경량 다중 모형 CNN 기반 응용 트래픽 분류

백의준, 김보선, 박재원, 최정우, 김명섭
고려대학교
{pb1069, boseon12, 2018270614, choigoya97, tmskim}@korea.ac.kr

Lightweight Multi-Shape CNN based Application Traffic Classification

Ui-jun Baek, Boseon Kim, Jae-Won Park, Jeong-Woo Choi, Myung-Sup Kim
Korea University

요 약

인터넷 환경의 규모가 급성장하고 응용 트래픽의 다양화로 인해 네트워크 트래픽 분류의 중요성이 나날이 증가하고 있다. 딥러닝 기술의 발전으로 페이로드 기반 분류 방식, 포트 기반 분류 방식 등의 기존 방법론 대비 딥러닝 기반 분류 방법이 높은 성능 및 발전 가능성을 보인다. 특히, 컴퓨터 비전 분야에서 많이 활용하는 CNN 기반의 트래픽 분류 방법들이 많이 연구되고 있으며 트래픽의 입력 형태를 고려하여 고정 길이의 입력 패킷을 다양한 모형으로 변환한 CNN 기반 분류 모델은 그중에서도 높은 분류 정확도를 보인다. 하지만, 하나의 입력을 여러 개의 이미지로 변환하여 특징을 추출하는 과정은 높은 연산량으로 인한 처리 속도 문제를 수반한다. 따라서, 우리는 기존의 모델 구조를 기반으로 정확도를 유지하면서도 처리 속도를 개선한 경량 다중 모형 CNN 기반 응용 트래픽 분류 방법을 제안한다. 우리가 제안하는 모델은 정확도의 하락 없이 기존 모델보다 2.2배 높은 처리 속도를 보였다.

Ⅰ. 서 론

인터넷 환경의 규모가 급성장하고 점점 더 다양한 응용 트래픽이 발생함에 따라 네트워크 트래픽 분류의 중요성이 나날이 증가하고 있다. 응용 트래픽 분류는 응용 타입 식별, 비정상 행위 탐지, QoS(Quality of Service) 보장 등을 목적으로 한다. 목적에 따라 응용 트래픽 분류는 높은 정확도로 수행되어야 하나 인터넷 환경 및 응용의 다양화, 트래픽 암호화 기술의 적용 등은 기존 응용 트래픽 분류 기술의 한계점을 드러내었다. 이를 위해 많은 딥러닝 기반 응용 트래픽 분류 방법들이 연구되고 있으며 기존 방법 대비 매우 높은 분류 정확도를 나타내고 있다. 그러나, 딥러닝 기반 방법은 기존 방법 대비 많은 계산량을 요구한다는 단점을 가진다. 따라서, 딥러닝 기반 모델의 정확도를 유지하면서도 계산량을 낮추는 방법이 필요하다. 따라서, 우리는 기존 연구 중 높은 정확도를 보였던 다중 모형 CNN 기반 응용 트래픽 분류 모델의 경량화 방법을 제안한다.

본 논문은 1장 서론에 이어 2장에서 관련 연구를 제시하고 3장에서 딥러닝 모델의 경량화 방법을 설명한다. 4장에서는 실험 결과를 설명하고 5장에서 결론 및 향후 연구를 제시하는 것으로 본 논문을 마친다.

Ⅱ. 관련 연구

본 장에서는 딥러닝 기반 모델 중 다중 모형 CNN 기반 응용 트래픽 분류 방법에 관해 설명한다. [1]은 1차원 데이터인 네트워크 패킷은 기존 CNN에서 활용하던 정사각형 모양의 2차원 데이터와는 구조가 다르며 네트워크 패킷의 일부 필드 값은 1차원 연속성을 지니고 있어 패킷 데이터가 2차원으로 변환하는 과정에서 필드 값이 손실되는 점을 근거로 정사각형 기반 2D-CNN, 1D-CNN뿐만 아니라 다양한 모형을 통해 특징을 추출하여 손실을 완화하는 방법을 제안하였다. 본 연구에서는 해당 모델을 baseline으로 사용한다.

Baseline의 구조는 그림 1와 같으며 다수 개의 패킷으로 이루어진 양방향 플로우 데이터가 모델에 입력된다. 입력 플로우에 포함된 각 패킷들은 컨볼루션 연산을 1회 거친 후 Reshape 레이어를 통해 다수 개의 모형으로 변환된다. 각기 다른 모형들은 하나의 잔차 블록[2]을 통과하여 특징이 추출되며 이는 Concatenate 레이어를 통해 패킷 하나를 대표하는 특징으로 합쳐진다. 각 패킷에서 추출된 특징들은 다시 한번 병합되고 GRU 레이어를 통해 시계열 특징을 추출하게 되며 마지막 Softmax 연산을 통해 각 카테고리에 해당하는 확률을 출력한다.

Ⅲ. 다중 모형 CNN 기반 분류 모델의 경량화 방법

우리는 Baseline의 경량화를 위하여 두 가지 방법을 제안한다. 첫 번째 방법은 변환된 다수 개의 모형 중 일부만 채택하는 것이다. 예를 들어, 입력 패킷의 크기가 484 bytes이면 8개의 다른 모형을 생성할 수 있으며 모든 모형이 아닌 8개 중 일부 모형만 채택하는 방법이다. 두 번째는 모델을 무겁게하는 특정 FC(Fully-connected) 레이어를 GAP(Global Average Pooling)또는 GMP(Global Max Pooling) 레이어로 교체하는 것이다. GMP의 경우 추출한 특징으로부터 하나의 뚜렷한 특징을 찾아내며 GAP의 경우 전체적으로 뚜렷한 특징이 있는지를 찾는 것으로 알려져 있으며 과적합을 방지하는데 도움이 된다.

우리는 제안한 방법의 평가를 위하여 공용 데이터셋인 ISCX VPN-nonVPN 2016를 사용하였으며 입력 플로우를 6개의 카테고리로 분류하는 데이터 세트다. 간략한 데이터 세트 개요는 표1에 나타나 있다. 실험에 사용한 baseline의 하이퍼파라미터는 [1]과 똑같이 설정하였으며 생성되는 입력 모형은 표 2에 나타내었다. 실험 데이터 세트는 8:2의 비율로 학습 데이터와 테스트 데이터로 분할하였다.

1) 이 연구는 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구(No. 20008902, IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발)이며 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2021RIS-004)

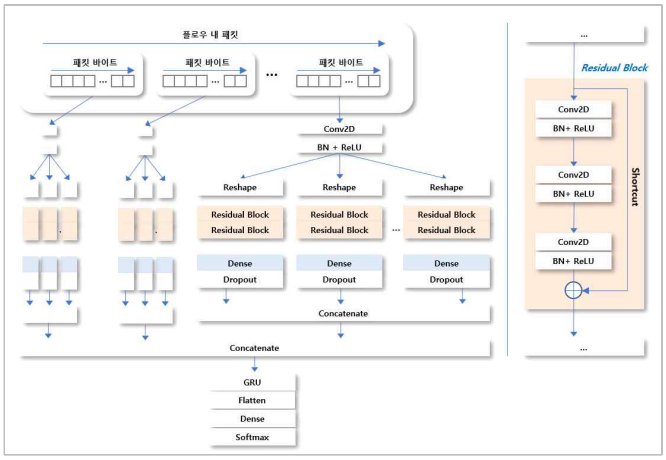

그림 1 Baseline 구조

표 1 데이터 세트 개요

카테고리	개수	비율(%)
Chat	2,438	8.8
Email	1,734	6.2
Streaming	1,664	6.0
File transfer	6,603	23.7
VoIP	15,111	54.3
P2P	261	0.9
Total	27,811	100.0

표 2 생성되는 입력 모형 개요

인덱스	모형	커널 크기	필터 개수
1	(1, 484)	(1, 60)	
2	(2, 242)	(1, 30)	초깃값 :4
3	(4, 121)	(1, 15)	마지막 값: 16
4	(11, 44)	(1, 5)	
5	(22, 22)	(2, 2)	
6	(44, 11)	(5, 1)	잔차블록을
7	(121, 4)	(15, 1)	거칠 때마다 x2
8	(2, 242)	(30, 1)	

Ⅳ. 실험 결과

표 3 입력 모형 개수에 따른 정확도 및 추론속도

Model	입력 모형 인덱스								정확도	추론속도
	1	2	3	4	5	6	7	8	(%)	(flow/s)
MISCNN[1]	√	√	√	√	√	√	√	√	92.0	95
1D CNN	√								91.2	264
2D CNN					√				87.2	277
Partial	√				√				90.02	206
Partial			√					√	90.49	213
Partial		√				√	√		90.34	173
Partial				√		√		√	89.97	179

표 3은 입력 모형 개수에 따른 정확도와 추론속도를 나타낸다. 분할된 모형을 모두 사용하는 기존 방법의 경우 1D CNN, 2D CNN보다 높은 성능을 나타내었지만 많은 모형으로 인해 가장 느린 추론속도를 나타내었다. 1D CNN은 두 번째로 좋은 성능을 보여주었으며 추론속도 또한 2D CNN을 제외하고 가장 빨랐다. 일부 모형만을 사용한 경우는 2D CNN 보다는 높은 정확도를 보여주었지만 1D CNN과 비교하였을 때 낮은 정확도와 추론속도를 보여주었다.

표 4 GAP와 GMP 적용 비교결과

MODEL	GMP	GAP	정확도 (%)	추론속도 (flow/s)
MISCNN			92.0	95
1D CNN			91.2	264
2D CNN			87.2	277
Partial			90.5	213
Partial		√	91.7	219
Partial	√		91.6	219
Partial	√	√	93.7	213

표 4는 Partial 모델 중 모형 3과 모형 8을 선택한 모델을 기반으로 GAP와 GMP를 적용하고 비교 결과를 나타낸다. 실험 결과, Partial 모델을 기반으로 GMP와 GAP를 모두 적용한 모델의 경우 MISCNN 및 1D CNN 모델 보다 높은 정확도를 보여주었다. 추론속도 측면에서는 1D CNN보다 다소 낮은 속도를 보여주나 baseline인 MISCNN 대비 2.2배 가량 빠른 속도를 보여주었다.

Ⅴ. 결론

본 논문은 높은 정확도를 보이지만 너무 많은 연산량을 인해 느린 추론속도를 가지는 기존 모델의 한계점을 해결할 수 있는 두 가지 방법을 제안하였다. GMP 또는 GAP 적용 없이 변환된 모형 중 일부만 사용한 방법은 추론속도를 향상하였으나 정확도 하락이 발생하였다. 우리는 이 문제를 해결하기 위하여 GAP와 GMP 레이어를 적용하여 하락한 정확도를 기존 baseline 이상으로 끌어올렸다.

우리는 향후 연구로 MISCNN에서 사용되었던 백본 네트워크보다 높은 성능을 보일 수 있는 백본 네트워크 또는 딥러닝 기법을 적용하여 분류 성능 고도화를 할 예정이다.

참 고 문 헌

[1] BAEK, Ui-Jun, et al. MISCNN: A Novel Learning Scheme for CNN-Based Network Traffic Classification. In: 2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2022. p. 01-06.

[2] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.