딥러닝 기반 응용 트래픽 분류에서 데이터셋과 파라미터의 수량적 관계성에 대한 연구

신창의*, 최정우**, 백의준**, 박지태**, 김명섭**

국방기술품질원*, 고려대학교**

*superego99@dtaq.re.kr, **{choigoya97, pb1069, pjj5846, tmskim}@korea.ac.kr

A study on the quantitative relationship between datasets and parameters in deep learning-based application traffic classification

Chang-Yui Shin*, Jeong-Woo Choi**, Ui-jun Baek**, Jee-Tae Park**, Myung-Sup Kim**

*Defense Agency for Technology and Quality, **Korea University

요 약

현대 문명의 흐름인 인터넷을 통해 전송되는 데이터의 엄청난 증가와 더불어 여러 가지 보안적 이슈로 인해 트래픽 암호화가 메인흐름이 되었다. 암호화된 트래픽은 동전의 앞뒷면과 같이 개별정보의 보안성을 강화시키기도 하지만 와 실용적인 트래픽분류를 방해하는 효과를 동시에 가져왔다. 이에 따라 유사영역인 트래픽 탐지와 분류는 급속한 관심과 성장을 이루고 있는데, 이전에는 머신러닝 기법을 통해 트래픽 탐지와 분류가 성장했다면 현재는 딥러닝 기법을 활용하는 쪽으로 관심과 성장이 전환되고 있다. 본 논문에서는 딥러닝을 기반으로 하는 트래픽 분류에서 분류의 성능을 높이기 위해 엔지니어 관점에서 추가하는 통계적 정보의 선택이 파라미터의 증가에 영향을 미치는데, 이때 데이터셋이 한정적일때 모델의 적합도에 어떠한 영향을 미치는가에 대해 유추 가능한 이론적 배경을 바탕으로 실험해보았다. 이를 통해 데이터셋의 수 한정적일 때 이를 고려하여, 파라미터가 구성되도록 것이 분류모델의 적합성에 영향을 미치는 요소가 될 수 있음을 확인했다.

Ⅰ. 서 론

IoT(Internet of Things), 즉 인터넷에 연결되어 애플리케이션이나 네트워크에 연결된 장치, 또는 산업 장비 등의 다른 사물들과 데이터를 공유할 수 있는 수 많은 사물들을 통한 헬스케어, 교통, 에너지 등 다양한 서비스 분야의 급격한 성장이 지속되고 있다. 이런 가운데 컴퓨터 네트워크의 QOS, 지능형 네트워크 운영·유지관리 및 네트워크 보안 등 다양한 분야에서 네트워크 트래픽 분류가 활용되고 있다. 최근 네트워크 기술의 발전과 더불어 여러 가지 보안취약성을 뒷받침하고자 암호화 트래픽을 사용하는 추세이다. 암호화된 트래픽으로 인해 개별 정보의 보안성은 강화된 반면 다양한 응용서비스의 실용성을 강화시키는 트래픽분류를 방해하는 효과를 동시에 가져와, 암호화된 트래픽 분류 및 탐지는 중요성이 보다 커지고 있다.

이미지 처리에서 급격히 성장한 딥러닝 기법은 다른 컴퓨팅영역에서의 확장으로 이어지고 있다. AI분야의 진화에 발맞춰 네트워크 트래픽 분류에서도 머신러닝 기법에서 딥러닝 기법으로 관심의 전환이라는 큰 흐름이 지속되고 있다. 머신러닝에 비교했을 때 딥러닝이 보다 높은 성능을 보이기 위해서는 대량의 데이터를 기반으로하는 학습을 통해 가능한데, 실제 환경에서 악성트래픽의 경우 대량의 데이터를 얻기도 어렵고, 대량의 데이터인 경우 연산에 많은 시간과 자원이 소비되는 제한사항이 있다.

본 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.

이에 본 논문에서는 딥러닝기반 응용트래픽 분류에서 데이터셋이 한정적일 때, 분류모델 연산의 중요요소인 파라미터의 증가가 모델의 적합성에 미치는 영향에 대해 접근하였으며, 실험결과를 통해 파라미터와 데이터셋의 수를 연계한 가운데 모델을 설계하고 학습시켜야 함을 확인하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 관련된 주요기술에 대해 제시하고, 3장 본론에서는 이론적 배경, 데이터셋의 구성, 시험환경 등에 대해 설명하고, 4장에서는 이에 대한 실험을 수행한 결과와 분석을 기술하고, 5장에서는 결론과 향후 연구방향을 제시하고자 한다.

Ⅱ. 관련연구

심층 신경망(Deep Neural Networks)은 기존의 머신러닝 모델이 가져올 수 있는 성능을 능가함에 따라 가장 인기있는 방법론 중 하나로 대두되어 지금까지도 급속히 발전되어왔으며 앞으로도 지속될 것으로 전망된다. 심층 신경망 중에서도 가장 널리 사용되는 유형은 LeCun et al.[1][2]에서 역사가 시작되는 CNN(Convolutional Neural Networks)이다. CNN은 특징을 추출하는 Convolutional layer와 데이터크기를 줄이는 Pooling layer로 구성된다.

심층신경망의 정확도를 높이기 위한 다른 추가적인 기법들도 같이 사용되고 있는데, Batch normalization과 Dropout layer가 공통적이다. Ioffe and Szegedy.[3]가 제시하는 Batch normalization은 훈련하는 동안 모든 피처에 대해 배치레벨에서의 정규화를 하고 이어서 전체 훈련데이터 세트레벨에서 크기 재조정을 한다. 이로써 훈련을 통해 얻어진 새로운 평균과 분산은 배치수준에서 습득한 평균과 분산을 대체하게 되는데, 훈련 수렴을 보다 빠르게 이끌어 성능결과를 개선하는데 기여한다.

Srivastava et al.[4]가 제시하는 Dropout layer는 이전 레이어의 출력비율을 0으로 설정하여 보이지 않는 데이터에 대해 일반화시킨다. 이 정규화 기능을 신경망이 특정 입력에 과도하게 의존하지 않게하여 과적합문제를 개선하는데 기여한다.

CNN과 더불어 최근 각광 받고있는 알고리즘이 RNN(Recurrent Neural Networks)인데, RNN은 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루며, 시퀀스 길이에 관계없이 인풋과 아웃풋을 받아들일 수 있는 게 장점이어서 유연하게 구조를 가질 수 있다. 하지만 RNN도 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하되는 단점이 있어 이를 개선하여 고안된 것이 LSTM(Long Short Term Memory)과 GRU(Gated Recurrent Unit)[5]이다. 이들은 RNN의 특성을 가지면서 동시에 역전파시 그래디언트가 잘 전파되도록 기능을 개선한 것이다.

최근에는 자연어처리 분야에서 RNN에 기반한 seq2seq 모델이 고정된 크의 벡터에 모든 정보를 압축함에 따른 정보손실과 고질적인 문제인 기울기 소실 문제를 발생시키는 문제에 대한 대안으로, Attention 기법이 등장하면서 텍스트처리 관점에서의 딥러닝 기법의 발전은 지속되고 있다.

Ⅲ. 본론

일반적으로 딥러닝 기법은 머신러닝 기법보다는 학습에 사용되는 파라미터(parameter) 개수가 현저히 많다. 머신러닝 기법은 엔지니어레벨에서 여러 가지 피처(feature)를 선정해 학습에 사용하지만, 딥러닝에서는 피처를 알고리즘 자체에서 추출한다고 이해하면 된다.

응용 트래픽 분류에 있어서 플로우간 관계성을 고려한 모델의 성능이 높다는 것[6]을 기반으로, 패킷 전체의 플로우 통계정보를 기본으로 추가적인 플로우 통계정보(첫번째 패킷으로부터 N번째 패킷까지의 통계정보)를 통한 분류가 뛰어날 것이라는 가정을 하였고, 실험을 통해 확인하려고 한다.

또한 심층신경망을 통해 만들어내는 모델은 히든레이어를 통한 피처의 다양한 선택을 가능하게 하므로, 다중선형회귀 특성의 추정모델이 유사성을 가질 수 있다는 가정을 전제로 하였다. 따라서

$$f(x) \approx \hat{f}(x) \tag{1}$$

$$\hat{f}(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{2}$$

이므로, RSE(Residual standard error)은

$$RSE = \sqrt{\frac{1}{n-p-1} RSS} \tag{3}$$

이고, 여기서의 RSS(residual sum squares)는

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \hat{\beta_1} x_{i1} - \hat{\beta_2} x_{i2} - \ldots - \hat{\beta_p} x_{ip} \right)^2 \end{aligned} \tag{4}$$

이므로, RSS의 감소가 p의 증가에 비해 작으면 더 많은 변수를 가진 모델이 더 높은 RSE를 가질 수 있다는 것을 알 수 있는데, 결론적으로 변수의 개수가 증가할수록 표본의 분산은 커지지만 RSE는 커지므로 언더피팅 또는 오버피팅 문제가 발생하지 않도록 적절한 개수의 변수를 선택하는 것이 필요하다[7]. 본 논문에서의 딥러닝모델을 기반으로 시험할 것이나, 피처의 선택을 모델에 위임하면 파라미터(p)의 수가 통제범위에서 벗어나기 때문에 성능 및 연구목적에 따라 p의 증가량을 통제변수로 넣고, 그 결과에 대해 실험을 통해 확인하는데 목적을 두었다.

본 논문에서는 딥러닝기반 응용트래픽 분류에서 한정된 데이터셋과 파라미터 수와의 관계에 대해 살펴보기 위해, 파라미터에 영향을 줄 수 있는 요소를 입력단계와 처리단계 2가지 경우로 구분해 보았다. 입력단계에는 패킷 헤더정보와 플로우 통계정보가 있으며, 처리단계에는 모델의 하이퍼파라미터가 있다. 처리단계의 하이퍼 파라미터는 입력정보의 피처를 추출하는 단계이므로 앞 단계인 입력단계에 의존적이고, 레이어가 깊을수록 학습효과가 뛰어나므로 모델링과정에서는 고정함으로써 고려사항에서는 제외했다. 딥러닝 기법에서 피처를 선정하지 않는 것이 일반적이나 입력단계의 정보에 변화를 주어, 결과적으로 파라미터의 수에 영향을 줄 수 있도록 하기 위해서 플로우 통계정보를 선택적으로 반영하였다. 사용된 플로우 통계정보는 패킷 길이와 발생 간격(Inter-Arrival Time)에 대한 정보이며, 패킷 전체 또는 첫 N개의 패킷의 길이 및 간격 정보 집계하여 13가지 통계 항목을 추출한 것이다. 통계 항목은 합 · 최댓값 · 최솟값 · 산술평균 · 기하평균 · 1분위 수 · 중앙값 · 3분위수 · 4분범위 · 모표준편차 · 표본표준편차 · 왜도 · 첨도를 포함한다. 패킷 헤더정보는 패킷 헤더정보를 대부분 수정 없이 반영하였으며, 암호화된 패킷이므로 패킷 길이정보로만 부가적으로 변형했다. 아래의 표 1은 수집한 응용 데이터셋의 종류를 나타낸다.

표 1. 수집한 응용 데이터셋

Bithumb	Coineone	Upbit	Excel
Teams	Excel	PPT	Word
Onenote	WEB_Excel	WEB_PPT	WEB_Word
Daum	Gmail	Nate Mail	Naver mail
Naver band	kakaotalk	Skype	Agoda
Airbnb	Goodchoice	Hotels.com	Netflix
Yanolja	11st	Coupang	Gmarket
Musinsa	Tmon	Wemakeprice	Facebook
Instagram	Kakaostroy	Naver blog	Tistory
Twitter	Disney+	Melon Music	Naver TV
Tving	Twitch	Wavve	Naver Series
Youtube Music	Hotels Combined	Kakao webtoon	Naver Webtoon
Kakao page	Youtube		

그리고 아래의 표 2는 전체 통계 정보에 포함된 첫 k개 통계 정보들 중, k를 변화시켜 만든 서브데이터셋을 나타낸다. 표 2의 플로우 통계정보 구성에서 k는 숫자 범위를 의미한다.

표 2. 데이터셋 별 플로우 통계정보와 파라미터 수

데이터셋	플로우 통계정보 구성	특징 수(개)	파라미터 수(개)
Set 1	패킷전체,	78	63,621
Set 2	패킷전체, 1~5	156	83,589
Set 3	패킷전체, 1~5, 1~10	234	103,557
Set 4	패킷전체, 1~5, 1~10, 1~15, 1~20, 1~25, 1~30	546	183,429

심층신경망은 가장 기본적인 MLP(Multi Layer Perceptron)로 선정했으므로, 실험결과에서는 모델의 성능지표 개별 값이 가지는 고저의 의미보다 성능지표별 값의 변화를 통해 p의 증가량 대비 모델의 적합성 변화를 살펴볼 예정이다.

Ⅳ. 실험 및 분석

데이터셋은 각 응용별 플로우 500개로 구성되어 있는데, 실험을 통해 5번째 패킷까지의 통계정보가 응용트래픽 분류모델의 성능을 높이는 정보이고, 다른 통계정보는 상대적으로 그렇지 않음을 우선적으로 확인했다. 물론 모델링에 사용되는 심층신경망의 종류와 형태에 따라 결과는 달라질 수는 있지만, 실험에서는 이것을 기반으로 표 3과 그림 1 같이 수집데이터 Set 1~4 각각에 대해 모델의 성능지표를 확인하는데 중점을 두었다.

아래의 표 3에서는 수집한 응용 데이터셋의 패킷 헤더정보와 플로우 통계정보를 Set 1~4로 구분한 후, MLP를 통해 각 응용별 분류한 결과를 의미하는 성능지표를 나타낸 것이다.

표 3. 데이터셋 별 성능지표

데이터셋	Precision(%)	Recall(%)	Accuracy(%)
Set 1	60.7	29.1	39.0
Set 2	65.6	36.7	45.5
Set 3	63.9	33.3	43.9
Set 4	59.4	33.4	42.2

아래의 그림 1은 표 3에서 살펴본 것을 재해석한 것인데, 데이터 Set 1에서 Set 4로의 순차적 구성은 추가적인 통계정보가 더해지는 추세를 의미한다고 볼 수 있는데, 통계정보가 추가되어도 모델의 성능지표는 감소하는 것을 확인할 수 있다.

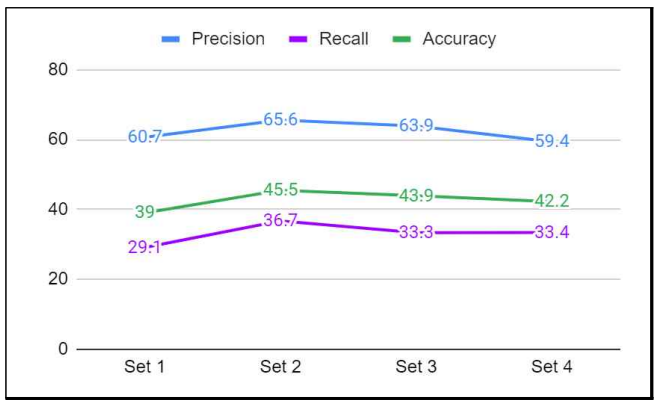

그림 1. 데이터셋 별 성능지표

결과적으로 데이터셋의 수가 한정적일 때, 추가적인 통계정보 중 일부는 정확도를 높이는데 기여했으나, 다른 통계정보들은 추정모델의 파라미터의 수를 늘리기만 할 뿐 오차를 크게 하여, 모델의 적합도를 떨어뜨리는 결과를 나타낸다고 볼 수 있었다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 딥러닝을 기반으로 하는 트래픽 분류에서 데이터셋이 한정적 일때 파라미터의 증가량이 모델의 적합도에 어떠한 영향을 미치는가에 접근하기 위해, 또한 분류의 성능을 높이기 위해 엔지니어 관점에서 추가하는 통계정보(피처)의 선택이 파라미터의 수를 증가시킬 수 있도록 했다. 일부 통계정보는 모델의 적합도를 향상시킬 수 있지만 데이터셋의 수량을 고려치 않고 파라미터의 수(통계정보)를 계속적으로 늘리면 모델의 적합성을 하향시킬 수도 있음을 실험적으로 확인하였다. 나아가 딥러닝 모델이 선택하는 피처의 수가 파라미터의 수를 구성하므로 데이터셋의 수를 고려하여 딥러닝 모델을 모델링하는 것이 중요할 수 있음을 유추해보았다.

향후 연구주제는 다음의 방향으로 확장하려고 한다. 첫째는 고정요소로 판단했던 통계항목을 모델의 적합성 향상에 대한 기여도에 대해 본 논문의 관점에서 딥러닝 모델을 통해 실험해보고자 한다. 둘째로 데이터셋을 한정적 요소로 두고 파라미터의 수에 변화를 주어 결과를 도출한 것에서 나아가, 파라미터를 고정한 가운데 한정적이었던 데이터셋을 단계적으로 증가시켜 보는 것을 딥러닝 모델별로 실험해보고자 한다.

데이터셋의 수가 늘어나면 늘어날수록 분류모델의 정확도가 높아지는 것으로 예측할 수도 있지만 구현된 모델에 따라 정확도의 증가가 둔감해지는 시점이 존재할 것으로 생각되어, 구현하는 모델별 데이터셋과 파라미터의 수는 합리적인 비율이 있음을 찾아보는 주제로 확장하는 것도 의미가 있을 수 있어 연구를 진행하려고 한다.

참 고 문 헌

[1] Y. LeCun et al., “Backpropagation Applied to Handwritten Zip Code Recognition,” Neural Computation, vol. 1, no. 4, pp. 541 - 551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[3] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” p. 9.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” p. 30.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” arXiv:1412.3555 [cs], Dec. 2014, Accessed: Mar. 26, 2022. [Online]. Available: http://arxiv.org/abs/1412.3555

[6] Ui-jun Baek et al, “DL-based Traffic Classification considering Relationship between Flows”, KICS 2022 Winter

[7] F. Sohil, M. U. Sohali, and J. Shabbir, “An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7,” Statistical Theory and Related Fields, vol. 6, no. 1, pp. 87 - 87, Jan. 2022, doi: 10.1080/24754269.2021.1980261.