https://doi.org/10.7840/kics.2019.44.6.1113

합성 곱 신경망 기반 웹 응용 트래픽 분류 모델 설계

지 세 현◆, 백 의 준*, 신 무 곤*, 채 병 민**, 문 호 원**, 김 명 섭°

Design of Web Application Traffic Classification Model Based on Convolution Neural Network

Se-Hyun Ji◆, Ui-Jun Baek*, Mu-Gon Shin*, Byeong-Min Chae**, Ho-Won Moon**, Myung-Sup Kim°

요 약

네트워크 관리의 기본 역할은 사용자에게 적합한 QoS(Quality of Service)를 제공하는 것이다. 적합한 QoS를 제공하고 안전한 네트워크 환경을 만들기 위해 정확한 응용 트래픽 분류는 필수적이다. 기존의 트래픽 분류 기법으로는 포트기반의 분류 기법, 페이로드 기반의 분류 기법, 통계정보 기반의 분류 기법이 있다. 그러나 동적인 포트 혹은 암호화된 페이로드를 갖는 패킷을 발생시키는 응용의 등장으로 인해 기존의 트래픽 분류 기법의 한계점이 발생하고 있다. 기존의 트래픽 분류 기법에 대한 한계점을 해결하기 위해 본 논문은 10종류의 웹 응용 트래픽(Baidu, Bing, Daum, Google, Kakaotalk, Nate, Naver, Yahoo, Youtube, Zum)에 대해 머신러닝 알고리즘 중 하나인 합성 곱 신경망(Convolution Neural Network) 알고리즘을 적용한 응용 트래픽 분류 모델 설계 방법을 제안한다. 제안한 모델의 학습 분류 정확도는 100%, 검증 분류 정확도는 99.6%의 성능을 달성하였다.

키워드 : 트래픽 분류, 머신러닝, 합성 곱 신경망, 응용 트래픽

Key Words : Traffic Classification, Machine Learning, Convolution Neural Network, Application Traffic

ABSTRACT

The basic role of network management is to provide quality of service suitable for users. Accurate application traffic classification is essential to provide adequate quality of service and to ensure a secure network environment. The existing traffic classification methods are port-based classification methods, payload-based classification methods and statistic information-based classification methods. However, due to the emergence of applications that generate packets with dynamic ports or encrypted payloads, the limitations of existing traffic classification techniques are occurred. In this paper, in order to address these limitations, we propose an application traffic classification model applying the convolution neural network algorithm which is one of the machine learning algorithms for 10 kinds of web application traffic(Baidu, Bing, Daum, Google, Kakaotalk, Nate, Naver, Yahoo, Youtube, Zum). The proposed model achieves 100% train classification accuracy and 99.4% validation classification accuracy.

※ 이 논문은 2016년도 국방과학연구소와 한화시스템(주)의 재원을 받아 수행된 연구임(UE161105ED)
◆ First Author : Department of Computer and Information Science, Korea University, sxzer@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, {pb1069, tm0309}@gmail.com, 학생회원
** Hanwha Systems, {byeonmin.chae, moon1000}@hanwha.com
논문번호 : 201901-421-B-RN, Received January 22, 2019; Revised March 13, 2019; Accepted April 11, 2019

Ⅰ. 서 론

오늘날의 네트워크는 고속화와 더불어 다양한 서비스와 응용이 개발됨에 따라 트래픽이 다양해지고 있으며 이러한 상황에 발맞춰 효율적으로 네트워크를 운용하기 위한 방안이 연구되고 있다. 네트워크 사용자는 고품질의 서비스를 제공받고, 네트워크 운영자는 서비스 제공의 신뢰성 확보 및 서비스의 안정적인 제공을 위해 네트워크를 관리하는 것이 필요하다[1][2]. 네트워크 관리의 기본역할은 사용자에게 적합한 QoS(Quality of Service)를 제공하는 것이다. 적합한 QoS를 제공하기 위해 정확한 응용 트래픽 분류를 하는 것은 필수적이다[3].

다양한 트래픽 분류기법들이 연구되고 있는 가운데 아직까지 트래픽을 응용 단위로 완벽하게 분류해내는 방법론은 개발되지 않았다. 응용 트래픽을 분류하기 위한 보편적인 기법으로는 표1과 같다. 포트 시그니처, 페이로드 시그니처, 통계정보 시그니처 기반 트래픽 분류 기법이 있다. 시그니처 기반 분류 기법은 특정 응용 프로그램에서 발생시킨 트래픽을 분석하여 다른 응용 프로그램과 구분 지을 수 있는 시그니처라고 하는 특정 응용만의 특징을 추출하고, 이를 통해 트래픽을 분류하는 방법이다. 포트 시그니처 기반의 분류기법은 두 개 이상의 포트 또는 임의의 포토를 설정할 수 있는 기능을 제공하는 응용 등 복잡한 구조를 갖는 응용이 등장함에 따라 정확한 분류를 어렵게 하고 있다. 페이로드 시그니처 기반의 분류기법은 분석률과 정확도 측면에서 가장 높은 분석 성능을 보이지만 수작업으로 시그니처를 추출하는 어려움이 있고, 응용 프로그램의 변화에 신속하게 대처할 수 없으며, 암호화된 페이로드를 갖는 응용에 대해서는 분류를 수행할 수 없다. 통계정보 시그니처 기반의 분류기법은 암호화된 트래픽을 분류할 수 있지만, 통계 정보를 사용할 경우 특정 응용 프로그램에서 사용한 프로토콜에 의존적인 시그니처가 생성될 가능성이 높다.

앞서 언급한 세 가지 분류방법은 모두 시그니처가 매칭 되는 트래픽만을 분류한다. 분류되지 않는 트래픽을 줄이기 위해 많은 시그니처를 사용하여야 하며, 이는 처리 시간의 증가를 야기한다[4].

표 1. 시그니처 기반 트래픽 분류
Table 1. Traffic classification based on signature

Signature	Example	Properties
Port	80:HTTP 21:FTP	고정된 포트번호 사용
Payload	"GET" "host"	페이로드 내의 특정한 패턴 사용
Statistic	Packet Size	통계 정보사용

본 논문에서는 기존의 분류기법의 한계점을 극복하기 위한 머신러닝 알고리즘 중 하나인 합성 곱 신경망(Convolution Neural Network) 기반 웹 응용 트래픽 분류 모델 설계 방법을 제안한다. 설계 방법을 통해 완성된 모델은 페이로드 시그니처 분류기법과는 대조적으로 자동으로 특징을 추출하는 것이 수월하고, 웹 응용을 분류하기 위해 학습을 통해 웹 응용의 고유성을 반영하는 고차원의 패턴을 추출한다[5].

본 논문은 10개의 웹 응용 트래픽을 대상으로 머신러닝 알고리즘 중 하나인 합성 곱 신경망 알고리즘을 적용한 모델을 설계하고, 10개의 응용에 대한 분류 실험의 결과를 통해 제안하는 방법의 적합성을 검증한다.

본 논문은 서론에 이어 2장에서는 합성 곱 신경망에 대해 기술하고, 3장에서는 제안하는 방법론에 대해 기술 한다. 4장에서는 실험을 통해 제안하는 방법론의 적합성을 검증하고, 마지막으로 5장에서는 결론 및 향후 연구를 기술한다.

Ⅱ. 합성 곱 신경망

본 장에서는 제안하는 모델의 기반인 합성 곱 신경망의 구조와, 합성 곱 신경망에서 수행하는 연산에 대하여 설명한다.

2.1 합성 곱 신경망 구조

합성 곱 신경망이란 기존의 Neural Network에 Convolution연산이 합쳐진 기술로 이미지 처리에 강력

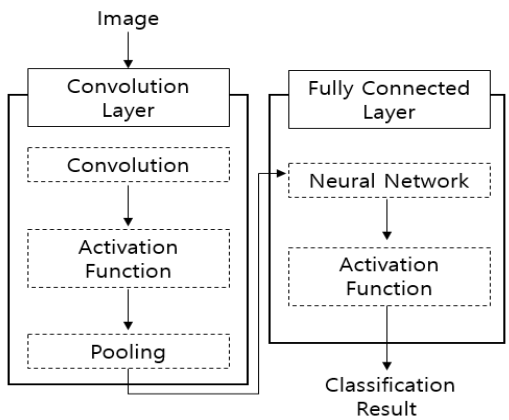


그림 1. 합성 곱 신경망 구조
Fig 1. Structure of Convolution Neural Network

한 성능을 보이는 알고리즘이다. 합성 곱 신경망의 구조는 그림 1과 같다. Convolution연산, Activation Function, Pooling연산을 통해 Feature Map을 추출하는 Convolution Layer와 Neural Network, Activation Function으로 이루어진 Fully Connected Layer로 구성된다.

2.2 Convolution연산

합성 곱 신경망에서 수행하는 첫 번째로 수행 되는 Convolution연산은 입력 데이터로부터 특징 추출을 수행한다. 그림 2와 같이 입력 받은 데이터로부터 임의의 Convolution Filter와의 Convolution연산을 통해 Convolved Map을 생성 한다.

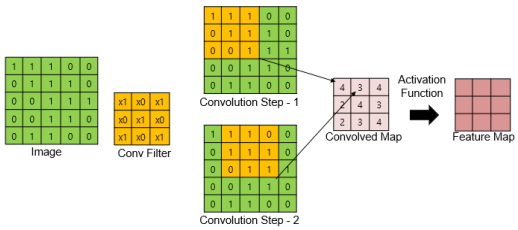


그림 2. Convolution연산과정
Fig. 2. Convolution Operation Process

2.3 Pooling연산

Pooling연산은 데이터의 크기를 인위적으로 줄이는 기법으로, Convolution연산을 통해 나온 Convolved Map에 대해 그림 3과 같이 Pooling Kernel안의 영역에서 하나의 값을 뽑아내는 연산을 수행한다.

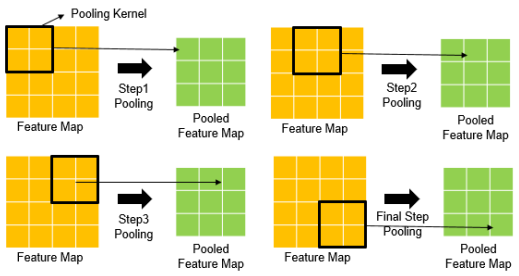


그림 3. Pooling연산과정
Fig. 3. Pooling Operation Process

2.4 Padding기법

Padding기법은 합성 곱 신경망 내의 연산을 수행하기 전, 그림 4와 같이 입력 데이터 주변을 특정 값으로 채워 늘리는 것이다. 이는 데이터의 크기가 줄어드는 것을 방지하기 위함이다.

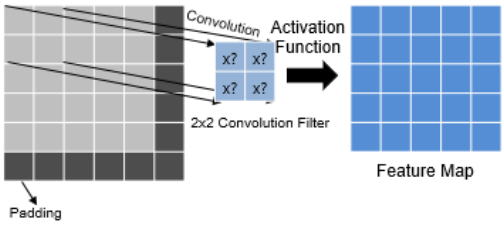


그림 4. Padding기법
Fig. 4. Padding Method

2.5 Activation Function

Activation Function은 Convolution연산을 통해 추출 된 Convolved Map의 정략적인 값을 비선형 값으로 바꿔주는 과정에서 적용되는 함수이고, 함수를 적용시켜 나온 결과는 Feature Map이다.

2.6 Fully Connected Layer

Fully Connected Layer는 그림 5와 같이 Convolution Layer의 연산을 통해 나온 Feature Map의 값에 대해 N개의 모든 Neural Network와의 연산을 한다. Neural Network와의 연산을 통해 나온 값에 대해 Activation Function을 적용한 최종 값을 통해 최종적인 분류를 수행한다.

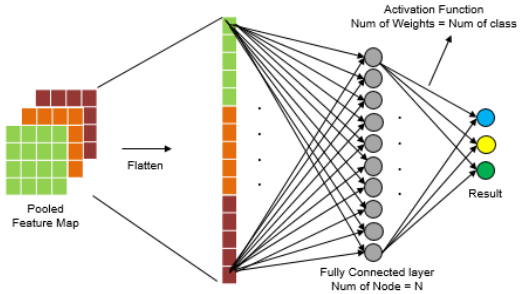


그림 5. Fully Connected Layer연산과정
Fig. 5. Fully Connected Layer Operation Process

Ⅲ. 웹 응용 트래픽 분류 모델 설계

본 장에서는 합성 곱 신경망 알고리즘을 적용한 웹 응용 트래픽 분류 모델 설계 방법에 대하여 설명한다.

3.1 Design Overview

본 논문에서 제안하는 웹 응용 트래픽 분류모델의 설계과정은 그림 6과 같다. Pre-Processing, Model Training, Model Validation과정을 통해 분류 모델을 완성한다.

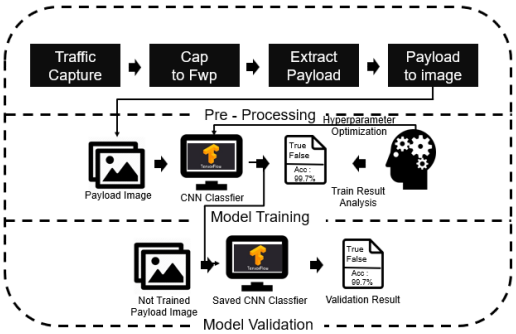


그림 6. 웹 응용 트래픽 분류 모델 설계
Fig. 6. Design of Web Application Traffic Classification

3.2 Pre-Processing

본 절에서는 합성 곱 신경망의 학습데이터를 생성하기 위한 Pre-Processing과정에 대해 단계별로 기술한다.

3.2.1 Traffic Capture

Microsoft에서 제공하는 MS Network Monitor프로그램의 Traffic Capture기능을 통해 웹 응용 트래픽 수집을 한다. 그림 7과 같이 정교한 트래픽 수집을 위해 대상 웹 서버의 IP정보를 기반으로 패킷들을 필터링하여 수집한다. 수집 된 트래픽은 패킷 단위로 구성된 cap파일의 형태로 저장이 된다.

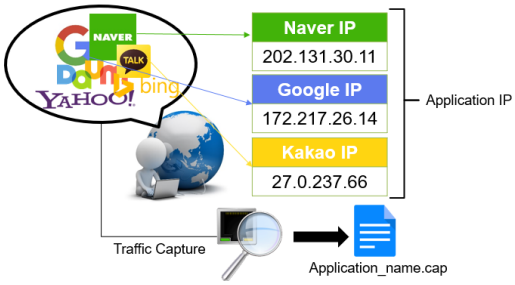


그림 7. 트래픽 수집
Fig. 7. Traffic Capture

3.2.2 Cap to Flow with Packet

Traffic Capture과정을 통해 패킷 단위로 구성된 cap 파일을 Flow with Packet형태로 구성하기 위한 단계이다. 그림 8에서의 Flow with Packet은 Client PC와 Application Server간의 하나의 세션에서 발생하는 패킷들의 집합이다.

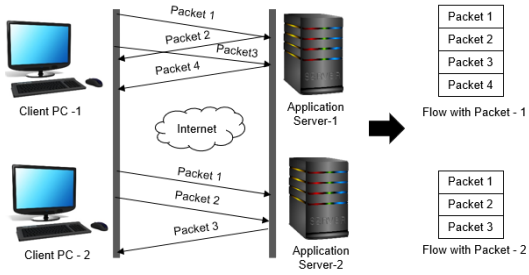


그림 8. Flow with Packet 개념
Fig. 8. Concept of Flow with Packet

3.2.3 Extract Payload

Flow with Packet의 구조에서 헤더를 제외한 데이터 부분만 추출하기 위한 단계이다. Flow with Packet의 페이로드의 형태는 그림 9와 같다. 각각의 Flow with Packet의 페이로드의 크기는 일정하지 않기 때문에 본 연구에서는 합성 곱 신경망의 대표적인 모델 중 하나인 Mnist 숫자 분류 모델의 입력 이미지의 크기와 일치시키기 위해 페이로드의 784bytes의 크기만 추출을 하고[6], 페이로드의 크기가 784bytes보다 작은 경우 다시 처음의 페이로드 값부터 채워나가는 방법을 취하였다.

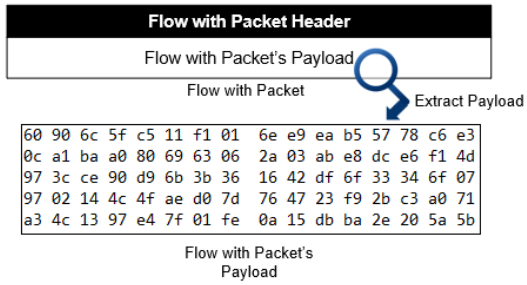


그림 9. Flow with Packet의 Payload
Fig. 9. Payload in Flow with Packet's Format

3.2.4 Payload to Image

Extract Payload과정을 통해 추출된 페이로드의 각 byte는 0~255사이의 값을 갖는다. 0~255 사이의 값을 흑백의 음영 이미지를 만드는 단계이다.

Payload to Image과정을 통해 그림 10과 같이 합성 곱 신경망에 쓰이는 학습 데이터가 완성이 된다. 페이로드의 값을 이미지로 표현한 학습 데이터는 고차원의 패턴을 추출하기 위한 이미지이므로, 단순하게 육안으로 식별하기에는 어려움이 있다.

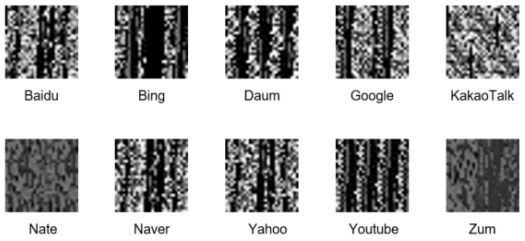


그림 10. 웹 응용 트래픽의 Payload 이미지
Fig. 10. Web Application Traffic's Payload Image

3.3 Model Training

본 절에서는 합성 곱 신경망 모델의 학습과정에 대해 기술한다. 합성 곱 신경망은 Pre-Processing과정을 통해 완성된 웹 응용 트래픽의 페이로드이미지를 입력받아 학습한다.

3.3.1 Hyper-parameter Properties

Hyper-parameter는 학습 모델을 구성하는데 임의로 설정할 수 있는 모든 요소이다. 본 연구에서 합성 곱 신경망의 학습에 쓰이는 Hyper-parameter는 그림 11과 같으며 이는 Convolution Filter의 크기, Activation Function의 선정, Pooling연산 선정, Padding기법 선정, Neural Network의 수이다.

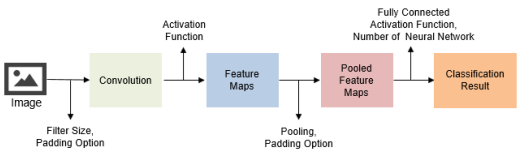


그림 11. 합성 곱 신경망의 Hyper-parameter
Fig. 11. Hyper-parameter for Convolution Neural Network

3.3.2 Hyper-parameter Optimization

Hyper-parameter Optimization은 합성 곱 신경망을 구성하는 Hyper-parameter를 임의로 선정 한 후, 각 Hyper-parameter의 모든 조합을 구성하여 모델들을 구성, 각 모델의 학습 결과로 나온 분류 정확도를 바탕으로 최적의 Hyper-parameter조합을 찾는 과정이다.

본 연구에서 쓰이는 Hyper-parameter구성은 표 2와 같이 구성하였으며. 모든 경우의 Hyper-parameter조합 중에서 학습 분류 정확도를 바탕으로 가장 성능이 좋은 1개의 모델을 선정하였다.

표 2. Hyper-parameter 구성
Table 2. Hyper-parameter Set

Activation Function	Sigmoid, Tanh, ReLU, Leaky_ReLU, Softmax
Convolution Filter Size	3x3, 5x5, 7x7
Pooling Methods	Max Pooling, Average Pooling
Padding Options	Zero Padding, No Padding
Number of Neural Network	256, 512, 1024

3.4 Model Validation

Model Training과정으로부터 완성된 최종 모델에 대해 학습에 쓰이지 않은 데이터를 통해 나온 분류 정확도를 바탕으로 검증을 하는 단계이다.

Ⅳ. 실 험

본 장에서는 제안하는 방법론을 적용하기 위해 Google에서 제공하는 Tensorflow를 이용하여 모델을 구성한다. 구성된 합성 곱 신경망 기반 웹 응용 트래픽 분류모델에 대해 10개의 웹 응용 트래픽을 대상으로 분류 실험을 진행하여 방법론의 적합성을 검증한다.

4.1 Data Set

MS Network Monitor를 통해 수집한 10개의 웹 응용 트래픽의 정보는 표 3과 같다. 443번 포트와 80번 포트를 사용하는 웹 응용 트래픽을 대상으로 수집을 하였고 수집된 실험 트래픽의 443번 포트와 80번 포트

표 3. 웹 응용 트래픽 데이터
Table 3. Web Application Data Set

Web Application Traffic	443 port (%)	80 port (%)	Number of Train Set	Number of Validation Set
Baidu	61.1	38.9	1000	300
Bing	27.8	72.2		
Daum	68.5	31.5		
Google	92.1	7.9		
Kakaotalk	62.8	37.2		
Nate	27.2	72.8		
Naver	93.4	6.6		
Yahoo	99.7	0.3		
Youtube	97.8	2.2		
Zum	14.9	85.1		

표 4. 합성 곱 신경망 모델 학습 실험 결과
Table 4. Convolution Neural Network Model Train Experiment Result

Model No.	Convolution Filter Size & Number	Convolution Activation Function	Pooling Option	Padding Option	Neural Network Activation Function	Number of Neural Network	Train Accuracy
1	5x5x32	Leaky_ReLU	No Padding	Max Pooling	Leaky_ReLU	1024	99.54%
2	5x5x32	Leaky_ReLU	Padding	Max Pooling	Leaky_ReLU	1024	99.72%
3	7x7x32	Leaky_ReLU	Only Pooling	Max Pooling	Leaky_ReLU	1024	99.78%
4	7x7x32	ReLU	Padding	Max Pooling	Leaky_ReLU	1024	99.82%
5	7x7x32	ReLU	Only Conv.	Max Pooling	Leaky_ReLU	1024	100%

에 대한 비율을 나타내었다. 각 웹 응용별 1300개의 페이로드이미지 중 1000개의 이미지를 모델의 학습에 사용하였고, 300개의 페이로드이미지를 검증에 사용하였다.

4.2 학습 실험

Hyper-parameter Optimization과정을 거친 후 나온 1440가지의 조합 중에서 높은 분류 정확도를 나타낸 상위 5개의 합성 곱 신경망 분류모델의 학습 결과는 표 4와 같다. Filter의 크기는 5x5크기의 필터보다 7x7 크기의 필터가 높은 성능을 나타내었고, 필터의 개수는 모든 Model에 동일하게 32개씩 사용하였다. Pooling Option에서 Padding의 경우 데이터의 주변을 0으로 채워놓는 Zero-Padding기법을 적용하였고, Only Pooling은 Convolution 과정 혹은 Pooling 과정에만 Padding을 적용한 것이다. Pooling기법의 경우 최댓값을 추출하는 Max Pooling기법이 우수한 성능을 나타내었다. Convolution Activation Function은 ReLU와 Leaky_ReLU가 좋은 성능을 보였고, Neural Network Activation Function의 경우 Leaky_ReLU가 최적의 함수이다. 마지막으로 Neural Network의 수가 많은 모델이 더 좋은 성능을 나타내었고, 그 중에서도 Model 5의 Hyper-parameter조합이 100%의 학습 분류 정확도를 나타내었다.

4.3 모델 검증 실험

학습 과정을 거친 상위 5개의 모델의 성능을 검증하기 위한 실험이다. 학습에 사용되지 않은 각 웹 응용별 300개의 이미지를 학습을 마친 모델의 검증 데이터로 사용한 결과는 표 5와 같다. 학습 결과가 좋은 모델일수록 좋은 검증 분류 정확도를 나타내었으며, Model 5가 99.57%로 가장 높은 정확도를 나타내었다.

표 5. 합성 곱 신경망 모델 검증 실험 결과
Table 5. Experimental Results for Convolution Neural Network Model Validation

Model No.	Validation Accuracy
1	96.76%
2	97.23%
3	97.44%
4	98.85%
5	99.57%

V. 결론 및 향후연구

본 논문에서는 기존의 트래픽 분류기법의 한계점을 극복하고자 Pre-Processing, Model Training, Model Validation과정을 통한 합성 곱 신경망 기반 응용 트래픽 분류 모델 설계방법을 제안하였다.

기존의 분류 기법 중 가장 높은 분석률과 정확도를 나타내는 페이로드 시그니처 기반 분류 기법은 443번 포트를 사용하는 암호화된 페이로드를 갖는 응용에 대해서 시그니처 추출의 어려움이 있어 분류를 할 수 없으나, 본 논문에서는 443번 포트를 사용하는 10개의 웹 응용에 대해 육안으로는 구분하기 힘든 페이로드 이미지를 학습시켜 분류하는 실험을 통해 기존 페이로드 시그니처 분류 기법의 한계점을 극복하였다.

본 연구에서 제시한 Hyper-parameter의 모든 조합에 대해학습 실험을 진행하였고, 검증 실험 결과를 통해 Model 5의 Hyper-parameter조합을 통해 합성 곱 신경망을 설계하는 것이 10개의 웹 응용 트래픽을 분류하는 데 있어 적절하다는 것을 검증하였다.

향후 연구로는 더 많은 종류의 웹 응용 트래픽을 대상으로 분류하는 실험을 진행하여, 그에 맞는 모델을 설계하는 것에 대해 연구할 계획이다.

References

[1] J. Park, S. Yoon, J. Park, S. Lee, and M. Kim, “Statistic signature based application traffic classification,” *J. KICS*, vol. 34, no. 11, pp. 1234-1244, Nov. 2009.

[2] K.-S. Shim, Y.-H. Goo, J. T. Park, and M.-S. Kim, “A study on the automatic payload signature generation based on clustering for service-specific traffic classification,” in *Proc. Symp. KICS*, pp. 1284-1285, Jun. 2017.

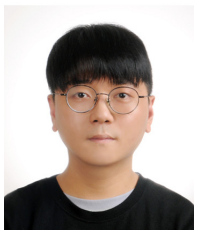
[3] Y.-H. Goo, K.-S. Shim, S. Lee, B. D. Sija, and M.-S. Kim, “A traffic-classification method using the correlation of the network flow,” *J. KIISE*, vol. 44, no. 4, pp. 433-438, Apr. 2017.

[4] S.-H. Ji, J.-T. Park, E.-J. Baek, and M.-S. Kim, “Malicious traffic detection based on convolution neural network, for secure network construction,” in *Proc. Symp. KICS*, pp. 861-862, Jun. 2018.

[5] J.-B. Kim, H.-K. Lim, J.-S. Heo, and Y.-H. Han, “Packet payload-based network traffic classification using convolutional neural network,” *KIPS 2018 Spring Conf.*, May 2018.

[6] W. Wang, X. Zeng, X. Ye, Y. Sheng, and M. Zhu, “Malware traffic classification using convolutional neural networks for representation learning,” in 31st ICOIN, Accepted, 2017.

지 세 현 (Se-Hyun JI)



2018년 : 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-2406-6958]

백 의 준 (Ui-Jun Baek)

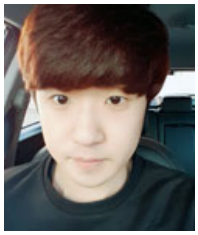
2018년 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-4358-7839]

신 무 곤 (Mu-Gon Shin)



2019년 : 고려대학교 컴퓨터정보학과 학사

2019년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0003-3703-3319]

채 병 민 (Byeong-Min Chae)

2007년 : 충남대학교 물리학과 학사

2012년 : 충남대학교 컴퓨터공학과 석사

2007년~2008년 : 삼성전자 연구원

2008년~현재 : 한화시스템 연구원

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0003-3299-6693]

문 호 원 (Ho-Won Moon)

2000년 : 한양대학교 수학과 학사

2000년~2011년 : 삼성전자 연구원

2011년~현재 : 한화시스템 연구원

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 라우팅

[ORCID:0000-0002-5668-7265]

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산학과 학사
2000년 : 포항공과대학교 전자계산학과 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]