패킷 표현 학습에서 플로우 별 분할 기반 평가: 암호화 트래픽 분류에서 데이터 스누핑과 페이로드 의존성

Per-Flow Split Evaluation for Packet Representation Learning: Data Snooping and Payload Dependency in Encrypted Traffic Classification

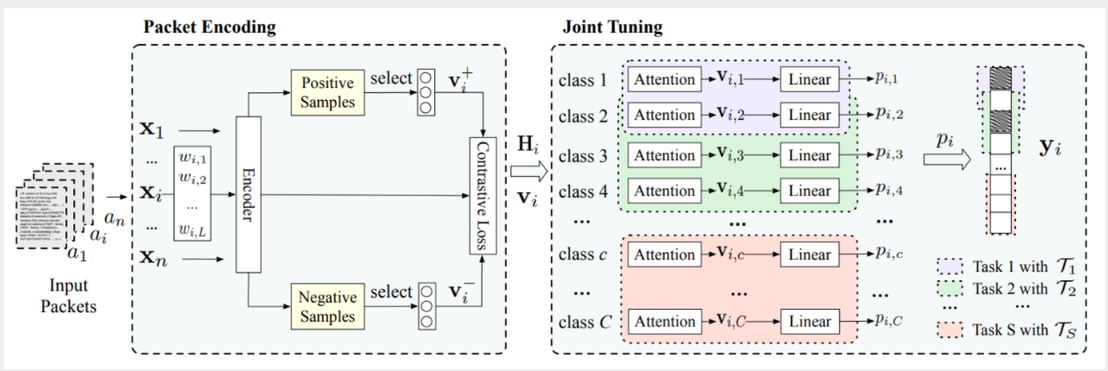
인하대학교 컴퓨터공학과 김강욱(ices17@inha.edu), 노희준*(hjroh@inha.ac.kr)

인하대학교 INHA UNIVERSITY

서론

현대의 인터넷에서 대부분의 패킷의 페이로드가 암호화되어 교환되면서, 기존 심층 패킷 검사 기반의 트래픽 분류 방식의 한계를 극복하고자 딥러닝을 활용한 트래픽 분류 연구가 활발히 진행되고 있다. 특히 트랜스포머 구조의 등장 이후, 암호화된 트래픽 분류 문제에 이를 적용한 분류 프레임워크가 제안되고 있다. 이중 사전학습된 인코더를 활용한 모델들은 여러 태스크에서 높은 분류 성능을 보인다고 알려졌었으나, 최근 연구에서 ET-BERT의 성능 평가 과정에 문제가 있음이 보고되었다. 본 연구에서는 사전학습된 인코더를 활용하면서도 ET-BERT와는 다른 방식으로 동작하는 PacRep 또한 성능 평가 과정에서 데이터 스누핑 오류를 범했음을 간접적으로 확인하는 실험을 설계, 과적합 가능성이 있음을 확인한다. 또한, 암호화된 페이로드를 입력으로 활용하는 PacRep의 분류 성능이 페이로드에 얼마나 의존하는지 실험을 통해 살펴보고, 과적합 가능성에 대해 논의한다.

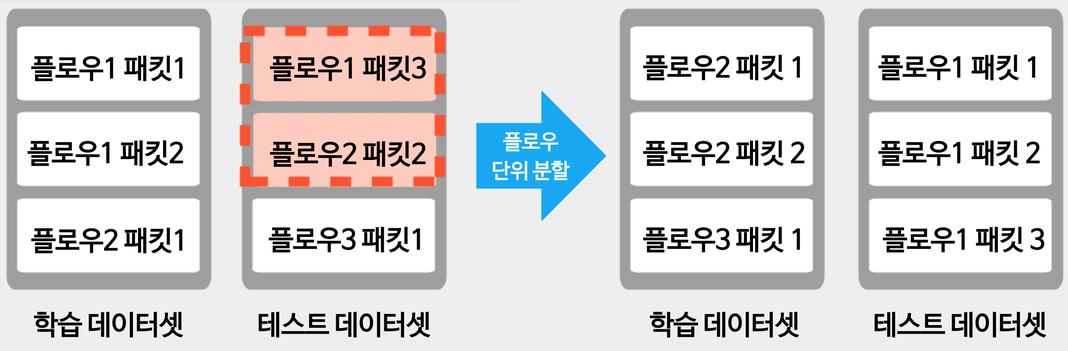
PacRep



PacRep 모델 구조[1]

PacRep [1]은 라벨링된 패킷을 사용하여 대조 학습 기반의 패킷 표현을 학습한다. 먼저 기준 패킷(anchor)을 하나 선택한 뒤, 같은 라벨의 패킷(positive)과 다른 라벨의 패킷(negative)을 각각 하나씩 샘플링하여 트리플렛을 구성한다. 트리플렛에서 각 패킷은 Scapy가 디코딩 가능한 헤더가 있는 경우 필드명과 값을 문자열로 변환하고, 디코딩 가능하지 않는 (암호화된) 페이로드는 16진수 문자열로 변환하여, 패킷 하나가 하나의 문자열에 대응되도록 한 뒤, 사전학습된 트랜스포머 인코더에 입력된다. 모델은 anchor와 positive는 가깝고 negative와는 멀어지도록 학습하며 패킷의 표현을 추출한다. 추출한 패킷의 표현을 분류기에 입력하여 활성화된 태스크에 대해 각 클래스의 확률을 계산하고, 가장 높은 확률의 클래스로 최종 분류한다.

데이터 스누핑



학습 데이터셋과 테스트 데이터셋 모두에 같은 플로우에 속한 패킷이 존재하게 되면 **학습 과정에서 테스트셋의 정보가 노출되는 문제**가 발생한다. 이러한 노출은 **모델의 실제 분류 성능을 왜곡**할 가능성이 있다. 기존 PacRep 실험에서는 **데이터셋을 패킷 단위로 분할**하였는데, 이러한 **패킷 단위 분할 방식은 데이터 스누핑 오류가 발생할 수 있으므로** 이를 검증하기 위한 **플로우 단위로 분할 실험을 진행**한다.

데이터셋

데이터셋	ISCXVPN2016		DoHBrw2020		USTCTFC2016	
태스크	태스크1	태스크2	태스크3	태스크4	태스크5	태스크6
분류 기준	VPN/ NonVPN	애플리 케이션	정상/공격	공격 트래픽 발생 도구	정상/공격	애플리 케이션
클래스 수	2	16	2	5	2	20

플로우 단위 분할

패킷 단위 분할	ISCXVPN2016		DoHBrw2020		USTCTFC2016	
	태스크1	태스크2	태스크3	태스크4	태스크5	태스크6
정확도	1.0000	0.9833	1.0000	0.9680	1.0000	0.9633
F1	1.0000	0.9828	1.0000	0.9744	1.0000	0.9744

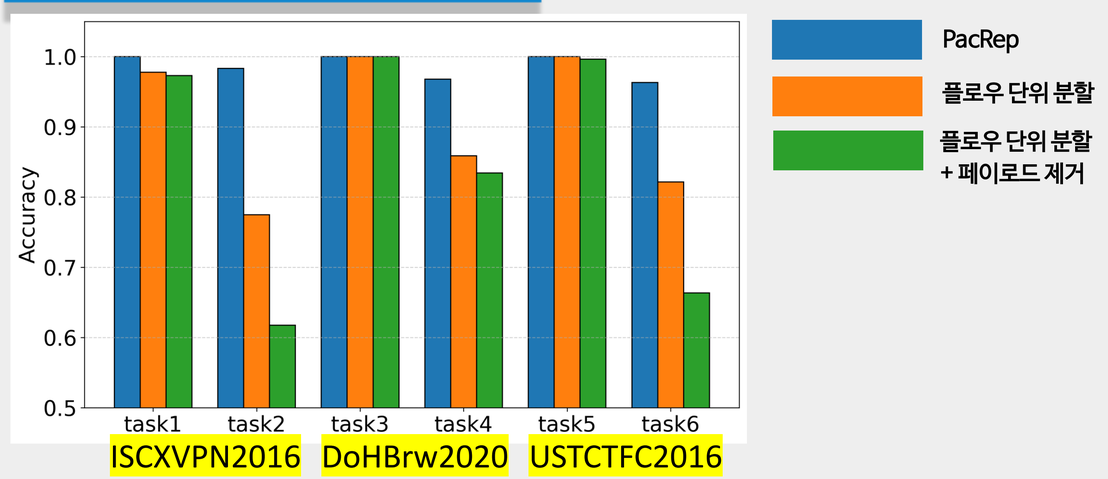
PacRep 재현 실험

플로우 단위 분할	ISCXVPN2016		DoHBrw2020		USTCTFC2016	
	태스크1	태스크2	태스크3	태스크4	태스크5	태스크6
정확도	0.9779	0.7747	1.0000	0.8588	1.0000	0.8215
F1	0.9774	0.7191	1.0000	0.8622	1.0000	0.8468

플로우 단위 분할 실험

PacRep의 실험을 논문대로 재현해본 결과, **분할된 데이터셋에서 데이터 스누핑이 발생하는 것을 확인**하였으며, 플로우 단위로 분할하여 추가 실험을 해본 결과 성능이 하락하였기에, **PacRep의 기존 성능은 데이터 스누핑 오류를 범했**음을 알 수 있다.

페이로드 제거



한편, 암호화 알고리즘이 강건하다는 전제하에 암호화된 페이로드의 학습은 타당하지 않을 수 있다 [2]. PacRep은 패킷의 헤더 정보뿐만 아니라 페이로드를 학습에 사용하고 있으므로, 페이로드를 제거한 상황의 성능을 평가해보았는데, 다중 클래스 분류(태스크 2, 4, 6)에서 상당한 성능 하락이 관찰되었다. 이는 기존의 PacRep이 암호화된 페이로드로부터 특정 패턴을 학습하여 과적합이 발생했을 가능성을 시사한다. 실제로, ISCXVPN2016과 USTCTFC2016에는 암호화되지 않는 트래픽이 다수 포함되어 있음이 알려져있다.

결론

PacRep을 플로우 단위 기반으로 데이터를 재구성하여 실험하였을 때 분류 성능이 하락하였으며, 이는 기존 PcaRep의 실험에서 데이터 스누핑으로 인한 과적합이 발생했을 가능성을 시사한다. 또한, PacRep이 암호화된 페이로드에 대한 의존성을 확인하여, 데이터셋의 암호화된 페이로드에 내포된 구조적 패턴을 학습했을 가능성을 논의하였다. 후속 연구에서는 보다 통제된 데이터셋을 통해 모델의 성능 평가가 이루어질 필요가 있다.

이 성과는 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00235509, ICT융합 공공 서비스·인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)(*:교신 저자)

참고 문헌.
[1] X. Meng, Y. Wang, R. Ma, H. Luo, X. Li, and Y. Zhang, “Packet Representation Learning for Traffic Classification,” in *Proc. of ACM KDD*, 2022.
[2] Y. Zhao, G. Dettori, M. Boffa, L. Vassio, and M. Mellia, “The Sweet Danger of Sugar: Debunking Representation Learning for Encrypted Traffic Classification,” in *Proc. of ACM SIGCOMM*, September 2025.