2D CNN과 STFT를 활용한 암호화된 WeChat 트래픽의 행위 분류

김지민, 남승우, 유경민, 장윤성, 김주성, 김명섭*

고려대학교

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

STFT-based Behavior Classification in Encrypted Wechat Traffic Using 2D CNN

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

Myung-Sup Kim*

Korea Univ.

요 약

TLS 및 mmTLS와 같은 강력한 암호화 기법의 보편화로 인해 페이로드 기반 트래픽 분석이 어려워지고 있다. 특히 WeChat과 같은 통합 메신저는 동일 세션 내에서 여러 행위가 혼재되어 발생하므로, 기존의 세션 단위 분석으로는 행위 단위 식별이 불가능하다. 본 연구에서는 네트워크 트래픽을 시간 신호로 해석하고, Short-Time Fourier Transform (STFT)을 적용하여 시간-주파수 도메인에서 행위별 고유 패턴을 추출하는 방법을 제안한다. 추출된 5 채널 스펙트로그램을 2D CNN으로 분류한 결과, WeChat 트래픽에서 4개 행위 클래스에 대해 98.9%의 분류 정확도를 달성하였다.

Ⅰ. 서 론

현대의 메신저 애플리케이션 중 하나인 WeChat은 텍스트 채팅과 파일 전송이 단일 세션 내에서 통합적으로 이루어진다. TLS v1.3 및 mmTLS 등 강력한 암호화 기법이 표준화되면서 패킷 페이로드를 통한 직접적인 트래픽 분석이 사실상 불가능해졌다. 이러한 환경에서도 네트워크 운영, 보안 모니터링, QoS 정책 적용 등을 위해 트래픽의 종류를 식별할 필요성은 지속적으로 증가하고 있다.

기존 트래픽 분류 연구는 주로 플로우 또는 세션 단위에서 이루어져 왔으나, 이 접근법은 동일 세션 내에서 발생하는 서로 다른 사용자 행위를 구분하지 못한다는 근본적인 한계를 갖는다. 본 연구는 행위마다 서로 다른 패킷 전송 주기성을 갖는다는 가정에 기반하여, 트래픽을 시간 영역의 신호로 해석하고, STFT를 통해 시간-주파수 도메인으로 변환하는 새로운 분석 방법을 제안한다.

Ⅱ. 관련연구

Wang 등[1]은 세션 단위로 트래픽을 2D 이미지로 변환한 뒤 CNN으로 분류하는 방법을 제안하였으나, 동일 세션 내 행위 구분은 불가능하다. 패킷 번들 단위 1D CNN 기반 연구[2]는 시간 bin 별 통계 특징을 활용하여 약 97.6%의 행위 분류 정확도를 보고하였으나, 시간 도메인의 특성만 사용하여 주기적 패턴이 명시적으로 드러나지 않는다는 한계가 있다. KL divergence 기반 sliding window 기법[3]은 분포 변화 시점만을 탐지할 뿐 행위별 고유 특성을 학습하지 못한다.

본 연구는 음성·영상 처리 분야에서 검증된 STFT 스펙트로그램[4]을 네트워크 트래픽 분석에 적용함으로써, 시간-주파수 동시 분석을 통한 해석 가능한 행위 분류를 실현한다는 점에서 차별성을 갖는다.

Ⅲ. 본론

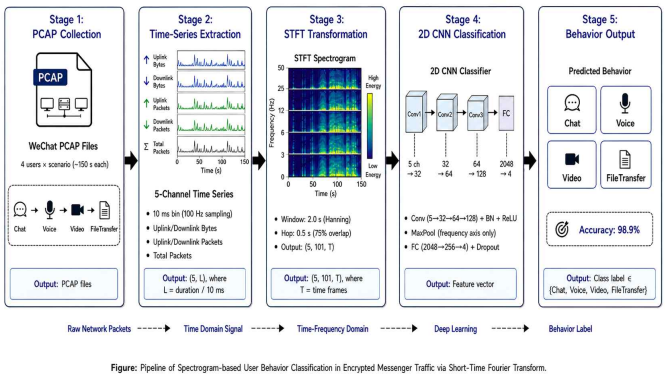

Figure: Pipeline of Spectrogram-based User Behavior Classification in Encrypted Messenger Traffic via Short-Time Fourier Transform.

그림 1. STFT 기반 행위 분류 시스템 구조도

3.1 시계열 정보 추출

수집된 PCAP 파일을 10ms 단위 bin으로 분할하여 5채널 시계열 정보를 추출한다. 채널 구성은 (1) 업링크 바이트, (2) 다운링크 바이트, (3) 업링크 패킷 수, (4) 다운링크 패킷 수, (5) 전체 패킷 수이다. 추출된 시계열 정보 그래프는 아래 그림 2와 같다.

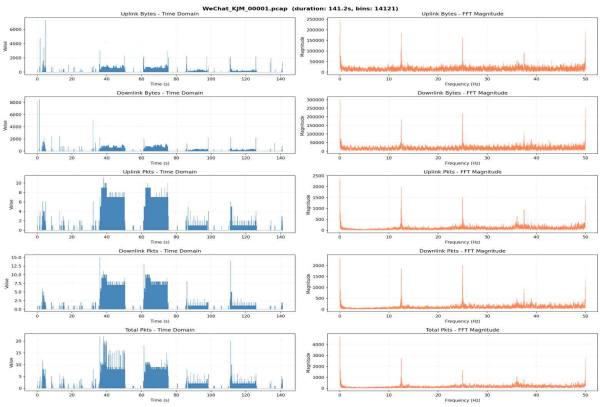

그림 2. 시계열 정보 추출 그래프

이 논문은 2026년도 정부(산업통상부)의 재원으로 한국산업기술진흥원의 지원(P0031136, 2026년 중부권 지역혁신클러스터육성)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-16067207, 양자 네트워크 라우팅 및 제어/관리 핵심 기술과 표준 개발)

3.2 STFT 변환

각 채널 시계열에 대해 길이 2초(200샘플)의 Hanning 윈도우와 0.5초 hop(75% 오버랩)을 적용하여 STFT를 수행한다. 채널별로 0~50Hz 범위의 101개 주파수 bin을 갖는 스펙트로그램이 생성되고, 최종 입력은 (5,101,T) 형태의 3차원 텐서이다. 생성된 스펙트로그램은 그림 3과 같다.

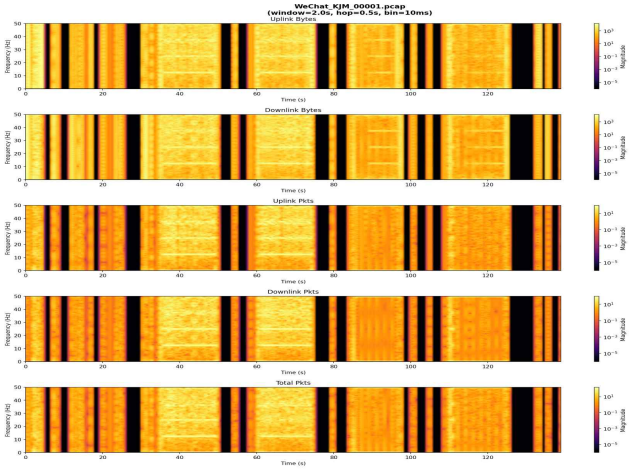

그림 3. STFT 변환 스펙트로그램

3.3 라벨링 및 데이터셋

WeChat을 대상으로 약 150초 길이의 시나리오를 4 명의 사용자로부터 수집하였다. 사람에 의한 시나리오 진행에서 발생하는 1~2초 시간 오차를 보정하기 위해 각 행위 경계에 3초의 마진을 적용하여 확실한 구간만 학습에 사용하였다. STFT 스펙트로그램에서 4프레임 단위 슬라이딩 윈도우로 샘플을 추출한 결과, Chat 170개, Voice 1,360개, Video 1,020개, FileTransfer 163개, 총 2,713개의 샘플을 얻었다.

3.4 2D CNN 분류 모델

분류기는 3개의 Convolution Layer와 2개의 Fully Connected Layer로 구성된다. 각 Layer는 3x3 Convolution, BatchNorm, ReLU로 이루어지며, 채널 수는 5→32→64→128로 증가한다. 풀링은 주파수 축에 대해서만 2x1 MaxPool을 적용하여 시간 축을 보존하였다. 이는 입력 텐서의 시간 차원이 4프레임으로 매우 작아 추가 다운샘플링 시 정보 손실이 발생하기 때문이다. 마지막 레이어 이후에는 4x4 크기의 AdaptiveAvgPool을 적용하여 다양한 시간 길이의 입력에 대해 일정한 특징 맵 크기를 보장하였다.

분류기는 평탄화된 2,048차원 특징 벡터를 256차원으로 축소하는 Fully Connected Layer, 과적합 방지를 위한 50% Dropout, 그리고 4개 클래스로 매핑하는 최종 출력층으로 구성된다.

Ⅳ. 실험결과

8:2 stratified 분할로 학습/평가를 수행한 결과, 최고 테스트 정확도 98.9%를 달성하였다. 표 1은 클래스별 분류 성능을 나타낸다.

클래스	Precision	Recall	F1
Chat	1.00	0.88	0.94
FileTransfer	0.85	1.00	0.92
Video	1.00	0.99	0.99
Voice	1.00	1.00	1.00
Weighted Avg	**0.99**	**0.99**	**0.99**

표 1. 클래스별 분류 성능

클래스별로 살펴보면 Voice와 Video에서 두드러진 성능이 관찰되었다. Voice 클래스는 F1-score 1.00을 기록하였으며, Video 클래스는 F1-score 0.99를 달성하였다. 반면 Chat은 0.94, FileTransfer는 0.92로 상대적으로 낮은 성능을 나타냈다.

혼동 행렬 분석 결과, 모든 오분류는 FileTransfer로 집중되어 발생하였다. 이유는 Chat과 FileTransfer의 시계열 정보 추출 그래프를 관찰해보면, Chat과 혼동될 수 있는 유사한 비주기적 burst 패턴이 두 클래스에서 공통적으로 나타나기 때문이다. 따라서 두 클래스를 구분할 수 있는 추가적인 통계적 특징의 도입이 필요할 것으로 판단된다.

Ⅴ. 결론

본 연구는 암호화된 메신저 트래픽에서 페이로드를 사용하지 않고 사용자 행위를 분류하는 STFT 기반 시간-주파수 분석 방법을 제안하였다. 트래픽을 다채널 시계열 신호로 해석하고 STFT 스펙트로그램으로 변환한 뒤 2D CNN으로 분류한 결과, WeChat 트래픽 4개 행위에 대해 98.9%의 정확도를 달성하였다.

본 연구의 기여점은 (1) 네트워크 트래픽을 다채널 시간 신호로 해석하는 새로운 관점을 제시하였고, (2) 음성·영상 처리에서 검증된 STFT를 트래픽 행위 분류에 적용하여 시간-주파수 동시 분석의 유효성을 입증하였으며, (3) 동일 세션 내 다중 행위 식별이라는 기존 세션 단위 분석의 한계를 해소하였다는 점이다.

오분류는 대부분 Chat과 FileTransfer 사이에서 발생하였으며, 이는 STFT 기반 분석이 행위에 따라 식별 난이도가 다르게 나타남을 보여준다. 향후 연구에서는 Chat과 FileTransfer 간 구분 성능을 향상시키기 위한 추가 특징 추출 및 다양한 메신저 어플리케이션에서의 확장 실험이 필요하다.

참 고 문 헌

[1] Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware traffic classification using convolutional neural network for representation learning. In *Proceedings of the 2017 International Conference on Information Networking (ICOIN)* (pp. 712 - 717). IEEE.

[2] Ji-Min Kim, Myung-Sup Kim, Ran-A Kim, Yoon-Seong Jang and Ui-Jun Baek. (2025). WeChat Behavior Analysis with Session and Packet Bundle Level. KNOM Review, 28(2), 33-40.

[3] Li, Ding & Li, Wenzhong & Wang, Xiaoliang & Nguyen, Cam-Tu & Lu, Sanglu. (2020). App Trajectory Recognition over Encrypted Internet Traffic based on Deep Neural Network. Computer Networks. 179. 107372. 10.1016/j.comnet.2020.107372.

[4] HERSHEY, Shawn, et al. CNN architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017. p. 131-135