1과 같다. 필드는 값 생성 메커니즘을 기준으로 세 종류로 구분된다. 값이 프로토콜 상태나 의미에 의해 결정되는 Deterministic Field, 난수 생성기 또는 랜덤 알고리즘으로 생성되는 Random Field, 암호화 연산을 거쳐 생성되는 Encrypted Field가 이에 해당한다. 이 중 Deterministic Field만 포함된 윈도우는 Structured, Random Field 또는 Encrypted Field가 포함된 윈도우는 Opaque로 분류한다. 윈도우 경계가 필드 경계와 일치하지 않아 두 종류가 하나의 윈도우에 걸치는 경우 Mixed로 정의한다.

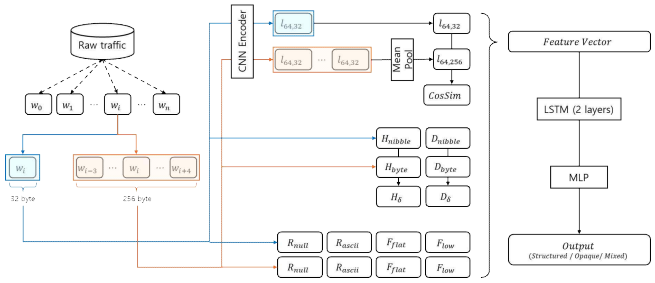

그림 1. 제안하는 방법론의 전체 구조

제안 모델은 세 가지 모듈로 구성되며 이는 그림 1과 같다. CNNEncoder는 32B 윈도우에서 로컬 바이트 패턴을 학습하고, 특징 모듈은 FFT 및 바이트 통계 기반 특징을 명시적으로 추출한다. 추출된 특징은 결합되어 LSTM에 입력되며, LSTM은 세션 내 윈도우 시퀀스 전체의 문맥을 학습하여 각 윈도우의 클래스를 출력한다.

CNNEncoder는 32B 윈도우를 입력으로 받아 로컬 바이트 패턴을 학습한다. 1D Convolution 레이어를 통해 채널을 32에서 64로 확대하고 BatchNorm을 적용하여 학습 안정성을 높이며, 최종적으로 64차원 latent vector를 출력한다. 이를 통해 Structured와 Opaque 구간에서 나타나는 바이트 수준의 지역적 특성을 압축적으로 표현한다. 명시적 특징은 CNN이 학습하기 어려운 구조적 특성을 수식으로 직접 계산하여 제공하며, 32B 윈도우와 256B 컨텍스트 각각에서 독립적으로 추출된다. 분포 기반 특징으로 null 비율과 ASCII 비율을 추출하고, 스펙트럼 기반 특징으로 FFT 저주파 성분, 주파수 집중도, 스펙트럼 평탄도를 추출한다. 엔트로피 기반 특징으로 Shannon 엔트로피와 Jensen-Shannon Divergence를 측정하며, 대비 기반 특징으로 32B 윈도우와 256B 컨텍스트 간의 null 비율, ASCII 비율, 엔트로피, JSD 차이값을 추출한다. CNNEncoder에서 추출한 64차원 latent vector, 39차원 명시적 특징을 결합한 103차원 특징 벡터를 구성한다. 이를 세션 내 윈도우 순서대로 시퀀스로 구성하여 2개 레이어로 구성된 LSTM에 입력한다. LSTM은 앞선 윈도우의 문맥이 현재 윈도우 분류에 반영될 수 있도록 하며, 각 윈도우에 대해 Structured, Opaque, Mixed 중 하나를 출력한다.

Ⅲ. 실험

데이터셋	세션 개수	
	전처리 전	전처리 후
ISCX VPN 2016	310,189	4,802
ISCX Tor 2016	31,367	26,165
CSTNET-TLS1.3	46,372	46,372

표 2. 데이터셋 주요 구성 정보

평가를 위해 총 3종의 네트워크 트래픽 공개 데이터셋을 사용하였으며 각각 ISCX VPN 2016[1], ISCX Tor 2016[2], CSTNET-TLS1.3[3]의 3가지이다. 전처리 단계에서는 Ethernet 헤더를 제거하고, IPv4 TCP/UDP 세션만을 사용하였다. 그리고 3-way 핸드셰이크가 불완전한 세션, 응용과 관계가 없는 세션들을 제거하였으며, 이를 다시 8:2로 분할 후 학습 및 테스트에 사용하였다. 표 2는 실험에 사용한 데이터셋들의 전처리 전후 세션 개수를 정리한 것이다.

데이터셋	클래스 별 윈도우 분포			
	Structured	Opaque	Mixed	Total
ISCX VPN 2016	397,353	103,838	3,460	504,651
ISCX Tor 2016	1,561,288	86,580	7,676	1,655,544
CSTNET-TLS1.3	3,268,322	3,009,299	279,108	6,556,729

표 3. 데이터셋 클래스 별 윈도우 분포

세션의 raw byte sequence를 32B 단위로 고정 분할하여 윈도우를 생성한다. 각 윈도우에 대해 앞 3개, 뒤 4개의 인접 윈도우를 결합하여 256B 컨텍스트 윈도우를 구성한다. 세션의 시작 또는 끝에 위치하여 인접 윈도우가 부족한 경우 0으로 패딩하였으며, 각 바이트 값은 0~1 범위로 정규화하여 모델에 입력한다. 표 3은 데이터셋 별로 전처리 후 생성된 윈도우의 클래스별 분포를 나타낸 것이다.

데이터셋	AC	PR	RC	F1
ISCX VPN 2016	94	94	93	93
ISCX Tor 2016	96	96	95	96
CSTNET-TLS1.3	98	97	98	97

표 4. 제안하는 방법론의 분류 실험 결과

표 4는 제안하는 방법론의 분류 성능을 Accuracy, Precison, Recall, F1-score 기준으로 나타낸 것이다. 제안하는 방법론은 전체 지표에서 고른 성능을 보이며 Structured와 Opaaque 구간을 효과적으로 분류함을 확인하였다. 이는 특정 프로토콜이나 데이터셋에 의존하지 않고 다양한 트래픽 환경에서도 일관된 성능을 보임을 나타낸다.

데이터셋	분류 정확도		
	XGB	CNN	Proposed
ISCX VPN 2016	51	62	94
ISCX Tor 2016	55	65	96
CSTNET-TLS1.3	59	67	98

표 5. 타 분류 모델과의 정확도 비교 실험 결과

표 5는 동일한 데이터셋, 동일한 특징 벡터를 대상으로 다른 분류 모델을 통해 비교 실험을 진행한 결과이다. 실험 결과, 3가지 데이터셋에서 XGB 및 CNN 대비 제안하는 방법론을 사용한 쪽이 모두 더 높은 정확도를 보였다. 이는 명시적 특징과 LSTM을 결합하여 세션 내 문맥을 반영한 것이 분류 성능 향상에 효과적임을 보여준다.

Ⅳ. 결론

본 논문에서는 네트워크 트래픽 페이로드의 바이트 구조적 특성을 분류하기 위해 CNNEncoder, 명시적 특징, LSTM을 결합한 모델을 제안하였다. 값 생성 메커니즘을 기준으로 Structured, Opaque, Mixed를 정의하고, 32B 윈도우와 256B 컨텍스트를 함께 활용하여 로컬 패턴과 문맥 정보를 동시에 학습하였다. 실험 결과 제안 모델은 XGBoost 및 CNN 대비 높은 정확도를 달성하였으며, 다양한 데이터셋에서도 일관된 성능을 보여 특정 프로토콜에 의존하지 않는 범용적 분류 가능성을 확인하였다. 다만 본 연구는 윈도우 크기를 32B로 고정하였으며, 세션 내 필드 경계 정보를 사전에 알고 있다는 가정 하에 레이블을 부여하였다는 한계가 있다. 향후 연구에서는 가변 크기 윈도우 적용 및 레이블 없는 환경에서의 비지도 학습 방식으로 확장하는 것을 목표로 한다.

참 고 문 헌

[1] 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). Setubal, Portugal: SciTePress, 2016.

[2] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun and Ali A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features", In the proceeding of the 3rd International Conference on Information System Security and Privacy, SCITEPRESS, Porto, Portugal, 2017.

[3] 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.