# 응용 트래픽 분류 공개 데이터셋의 전처리 방법

백의준, 박지태, 김주성, 남승우, 박재원, 김명섭
고려대학교
{pb1069, pjj5846, jsung0514, nam131119, 2018270614, tmskim}@korea.ac.kr

# Preprocessing Open Dataset in Application Traffic Classification

Ui-Jun Baek, Jee-Tae Park, Ju-Sung Kim,
Seong-Woo Nam, Jae-Won Park, Myung-Sup Kim
Korea University

요 약

데이터 전처리는 원시 데이터를 분석에 적합한 형식으로 변환하거나 잡음 또는 이상치를 처리하는 것으로 데이터 분석과 기계학습에서 중요한 단계이다. 적절한 데이터 전처리는 정확한 분석 결과 제공, 모델 성능 향상, 모델 일반화 능력 개선, 빠른 모델 학습 등의 이점을 제공한다. 공개 데이터셋인 ISCX-VPN-nonVPN 2016은 응용 트래픽 분류 분야에서 널리 사용되는 공개 데이터셋으로 많은 연구들에서 이 데이터셋을 사용하여 제안하는 방법의 성능을 평가하고 있다. 그러나, 각 연구마다 데이터셋 전처리에 다른 방법을 사용하며 자세한 설명이나 전처리된 데이터셋을 공유 하지 않아 방법론들 간 객관적인 성능 평가가 어렵다. 본 논문은 공개 데이터셋 ISCX-VPN-nonVPN 2016 전처리에 대한 기준을 제시하고 전처리 과정을 설명하고 전처리된 데이터셋을 공유하는데 목적을 둔다.

## Ⅰ. 서 론

복잡해지는 인터넷 환경과 급격하게 증가하는 응용 개발에 따라 네트워크 관리 분야에서 응용 트래픽 분류 기술에 대한 수요가 증가하고 있습니다. 응용 트래픽 분류의 공개 데이터셋인 ISCX VPN-nonVPN 2016은 응응 트래픽 분류 분야에서 검증 및 평가를 위해 사용되고 있습니다 [1]. 그러나, 각 연구마다 다른 전처리 방법이 적용되고 있으며 재현 가능성을 위한 자세한 설명이 누락되어 있으며 전처리된 데이터셋을 공개하지 않습니다. 기계학습 분야에선 데이터 전처리 방법에 따라 결과가 상당히 다를 수 있기에 각 연구간 비교를 위해 명확한 전처리 기준을 제시하고 해당 데이터셋을 공개할 필요가 있습니다.

본 논문에서는 응용 트래픽 분류 분야에서 널리 사용되는 ISCX VPN-nonVPN 2016 공개 데이터셋을 전처리하는 기준을 제시하고 데이터셋을 공개합니다. 제안된 전처리 방법은 프로토콜과 세션 내 3-hand-shake 여부를 고려하여 데이터를 전처리하여 응용 트래픽 분류의 성능을 향상시킵니다.

## Ⅱ. 관련 연구

[2]는 응용 트래픽 분류를 위한 다중 모달 멀티태스크 딥러닝 접근 방법인 DISTILLER 분류기를 제안하였으며 전처리를 위해 전체 플로우 내 단일 UDP 또는 대상 (IP 주소, 포트)가 (255.255.255.255, 10505)인 플로우를 제거하여 11.6K개의 양방향 플로우를 추출합니다. 그러나, 이는 Bluestack, Android Emulator와 같은 데이터 수집을 위한 기반 응용에서 생성된 플로우만 고려하였으며, 운영체제 또는 트래픽 수집을 위한 네트워크 구성에 따라 발생한 플로우는 고려하지 않았습니다.

[3]은 딥러닝 기반의 응용 트래픽 분류를 위한 초기 분류기로 ISCX VPN-nonVPN 2016을 사용하였으나 전처리 방법에 대해 설명하지 않았으며 샘플 수만 공개하였습니다. [4]는 플로우 데이터를 이미지 기반의 FlowPic으로 변환하여 CNN 기반 모델로 분류합니다. 이 연구에서는 응용 카테고리와 일치하지 않는 세션 또는 잘못된 패킷을 제거하였으나 세부 과정 및 기준이 명확히 명시되지 않습니다. 현재 응용 트래픽 분류 분야에서는 데이터 전처리에 대한 명확한 기준이 없으며 서로 다른 연구 간 객관적인 평가 및 비교가 어렵다는 도전 과제가 있습니다.

## Ⅲ. 프로토콜 및 3-hand-shake 기반 전처리 방법

공개 데이터셋인 ISCX VPN-nonVPN 2016은 각 응용 이름 별 파일 (<application_name>.pcap)로 구성되어 있으며 세 가지 작업으로 구성되어 있습니다. 각 pcap 파일은 여러 플로우로 구성되어 있으며 이를 세션 단위로 구분하기 위하여 pcap 분할 도구인 *Splitcap*을 사용하여 양방향 플로우로 분할합니다. 이후, 분할된 플로우에 대하여 두 가지 기준을 바탕으로 데이터 청소를 수행합니다. 첫 번째 기준은 각 플로우에서 응용 계층 프로토콜을 수집하고 해당 프로토콜이 응용의 실제 동작과 관련 있는지를 판단합니다. 예를 들어, nbss 프로토콜은 윈도우 운영체제에서만 사용되는 프로토콜로 응용의 실제 동작과 관련있다고 보기 어렵습니다. 두 번째로 TCP 플로우 중 3-hand-shake가 보존되지 않는 플로우를 제거합니다. 플로우 내 패킷의 오프셋이 일관되지 않으면 모델이 트래픽을 학습할 때 플로우 내 패킷의 시계열성을 고려하기 어렵습니다. 또한, 분류 모델이 실제 환경에 적용될 때 실시간으로 수집되는 트래픽 데이터셋은 3-hand-shake 과정이 보존되어 있다고 보는 것이 적합하므로 3-hand-shake가 보존되지 않는 플로우는 노이즈로 판단할 수 있습니다.

본 과제(결과물)는 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구 (No. 20008902. IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발) 이며 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과 (2021RIS-004)입니다.

| 네트워크 계층 | 응용 계층 | chat | email | file transfer | p2p | streaming | voip | 계 |
|---|---|---|---|---|---|---|---|---|
| TCP | data | 47 | 32 | 221 | 0 | 6 | 254 | 560 |
| | ftp | 0 | 0 | 22 | 0 | 0 | 0 | 22 |
| | http | 56 | 11 | 653 | 225 | 99 | 330 | 1374 |
| | stun | 0 | 0 | 0 | 0 | 0 | 93 | 93 |
| | tcp | 57 | 19 | 641 | 2 | 593 | 332 | 1644 |
| | tls | 361 | 401 | 463 | 8 | 854 | 1283 | 3370 |
| UDP | data | 2010 | 1353 | 5563 | 127 | 18 | 14651 | 23722 |
| | dtls | 0 | 0 | 1 | 0 | 0 | 14 | 15 |
| | gquic | 27 | 26 | 24 | 0 | 100 | 46 | 223 |
| | stun | 0 | 0 | 0 | 0 | 0 | 277 | 277 |
| 계 | | 2,558 | 1,842 | 7,588 | 364 | 1,670 | 17,280 | 31,300 |

표 1. 제거되지 않은 프로토콜 및 카테고리별 플로우 개수

| 네트워크 계층 | 응용 계층 |
|---|---|
| TCP | ancp, bittorrent, data, dns, ftp, http, nbss, reload-framing, rtmpt, ssh, stun, tcp, tls, vnc, x11, xmpp |
| UDP | bjnp, chargen, data, db-lsp-disc, dcp-etsi, dhcp, dhcpv6, dns, dtls, elasticsearch, enip, gquic, kip, llmnr, lsd, mdns, nbdgm, nbns, ntp, nxp_802154_sniffer, pathport, portcontrol, rtcp, sip, snmp, srvloc, ssdp, stun, teredo |

표 2. 데이터셋 내 전체 프로토콜 리스트

전체 프로토콜 리스트는 표 2과 같으며 빨간색 마킹되어 있는 프로토콜들은 수집을 위한 사전 준비 또는 수집 환경에서 발생했다고 추정되는 프로토콜들로 데이터셋에서 제외하였다. 표 1은 제거한 후 데이터셋에 남은 프로토콜들을 나타내며 전체 30만 개 가량의 플로우 중 31,300개의 플로우만을 데이터셋으로 사용한다. 표 3은 프로토콜 기준으로 전처리한 데이터셋을 3-hand-shake 보존 여부를 기준으로 한번 더 전처리한 결과를 나타낸다. 결과적으로 29,195개 플로우 데이터셋을 생성하였다.

| 카테고리 | 패킷 사이즈 총합 (MB) | | #패킷 (K) | | #플로우 (K) | |
|---|---|---|---|---|---|---|
| Chat | 19 | | 46.08 | | 2.34 | |
| | 0.71 | 0.29 | 0.66 | 0.34 | 0.13 | 0.87 |
| Email | 6 | | 26.84 | | 1.77 | |
| | 0.56 | 0.44 | 0.71 | 0.29 | 0.22 | 0.78 |
| File Transfer | 8691 | | 7524.59 | | 7.31 | |
| | 1 | 0 | 1 | 0 | 0.24 | 0.76 |
| P2P | 352 | | 421.83 | | 0.36 | |
| | 0.98 | 0.02 | 0.98 | 0.02 | 0.65 | 0.35 |
| Streaming | 2181 | | 1843.72 | | 0.77 | |
| | 1 | 0 | 1 | 0 | 0.85 | 0.15 |
| Voip | 3654 | | 8819.62 | | 16.63 | |
| | 0.18 | 0.82 | 0.08 | 0.92 | 0.1 | 0.9 |
| Total | 14903 | | 18682.69 | | 29.2 | |
| | 0.8 | 0.2 | 0.56 | 0.44 | 0.17 | 0.83 |

표 3. 데이터셋 최종 전처리 결과

## Ⅳ. 결론

본 논문은 응용 프로그램 트래픽 분류 분야에서 널리 사용되는 ISCX VPN-nonVPN 2016 공개 데이터셋에 대해 전처리 방법을 제안한다. 또한, 전처리된 데이터셋을 공개하여 이 공개 데이터셋을 활용하는 다른 논문과의 객관적인 성능 비교를 가능하게 합니다. 전처리된 데이터셋, 데이터셋 구조, 자세한 전처리 단계 및 관련 코드는 [5]에서 제공된다.

## 참 고 문 헌


[1] DRAPER-GIL, Gerard, et al. Characterization of encrypted and vpn traffic using time-related. In: Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016. p. 407-414.

[2] ACETO, Giuseppe, et al. DISTILLER: Encrypted traffic classification via multimodal multitask deep learning. Journal of Network and Computer Applications, 2021, 183: 102985.

[3] SHAPIRA, Tal; SHAVITT, Yuval. FlowPic: A generic representation for encrypted traffic classification and applications identification. IEEE Transactions on Network and Service Management, 2021, 18.2: 1218-1232.

[4] WANG, Wei, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In: 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2017. p. 43-48.

[5] Baek, U. (2023, May 24). Pb1069 / Preprocessing-of-ISCX-VPN-NonVPN-2016-. Github.Com. https://github.com/pb1069/preprocessing-of-ISCX-VPN-nonVPN-2016-