https://doi.org/10.7840/kics.2022.47.12.2096

순차적인 데이터 처리를 통한 딥 러닝 기반 트래픽 분류속도 개선

이 민 성◆, 박 지 태*, 백 의 준*, 최 정 우*, 신 창 의**, 김 명 섭°

Deep Learning-Based Traffic Classification Speed Improvement Through Sequential Data Processing

Min-Seong Lee◆, Jee-Tae Park*, Ui-Jun Baek*, Jung-woo Choi*, Chang-Yui Shin**, Myung-Sup Kim°

요 약

네트워크의 발전과 변화된 환경으로 인하여 다양한 응용 프로그램들이 개발되고 사용되고 있다. 이에 따라 네트워크 트래픽의 발생량도 증가하고 있으며 효율적인 네트워크의 관리를 위한 응용 트래픽 분류가 필요하다. 응용 트래픽 분류는 대부분 분류 정확도에 중점을 두고 있고, 실제 대용량 트래픽이 발생하는 네트워크 환경에서 트래픽 분류를 빠르게 처리하기 위한 연구가 필요하다. 본 논문에서는 순차적으로 데이터를 처리하여 딥 러닝 기반의 트래픽 분류속도를 개선하는 방법에 대하여 제안하고, 제안하는 방법에서 사용하는 임계값 및 신뢰도를 정의한다. 앙상블 모델에서의 적절한 임계값은 0.7로 99.78%의 신뢰도를 달성하고 전체 데이터의 58.99%의 데이터를 정확하게 분류하였다. 앙상블 모델에서 분류되고 남은 데이터들을 딥 러닝 모델에 적용하여 실험한 결과 제안한 방법의 전체 처리 속도는 1D CNN만을 사용한 결과보다 0.88초 빠른 처리 속도를 보여주었다.

키워드 : 트래픽 분류, 응용 트래픽, 머신 러닝, 딥 러닝, 처리 속도
Key Words : Traffic Classification, Application Traffic, Machine Learning, Deep Learning, Processing Speed

ABSTRACT

Due to the development of networks and the changed environment, various application programs are being developed and used. Accordingly, the amount of network traffic is also increasing, and application traffic classification is required for efficient network management. Application traffic classification focus on classification accuracy, and is needed to quickly process traffic classification in a network environment where large-capacity traffic. In this paper, we propose a method to improve the traffic classification speed based on deep learning by sequentially processing data, and define the threshold and reliability used in the proposed method. The appropriate threshold in the ensemble model was 0.7, which achieved a reliability of 99.78% and correctly classified 58.99% of the data. As a result of testing by applying the remaining data classified from the ensemble model to the deep learning model, the overall processing speed of the proposed method was 0.88 seconds faster than the result using only 1D CNN.

※ 이 논문은 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구(No.20008902, IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발)이고, 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과(2021RIS-004)이다.
◆ First Author : Department of Computer and Information Science, Korea University, min0764@korea.ac.kr, 학생회원
° Corresponding Author : Department of Computer and Information Science, Korea University, tmskim@korea.ac.kr, 종신회원
* Department of Computer and Information Science, Korea University, {pjj5846, pb1069, choigoya97}@korea.ac.kr, 학생회원
** Defense Agency for Technology and Quality Dajeon, Korea, {superego99}@dtaq.re.kr, 학생회원
논문번호 : 202209-210-B-RN, Received September 8, 2022; Revised October 5, 2022; Accepted October 5, 2022

Ⅰ. 서 론

코로나 19로 인한 환경 변화와 네트워크의 발전으로 네트워크의 사용량이 늘어나고 있다. 또한, 편의성을 제공하기 위하여 다양한 응용 프로그램들이 개발되고 사용되고 있다. 다양한 화상회의시스템이나 스트리밍 서비스들이 사용되고 있고, 이에 따라 네트워크 트래픽 발생량이 증가하고 있다. 이러한 대용량의 트래픽이 발생하는 상황에서 네트워크의 효율적인 운용과 관리를 위한 응용 트래픽 분류가 필수적이다.

응용 트래픽 분류는 네트워크의 효율적인 운용 및 관리, 서비스 품질 향상(QoS)[1,2], 네트워크 보안 향상 측면에서 연구가 필수적이다. 기존의 응용 트래픽 분류 기법은 포트 기반 분류[3], 페이로드 기반[4] 및 통계정보[5]를 사용한 트래픽 분류가 일반적이다. 하지만 네트워크의 보안이 중요해짐에 따라 트래픽의 암호화가 널리 사용되고 있고, 포트 기반 및 페이로드 기반의 트래픽 분류 방법들을 사용하기 어려운 문제가 있다. 암호화된 트래픽의 분류 문제를 해결하기 위하여 통계적 접근 방법을 사용하여 응용 트래픽 분류를 진행하게 되었다. 최근 연구로는 응용 트래픽 분류에 머신 러닝 및 딥 러닝 알고리즘을 사용하여 응용 트래픽 분류 분야에서 높은 분류 정확도를 보이고 있다.

머신 러닝 및 딥 러닝 알고리즘을 사용한 응용 트래픽 분류 연구가 지속해서 이루어지고 있다. 하지만 트래픽 분류 연구의 대부분이 분류 정확도에 중점을 두고 있으며, 실제 네트워크 환경에서 대용량 트래픽을 빠르게 처리할 수 있는 분류기의 분류속도와 관련된 연구는 부족하다. 응용 트래픽의 분류를 사용하는 여러 사용 환경이나 네트워크 상황은 다양하고, 각 환경에 맞는 적절한 분류기를 찾아 사용하는 것이 필요하다. 이에 따라 분류 정확도가 높지만, 처리 속도가 느린 모델을 사용할 것인지, 분류 정확도가 조금은 낮지만, 처리 속도가 빠른 모델을 사용할 것인지에 대한 선택지가 필요하다.

본 논문에서는 머신 러닝 중에서도 분류 성능이 좋은 앙상블 모델과 딥 러닝 모델을 사용하여 순차적으로 데이터를 처리하고 응용 트래픽의 분류속도를 개선하는 방법을 제안한다. 머신 러닝 모델을 사용하였을 경우 딥 러닝 모델보다 분류 정확도는 낮지만 빠른 처리 속도를 보여주는 장점이 있다. 따라서 머신 러닝 모델에서 학습된 모델을 사용하여 데이터를 분류하여 결과를 도출하고, 분류가 어려운 데이터는 딥 러닝 모델에서 분류하도록 하여 딥 러닝에서의 입력 데이터의 개수를 줄여 전체적인 처리 속도를 줄이는 것을 목표로 한다. 머신 러닝 모델에서 학습된 모델에서 임계값과 신뢰도를 새롭게 정의하여 모델에서 분류된 내용의 신뢰성을 보장할 수 있도록 한다.

1장의 서론에 이어 2장에서는 관련 연구에 대하여 설명하고, 3장에서는 제안하는 방법론에 대하여 설명한다. 4장에서는 적용한 방법론의 실험 결과를 보여주며, 마지막으로 5장에서는 결론 및 향후 연구에 대하여 설명한다.

Ⅱ. 관련 연구

본 장에서는 응용 트래픽 분류 분야에서 사용되는 연구 방법을 간략히 소개한다.

2.1 전통적인 트래픽 분류 방법

암호화된 트래픽이 나타나기 이전의 응용 트래픽 분류 분야에서는 포트 기반 분류 방법, 페이로드 기반 분류 방법 및 통계적 방법을 사용한 분류 방법을 사용하였다. 포트 기반 분류 방법의 경우 응용 트래픽과 포트 번호를 규칙에 따라 매핑하여 분류하는 방법이지만 임의의 포트를 사용하거나 암호화된 트래픽의 도입으로 사용하기 어려워졌다. 페이로드 기반 분류 방법은 분석 측면에서 가장 높은 분석 성능을 보이지만 페이로드를 직접 추출해야 하는 어려움이 있으며, 암호화된 트래픽에서는 페이로드를 정의하기 어렵다. 마지막으로 통계적 방법은 암호화 트래픽에서도 사용할 수 있는 장점이 있으나 실용성이 떨어지는 단점이 있지만 머신 러닝 및 딥 러닝 기반의 트래픽 분류 연구가 진행되면서 데이터의 입력으로 사용되고 있다.

2.2 머신 러닝 기반 트래픽 분류

트래픽 플로우의 시간적 요소를 고려하면서 응용 트래픽을 분류할 수 있는 머신 러닝 기반의 분류기를 사용한 연구들이 진행되었다. 머신 러닝 모델의 입력으로 트래픽의 지속 시간, 초당 바이트 수 등의 시간 관련 기능을 사용하였다. 각 패킷의 크기, 패킷 간 도착 시간, 패킷 길이 등의 플로우의 통계정보를 사용하거나 페이로드 정보를 사용하여 Supervised Machine Learning을 적용한 연구가 있으며 다양한 단일 및 앙상블 모델에 대하여 성능 실험을 진행한 연구들이 있다[6].

2.3 딥 러닝 기반 트래픽 분류

딥 러닝은 기계 학습에서 발전되어 컴퓨터 비전 분야에서 높은 성능을 보여주었다. 이를 계기로 컴퓨터

비전 분야뿐만 아니라 다양한 분야에서 딥 러닝을 적용하였으며, 네트워크 분야에서도 사용되고 있다. 암호화된 트래픽을 분류하기 위하여 CNN(Convolution Neural Network)을 적용하여 분류 방법을 제안하였다[7,8]. 트래픽 데이터를 학습하여 딥 러닝 모델에 적용하는 것뿐만 아니라 트래픽 특징 추출 단계에서 SAE(Stacked Auto encoder)를 사용하고 트래픽 분류 단계에서 CNN을 사용하여 연구한 논문이 있다[9].

Ⅲ. 본 론

본 장에서는 제안하는 처리 속도 개선 방법의 분류 시스템에 대하여 제안한다. 제안하는 분류 시스템은 분류 모델을 2가지를 사용한다. 제안하는 분류 시스템의 구성은 그림 1과 같다.

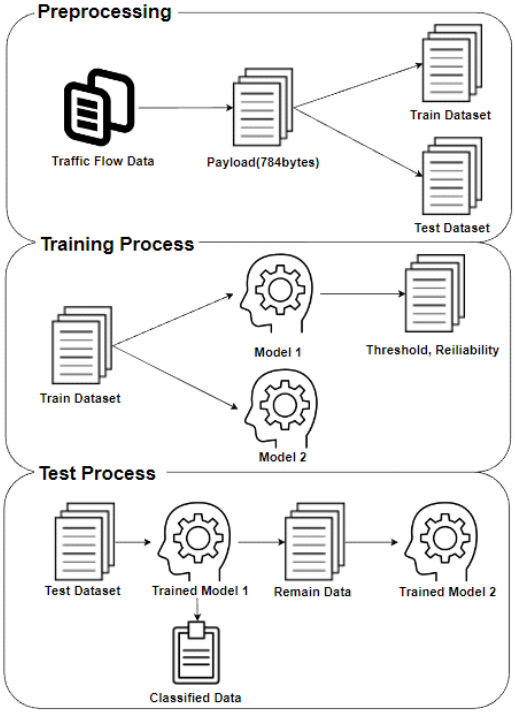


그림 1. 분류 시스템 구성도
Fig. 1. Classification System Architecture

3.1 Preprocessing

본 절에서는 분류 시스템에서의 전처리 과정에 대하여 설명한다. 실험에 사용되는 원시 데이터는 같은 IP, Port, Protocol에 대하여 플로우 단위로 정리할 수 있다. 플로우 단위에서 발생하는 패킷들의 페이로드만을 추출하여 학습 및 테스트 데이터로 사용한다. 784바이트의 페이로드를 사용하며 플로우 안에서 784바이트 이하의 페이로드가 발생하면 0으로 패딩을 하게 되고, 784바이트 이상의 페이로드가 발생하면 784바이트 이후의 페이로드는 잘라내어 사용한다[10]. 1개의 플로우 당 784바이트의 페이로드가 추출되면 학습 및 테스트 데이터로 분류하여 학습 과정 및 테스트 과정에서 사용된다.

3.2 Training Process

본 절에서는 학습 과정에 대하여 설명한다. 학습 과정에서 분류 모델 2가지를 사용하여 학습 모델을 생성해야 한다. 첫 번째로 사용되는 모델은 데이터를 분류하는 속도가 빨라야 한다. 그리고 분류가 쉬운 데이터를 먼저 처리하기 때문에 많은 수의 데이터를 정확하게 분류하기 위해 약 80% 이상의 검증 정확도가 필요하다. 첫 번째 모델에서 학습된 모델의 검증 결과를 바탕으로 임계값과 신뢰도를 정의한다. 임계값과 신뢰도는 첫 번째 분류 모델이 데이터를 먼저 분류하는 것에 대한 신뢰성을 보장하는 방법이다. 두 번째로 사용되는 분류 모델은 첫 번째 분류 모델보다 처리속도가 느리고 분류 정확도가 첫 번째 분류 정확도보다 높아야 한다. 2가지 분류 모델에 대한 조건은 표 1과 같다. 제안하는 방법을 통해 처리 속도가 느린 두 번째 분류 모델에 들어가는 입력 데이터를 줄여 전체 처리 속도를 낮추도록 한다.

표 1. 제안하는 분류 시스템의 조건
Table 1. Conditions of the Proposed Classification System

	Model 1	Model 2
Processing Speed	Model 1 < Model 2	
Accuracy	Over 80%	Model 1 < Model 2

3.2.1 임계값과 신뢰도

본 항에서는 임계값과 신뢰도에 대한 정의에 대하여 설명한다. 제안하는 분류 시스템의 학습 과정에서 첫 번째로 사용되는 분류 모델의 학습된 결과로 임계값과 신뢰도를 사용하게 된다.

임계값이란 첫 번째 분류 모델의 학습 모델 생성 시 검증 과정에서 데이터를 분류하였을 때 상세 결과에 대한 기준이다. 예를 들어, 학습 모델에서 한 개의 플로우가 클래스를 결정하게 되는 값들이 결과로 나타나게 된다. 임계값을 설정하고 임계값 이하의 값으로 플로우의 클래스가 결정되었을 때 분류된 플로우는 첫 번째 모델에서 분류가 어려운 문제로 판단하고

두 번째 모델에서 분류하게 된다. 설정된 임계값 이상으로 분류된 데이터들은 첫 번째 모델에서도 쉽게 분류할 수 있다고 판단하고 분류 결과로 도출한다. 첫 번째 모델에 데이터를 입력으로 하여 설정된 임계값 이상의 데이터들은 분류 결과로 나타내고 남은 데이터는 두 번째 모델에서 분류를 진행한다. 신뢰도는 첫 번째 분류 모델에서 임계값을 결정하고 임계값 이상으로 분류된 데이터 중에서 정답이 되는 분류 비율이다. 임계값을 작게 설정하면 데이터는 많이 처리할 수 있지만, 분류 정확도가 첫 번째 모델의 영향을 많이 받게 된다. 임계값을 높게 설정하면 처리되는 데이터의 개수는 줄어들게 되지만 전체 데이터의 분류속도에 차이가 없어진다. 따라서 학습 모델에서 임계값에 따른 신뢰도를 확인하여 적절한 임계값을 설정하여야 한다.

임계값에 따른 신뢰도의 예시로서 표 2은 학습 모델 생성 시 검증 결과를 임의로 나타낸 결과이다. 임계값을 0.7이라고 설정하였을 때 0번 플로우와 1번 플로우를 결정하는 값이 임계값보다 낮으므로 분류 결과로 나타내지 않고 두 번째 분류 모델에서 처리하도록 한다. 2번 플로우와 3번 플로우는 임계값 이상으로 분류 결과를 나타내기 때문에 분류 결과로 도출된다. 임계값이 0.7일때의 신뢰도는 총 2가지 플로우(2번, 3번 플로우) 중에서 2가지 모두 정확한 분류가 되기 때문에 100%의 신뢰도를 달성한다. 임계값이 0.5로 설정되면 1번 플로우도 분류 결과에 포함되지만, 클래스 분류에 실패했기 때문에 신뢰도는 67%가 된다.

표 2. 임계값에 따른 신뢰도 예시
Table 2. Example of Reiliability by Threshold

Data(Label)	Class1	Class2	Class3
0(Class2)	0.4	0.3	0.3
1(Class1)	0.1	0.5	0.4
2(Class3)	0.1	0.2	0.7
3(Class1)	0.8	0.5	0.15

Threshold	Classified	TP	Reliability
0.5	3	2	67%
0.7	2	2	100%

3.3 Test Process

본 절에서는 테스트 과정에 대하여 설명한다. 학습된 두 가지 모델을 기반으로 테스트 데이터를 사용한다. 학습 과정에서 첫 번째 분류 모델의 학습 모델이 생성되고 적절한 임계값이 설정되면 첫 번째 학습 모델에서 테스트 데이터의 분류를 진행한다. 분류를 진행하면서 임계값 이하의 데이터는 남기고 임계값 이상의 데이터는 분류 결과로 도출한다. 분류된 데이터를 제외하고 남은 데이터를 두 번째 모델에서 분류하고 전체 처리 시간 및 분류 정확도를 도출한다.

Ⅳ. 실 험

본 장에서는 제안한 방법에 대하여 실험한 결과에 대하여 설명한다. 실험에서는 2가지 모델에 데이터를 순차적으로 사용하여 전체 처리 속도를 개선하는 것을 목표로 한다. 두 가지 분류 모델을 선정하는 조건으로 첫 번째 모델이 두 번째 모델보다 처리 속도가 빨라야 한다. 그리고 두 번째 모델이 첫 번째 모델보다 분류 정확도가 높아야 한다. 따라서 첫 번째 모델은 앙상블 모델을 사용하고 두 번째 모델은 딥 러닝 모델을 사용하였다. 첫 번째 모델은 Random Forest(RF) 모델을 사용하였고, 두 번째 딥 러닝 모델로 1D CNN 모델을 사용하였다.

4.1 데이터셋

실험에 사용된 데이터셋은 ICSXVPN2016으로 암호화된 응용 트래픽을 모아둔 공공 데이터셋이다. 실험에서는 큰 범주에서 6가지의 응용 트래픽을 분류하였다. 원시 데이터에서 정리한 전체 플로우의 개수는 27811개이며, 한 개의 플로우 당 784바이트의 페이로드 데이터를 추출하여 학습 및 테스트 모델의 입력으로 사용하였다. 전체 플로우 데이터에서 80%인 22249개의 플로우 데이터는 학습 데이터로 사용하였으며 학습 데이터 중에서 20%인 5596개의 데이터를 검증 데이터로 사용하였다. 검증 데이터를 통해 임계값에 따른 신뢰도를 도출하고 적절한 임계값을 설정하였다. 학습 데이터를 제외한 나머지 20%의 데이터인 5563개의 플로우 데이터는 테스트 데이터로 사용

표 3. ISCXVPN2016 데이터 셋 정보
Table 3. Information of ISCXVPN2016 Dataset

	Train	Val	Test	Total
Email	1387	347	347	1734
CHAT	1950	487	488	2438
STREAM	1331	333	333	1664
FTP	5282	1321	1321	6603
P2P	209	52	52	261
VOIP	12089	3022	3022	15111
Total	22248	5596	5563	27811

하였다. ICSXVPN2016 데이터셋의 응용 트래픽의 정보와 학습 및 검증, 테스트 데이터의 개수는 표 3과 같다.

4.2 데이터 학습 및 학습 모델 생성

데이터셋에서 하나의 플로우 당 784바이트의 페이로드를 사용하여 앙상블 모델과 딥 러닝 모델의 입력으로 사용한다. 앙상블 모델로는 RF를 사용하였으며, 데이터를 분류하는 파라미터를 선정하였다. RF의 파라미터는 표 4와 같다. 딥 러닝 모델로는 1D CNN을 사용하였다. 1D CNN은 Convolution Layer 2개를 사용하여 실험에 적용하였다.

앙상블 모델 및 딥 러닝 모델을 학습하여 모델을 생성하고 모델 생성 시 사용된 검증 데이터를 통해 검증 결과를 도출하였다. 검증 결과는 표 5와 같다.

표 4. Random Forest의 파라미터 정보
Table 4. Random Forest Parameter Information

Random Forest Parameters	
n_estimators	100
max_depth	10
min_samples_leaf	2
min_samples_split	4

표 5. 분류 모델의 검증 결과
Table 5. Validation Results of Classification Model

Train Model	Validataion Accuracy
Random Forest	81.09%
1D CNN	91.23%

4.3 임계값과 신뢰도 결정

학습 데이터셋으로 앙상블 모델을 생성할 때의 검증 결과를 통해 임계값과 신뢰도를 결정할 수 있다. 임계값에 따른 신뢰도는 표 6과 같다. RF를 사용하였을 때 임계값은 학습 모델이 검증 데이터의 플로우를 분류할 때 클래스를 판단하는 수치에 대한 기준이다. RF의 결과로 모든 데이터의 클래스를 구분하는 수치가 주어졌을 때, 임계값은 각 수치의 기준을 정해준다. True Positive Threshold(TPT)는 주어진 임계값을 충족하는 데이터 중에서 실제 클래스와 일치하는 데이터이다. 신뢰도는 임계값을 충족하는 데이터들 중 실제 클래스와 일치하는 데이터의 비율이다. RF로 학습된 모델이 검증 시 나타내는 검증 정확도와는 별개로 임계값을 주기 때문에 신뢰도는 RF 모델이 전체 데이터셋 중에 확실하게 분류가 가능한 데이터의 수치를 보여 줄 수 있다. Classified Data Rate(CDR)는 테스트에 사용된 데이터 중 RF에서 확실하게 분류된 데이터의 비율이다. CDR이 높을수록 딥 러닝에서 분류해야 하는 데이터의 개수가 적어지며, 전체 처리 속도가 빨라질 수 있다. 그림 2는 임계값을 설정하였을 때 도출되는 신뢰도와 임계값 이상에서 정확하게 분류되는 데이터의 개수이다.

적절한 임계값을 설정하면 높은 신뢰도를 통해 RF 모델에서 확실한 데이터를 분류할 수 있다. 실험 결과 적절한 임계값은 0.7로서 신뢰도는 99.78%이며, 테스트 데이터의 58.99%의 데이터를 정확하게 분류할 수 있다. 임계값을 0.8로 설정하게 되면 신뢰도는 99.96%로 더 높지만 TPT의 개수가 적어져 두 번째 모델에서의 처리 속도는 증가하게 된다. 표 6은 설정된 임계값에 따라서 분류된 데이터의 개수와 전체 데이터에서 RF 모델이 데이터를 분류한 비율이다.

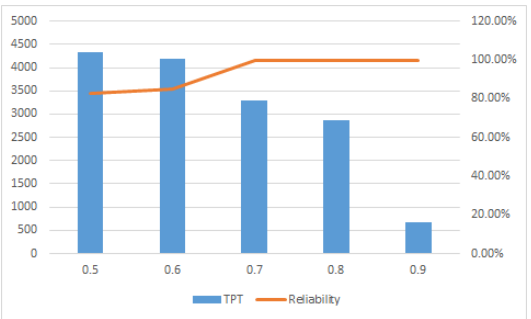


그림 2. 임계값에 대한 신뢰도 및 분류된 데이터
Fig. 2. Reliability and Classified Data for Thresholds

표 6. 임계값에 따라 분류된 데이터의 개수 및 비율
Table 6. Number and Rate of Classified Data by Threshold

Threshold	TPT	CDR
0.5	4339	77.99%
0.6	4183	75.19%
0.7	3282	58.99%
0.8	2871	51.60%
0.9	684	12.29%

4.4 실험 결과

테스트 데이터셋을 사용하여 학습된 모델을 테스트하였을 때의 결과는 표 7과 같다. 테스트 데이터셋은 5563개의 플로우 데이터의 페이로드 784바이트를 입력으로 한다. 앞서 학습된 RF 모델과 1D CNN 모델

표 7. 2가지 모델의 실험 결과
Table 7. Experiment Results of 2 Models

Test Model	Test Accuracy	Test Time(sec)
RF	81.00%	0.33
1D CNN	84.23%	2.51

을 사용하여 실험한 결과는 아래 표와 같다. RF를 사용하여 테스트 데이터셋을 처리하였을 때 분류 정확도는 81%가 도출되며, 0.33초의 시간이 걸렸다. 1D CNN 모델을 사용하였을 때 84.23%의 분류 정확도를 보이며, 2.51초의 시간이 걸렸다.

제안하는 방법을 사용하여 테스트 데이터를 실험하였을 때의 결과는 표 8과 같다. 임계값에 따라서 1D CNN 모델에서 처리하는 데이터의 양이 달라진다. 임계값이 높아질수록 분류해야 하는 데이터의 개수가 많아지게 되며, 그만큼 전체 처리 시간이 늘어나게 된다. RF 모델에서 임계값을 0.7로 설정하였을 때 가장 많은 데이터를 높은 신뢰도를 통해 분류할 수 있게 되고, 1D CNN 모델에서 처리하는 데이터의 양이 적어진다. 제안하는 방법을 사용하여 전체 분류 정확도와 전체 처리 시간을 고려하였을 때, 1D CNN 모델만 사용하여 테스트할 때보다 높은 85.88%의 분류 정확도를 보여준다. 또한, 전체 처리 시간도 0.88초 빨라지는 결과를 도출하였다.

표 8. 임계값에 따른 분류 정확도 및 처리 시간
Table 8. Classification Accuracy and Processing Time by Threshold

Threshold	Total Accuracy	Total Time(sec)
0.7	85.88%	1.63
0.8	84.82%	1.72
0.9	84.18%	2.24

V. 결 론

본 논문에서는 두 가지 모델에 데이터를 순차적으로 사용하여 응용 트래픽 분류에서의 전체 처리 속도를 개선하는 방법에 대하여 제안하였다. 앙상블 모델을 사용한 분류기로 쉽게 분류할 수 있는 데이터는 분류하고 분류하기 어려운 데이터는 딥 러닝 모델에서 분류하여 전체적인 처리 속도를 개선하였다. 이 과정에서 임계값과 신뢰도를 사용하여 적절하게 데이터를 처리할 수 있도록 하였다. 본 논문에서는 RF 앙상블 모델을 사용하여 테스트 데이터를 먼저 분류하였고, 1D CNN 모델을 사용하여 분류가 완료되고 남은 데이터를 처리하여 기존 모델을 사용한 것보다 높은 분류 정확도를 도출하였고, 처리 시간을 단축하였다. 향후 연구로는 머신 러닝 모델, 앙상블 모델, 딥 러닝 모델을 적절하게 사용하여 처리 시간을 단축하는 방법에 관하여 연구를 진행할 예정이다. 또한, 본 논문에서 사용한 테스트 데이터의 플로우 개수가 비교적 적기 때문에 실제 네트워크 환경에서 발생하는 대용량 트래픽의 데이터를 통해 전체 처리 시간을 단축하는 연구를 진행한다.

References

[1] M. S. Kim, Y. J. Won, and J. W. K. Hong, “Application-level traffic monitoring and an analysis on IP networks,” *ETRI J.*, vol. 27, pp. 22-42, 2005.

[2] B. Park, Y. Won, J. Chung, M. S. Kim, and J. W. K. Hong, “Fine-grained traffic classification based on functional separation,” *Int. J. Netw. Manag.*, vol. 23, pp. 350-381, Sep. 2013.

[3] *IANA port number list*, Available: http://www.iana.org/assignments/service-names-prt-numbers/service-names-port-numbers.xml

[4] T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, and H. Kim, et al., “Content-aware internet application traffic measurement and analysis,” *IEEE/IFIP NOMS 2004*, pp. 511-524, 2004.

[5] N. F. Huang, G. Y. Jai, H. C. Chao, Y. J. Tzang, and H. Y. Chang, “Application traffic classification at the early stage by characterizing application rounds,” *Inf. Sci.*, vol. 232, pp. 130-142, May 2013.

[6] A. A. Afuwape, Y. Xu, J. H. Anajemba, and G. Srivastava, “Performance evaluation of secured network traffic classification using a machine learning approach,” *Comput. Standards & Interfaces*, vol. 78, no. 103545, 2021, ISSN 0920-5489.

[7] Z. Zou, J. Ge, H. Zheng, Y. Wu, C. Han, and Z. Yao, “Encrypted traffic classification with a convolutional long short-term memory neural network,” *2018 IEEE 20th Int. Conf. High Performance Comput. and Commun.; IEEE 16th Int. Conf. Smart City; IEEE 4th Int.*

Conf. Data Sci. and Syst. (HPCC/SmartCity/DSS), pp. 329-334, 2018. (https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00074)

[8] Z. Han, et al., "An effective encrypted traffic classification method based on pruning convolutional neural networks for cloud platform," *CECIT*, pp. 206-211, 2021. (https://doi.org/10.1109/CECIT53797.2021.00043.

[9] M. Lotfollahi, M. J. Siavoshani, R. S. H. Zade, et al., "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Soft Comput.*, vol. 24, no. 1999-2012, 2020. (https://doi.org/10.1007/s00500-019-04030-2)

[10] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," *2017 IEEE Int. Conf. Intell. and Secur. Informatics (ISI)*, pp. 43-48, 2017. (https://doi.org/10.1109/ISI.2017.8004872)

박 지 태 (Jee-Tae Park)

2017년 : 고려대학교 컴퓨터정보학과 학사

2017년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-8515-6164]

백 의 준 (Ui-Jun Baek)

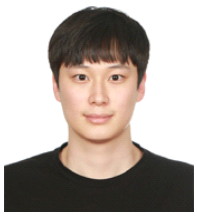
2018년 : 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-4358-7839]

이 민 성 (Min-Seong Lee)



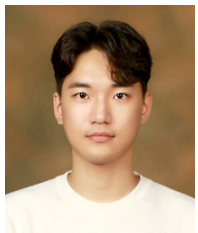
2020년 : 고려대학교 컴퓨터정보학과 학사

2020년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-1774-2831]

최 정 우 (Jung-woo Choi)



2018년 : 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-0492-8311]

신 창 의 (Chang-Yui Shin)

2003년 : 육군사관학교 운영분석학과 학사
2007년 : 고려대학교 전자컴퓨터공학과 석사
2022년~현재 : 고려대학교 컴퓨터정보학과 박사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, Ad-hoc
[ORCID:0000-0002-8410-0177]

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산 학과 학사
2000년 : 포항공과대학교 전자계산 학과 석사
2004년 : 포항공과대학교 전자계산 학과 박사
2006년 : Dept. of ECS, Univ of Toronto Canada
2006년~현재 : 고려대학교 컴퓨터정보학과 교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크
[ORCID:0000-0002-3809-2057]