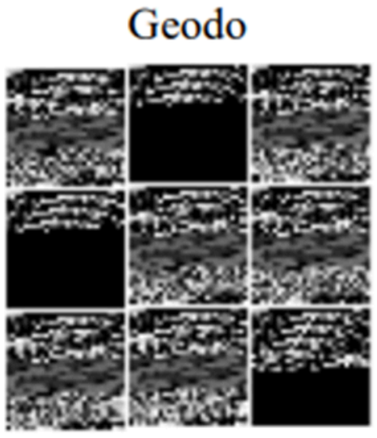
세션 슬라이싱 기법을 활용한 CNN 기반 악성 트래픽 분류 연구

지도교수 : 김명섭 교수님

2022270649 양지윤
2022271329 정고은

[연구 목표]
"세션 전체 정보를 활용한 입력 구조로 기존 연구의 한계를 개선하고, 악성 트래픽 분류 정확도를 향상시키는 것을 목표로 한다"



기존 연구의 한계

- 기존 논문은 세션의 앞부분 784B만을 CNN 입력으로 활용함
- 그러나 짧은 세션(예: Geodo)의 경우 과도한 패딩이 발생하여 학습이 왜곡됨
- 또한 세션 중·후반부의 중요한 특징을 반영하지 못해 모델의 일반화 성능에 한계가 존재함

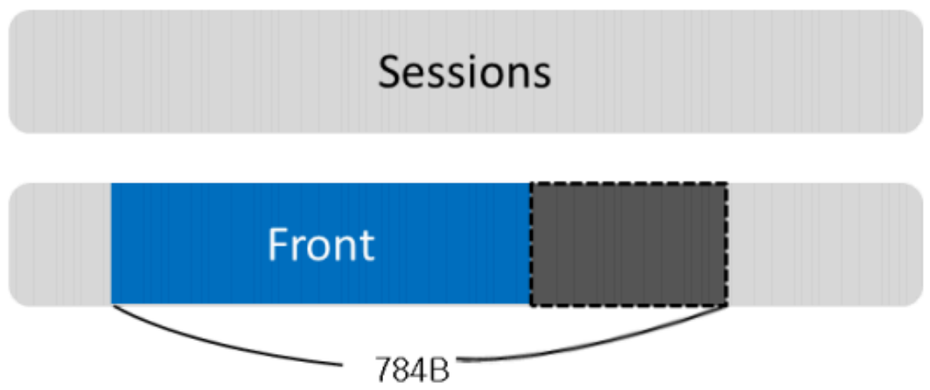
「cf) 기존논문 : Malware Traffic Classification using Convolutional Neural Network for Representation Learning」

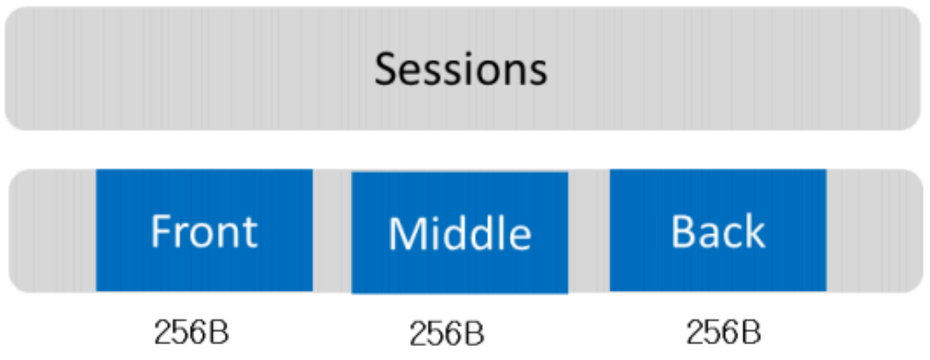
구현 방법

■ Payload 길이 정의
- TCP 3-way handshake 및 ACK-only 패킷 제거
- 실제 데이터(payload)만 이어붙인 총 길이 L 계산

■ 256B × 3 구간 슬라이싱 규칙
- Front: [0 ~ 256)
- Mid: [mid-128, mid+128], mid = L // 2
- Back: [L-256 ~ L)
- → 일부 겹침 허용 / 불필요한 패딩 제거하여 효율 극대화

■ CNN 입력 구성
- Front / Mid / Back 3구간을 각각 R·G·B 채널로 매핑
- → 3채널 이미지 형태의 입력 데이터 생성





연구 절차

1. 데이터 수집
   - USTC-TFC2016 데이터셋 사용
2. 데이터 전처리
   - 새로운 세션 슬라이싱 데이터 생성
   - Payload 재정의
   - Front/Mid/Back 256B 슬라이싱
   - RGB 3채널 변환
3. 모델 설계 및 학습
   - RGB 3채널 입력의 CNN
4. 성능 평가
   - 동일 조건에서 기존 방식 784B 입력 vs 제안 방식 256×3 입력 비교
   - 평가 지표: Accuracy, Precision, Recall, F1-score
   - 특히 짧은 세션에서 성능 개선 여부 집중 검증
5. 결과 분석

기대 효과

- 패딩 문제 완화 → 입력의 효율성 증가
- 세션의 다양한 위치 정보 반영 → 더 높은 정확도 및 일반화 성능 확보
- CNN이 실제 트래픽 패턴을 충실히 학습 → 악성 트래픽 탐지 성능 향상
- 입력 크기 단순화로 연산 효율 증가 → 실시간 탐지 시스템 적용성 확대

개발 tool

python™ PyTorch scikit learn scapy