기계 학습을 통한 네트워크 트래픽 변화 예측

고태진*, 양희규*, 샤이드 무하마드 라자*, 김문성**, 추현승*
*성균관대학교 소프트웨어대학
**서울신학대학교 교양학부
e-mail : {tjko9113, huigyu, s.moh.raza, choo}@skku.edu
moonseong@stu.ac.kr

Prediction of Change in Network Traffic with Machine Learning

Tae-Jin Ko*, Hui-Gyu Yang*, Syed Muhammad Raza*, Moon-Seong Kim**, Hyun-Seung Choo*
*College of Software, Sungkyunkwan University
**Dept. of Liberal Art, Seoul Theological University

요 약

본 논문은 네트워크 트래픽에 대한 동적인 변화에 대응하기위해 기존의 네트워크 트래픽 데이터를 이용하여 기계 학습을 사용하여 학습시킴으로써 이후 네트워크 트래픽 동향에 대해 분류하여 예측하는 연구에 관한 논문으로, 기계 학습의 종류 중 MLP(Multi-Layer Perceptron)를 이용하여 실험하였는데 MLP 의 구조와 학습 반복 횟수에 따른 정확도의 차이와 테스트 데이터 실험 결과를 정리하였다. 또한 이를 통해 얻어진 결과는 어떻게 사용 될 지와 정확도를 높이기 위해서는 어떤 요소가 영향을 끼치는지에 대해 논문의 방식과 비교하여 설명한다.

1. 서론

IoT(Internet of Things)디바이스의 증가로 인해 급증한 네트워크 트래픽을 관리하는데 응용 프로그램의 동적인 분류 및 예측, 식별은 관리 및 감시 등 여러 다양한 영역에 상당한 이점을 제공한다.

네트워크 트래픽 예측 연구에는 Proactive 한 방식과 Reactive 방식이 존재한다. 기존의 네트워크 트래픽 예측 연구에서는 Reactive 방식의 Threshold 감지 형태가 자주 연구 되었다. 이에 동적이고 방대한 네트워크 트래픽에 관리를 위해서는 동적인 Proactive 방식이 더 적합하다. 또한 현재 분류 방식에는 선택한 패킹 헤더 필드 (포트 번호) 또는 응용 프로그램 계층의 프로토콜 디코딩에 의존한다. 이러한 방식은 많은 응용 프로그램에서 예측할 수 없는 포트 번호를 사용하는 경우나 프로토콜을 알 수 없는 경우, 암호화 된 경우에는 분류가 불가능하다.

본 논문에서는 MLP 형태의 기계 학습을 사용하여 트래픽 변화 분류(예측)를 진행하고 결과에 대한 피드백과 얻어진 결과를 활용될 수 있는 분야를 제안한다.

2. 관련 연구

2.1 Network Traffic prediction [1]

네트워크 트래픽의 분석 및 예측은 광범위한 영역의 응용 프로그램에 적용 할 수 있으며 상당한 수의 연구가 지속되어왔다. 기존 컴퓨터 네트워크 응용 프로그램의 다양한 문제를 식별하기 위해 다양한 종류의 실험이 수행되고 있지만, IT 기술의 발달로 인해 증가하는 네트워크 트래픽에 대한 적절하지 못한 대처는 상업적으로 매우 큰 피해를 가져올 수 있다. 네트워크 트래픽 분석 및 예측은 신뢰할 수 있으며 정상적인 네트워크 통신을 보장하기 위해서는 사전 접근 방식이 적합하다.

2.2 Multi-Layer Perceptron [2]

MLP 는 피드 포워드 인공 신경망 클래스이다. MLP 는 입력 레이어, 숨겨진 레이어 및 출력 레이어의 최소 3 개의 노드 레이어로 구성된다. 입력 노드를 제외하고 각 노드는 비선형 활성화 기능을 사용하는 뉴런이라는 것으로 구성된다. MLP 는 교육을 위해 역전파라는 지도 학습을 활용한다. 지도 학습이란 입력데이터의 결과 값을 알고있는 상태에서 학습을 진행하여 입력 데이터에 대한 결과값과 출력 값의 차이를 사용하여 학습하는 과정을 의미한다. 선형으로는 분리할 수 없는 데이터를 구별 할 수 있다.

3. 제안 시스템

3.1 시스템 구성도

시스템은 패킷으로 구성된 트래픽을 측정할 수 만 있다면 어느 응용프로그램에도 적용 될 수 있다. 그림 1 은 추상적으로 시스템의 구성 한 것이다.

네트워크 트래픽을 수집하는 모듈에서 데이터는 주기적으로 수집 되고 데이터가 일정치 이상 수집되면 저장된 데이터는 최종적으로 학습 모듈로 이동한다.

학습 모듈로 이동하기 전에 데이터는 프로세싱 모듈에서 학습 모듈에 입력으로 사용되기 위해 정제된 후 MLP 로 구성된 학습 모듈로 이동한다. 이때 정제

되는 방식은 기준이 되는 시간 때의 패킷량과 5 초 후의 패킷량의 차이를 비교해서 변화 추이를 5 가지 기준으로 분류한다. 매우 높음, 높음, 보통, 낮음, 매우 낮음 순이다. 분류된 데이터는 5 초 간격으로 모아진다. 예를 들어 t 초에 측정되어 분류된 데이터는 t+5 초후의 데이터까지 묶여져서 입력으로 사용된다. 그리고 t+6 초후의 데이터가 라벨 데이터로 사용된다.

학습 모듈에서 학습이 완료된 예측 모듈은 데이터 센터에서 프로세싱 모듈을 거쳐 분류된 데이터를 입력 받는다. 학습한 예측 모듈은 입력 받은 데이터를 결과로서 다시 데이터 센터에 보낸 후 데이터 센터에서는 피드백 모듈로 예측이 성공적인지 아닌지에 대한 데이터를 전송한다. 피드백 모듈에서는 이러한 데이터를 일정치 모은 후 다시 MLP 로 전송한다.

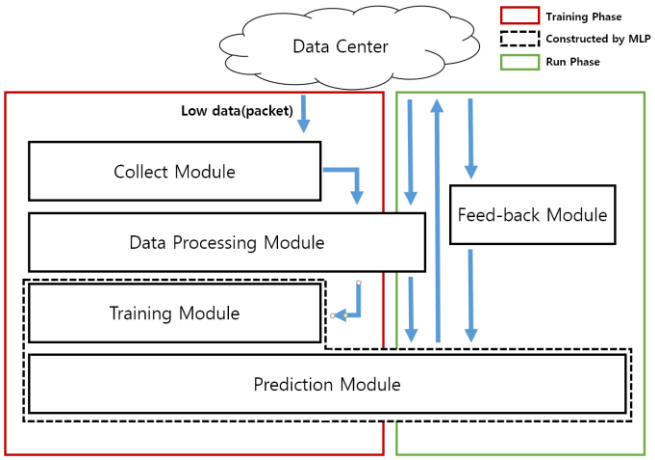

Figure 1 트래픽 예측 시스템 구성도

3.2 예측 모듈의 구조

프로세싱 모듈에서 분류가 5 종류이기 때문에 입력층과 출력 층의 개수도 5 개이다. 은닉 층은 128 개의 뉴런으로 구성되어 있는 은닉 층을 두 층으로 구성하였다.

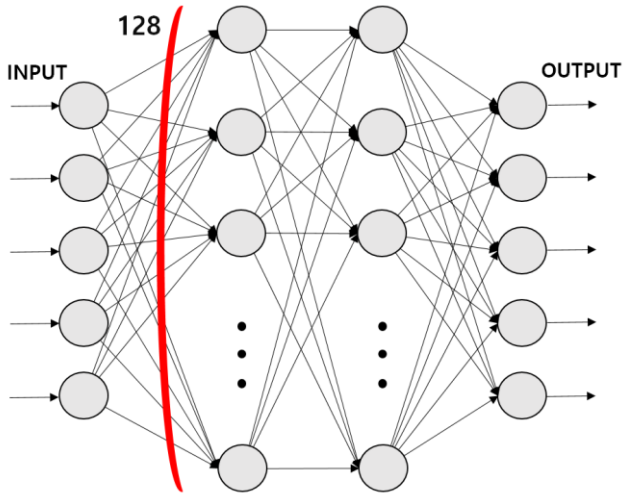

Figure 2 MLP 구조

4. 시스템 구현 및 평가

4.1 시스템 구현 방법

시스템을 구현할 때 시스템 언어로는 Python 을 사용하여 구현하였다. MLP 를 구현 할 때는 라이브러리로서 Tensorflow Keras, Numpy 를 사용하였다. MLP 의 계층적인 형태를 구현할 때 Keras 의 Dense Function 을 사용하여 각 레이어를 (5, 128), (128,128), (128, 5)와 같이 구성하였다. 학습의 epoch 는 2000, 3000 으로 실험 하였다. 배치 사이즈는 10 으로 설정 하였으며 비용 함수로는 CC(Categorical Crossentropy)를 사용 하였다.

4.2 예측 모델 평가를 위한 데이터

패킷 데이터는 " Network Traffic Characteristcs of Data Center in Wild" IMC 2010 paper 에 사용된 데이터를 사용하였다. 이 데이터는 pcap file 형태로 배포되었으며 UNIV1, UNIV2 라는 데이터 센터에서의 Packet Traces 가 담긴 pcap file 과 SNMP Data 와 Topology Data 를 함께 배포하였다. 이중에서 실험에는 Packet Traces 를 한 pcap file 이 사용되었다.

4.3 학습 결과 평가

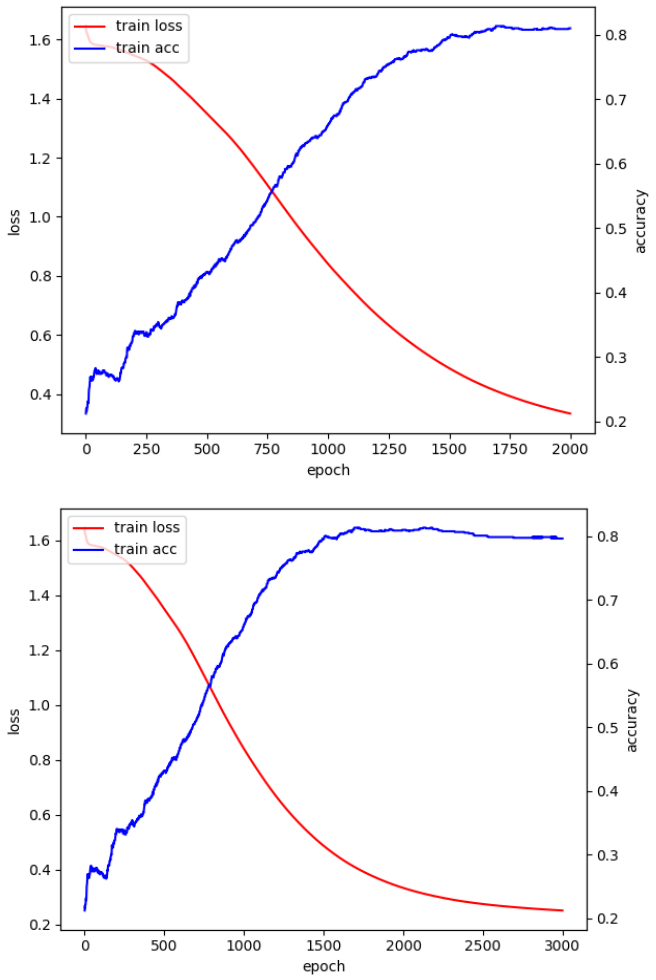

Figure 3 예측 모듈의 Loss 와 Accuracy 비교 (위 epoch 2000 아래 epoch 3000)

그림 3 은 학습 반복 횟수인 epoch 의 횟수에 따른 정확도와 손실 이력에 대한 그래프이다. Epoch 이 증가할수록 손실 이력은 감소, 정확도는 증가 한다. 반복 횟수가 2000 가 넘어갈수록 정확도의 증가 폭은 감소했는데 2300 을 넘어갈 쯤엔 정확도가 감소하는 모습을 볼 수 있다. 2000 의 정확도는 81%이고 3000 의 정확도는 79%로 측정되었다. 가장 최대치를 찍었을 2300 일 때 이후로의 학습은 오히려 역효과를 볼 수 도 있음을 알 수 있다. 최대 정확도는 82%로 2300 회 정도에 가장 큰 정확도를 보였다.

5. 결론 및 향후 연구 계획

본 논문의 시스템의 네트워크 트래픽 예측에 대한 정확도는 80%로 Proactive 한 사방 예방적 형태에서 활용할만한 정확도를 보인다.

시스템에서 제안한 네트워크 트래픽에 대한 예측은 매우 다양한 분야에 사용될 수 있다. 예를 들어 최신의 NFV (Network Function Virtualization)기술을 이용하여 부하가 예측되는 상황에 추가적인 VNF 를 신속히 보급해주어 부하를 줄이는 상황에 사용 될 수 있다[3]. 이외에도 네트워크 트래픽이 예측이 가능하다면 관리 및 감시 등 다양한 분야에 사용이 가능하다.

향후에는 단순한 MLP 모델이 아닌 Convolution Neural Network, Recurrent Neural Network 를 활용하여 정확도 및 신뢰성을 올리는 형태로 연구를 지속하고자 한다.

ACKNOWLEDGEMENT

본 논문은 과학기술정보통신부 및 정보통신기획평가원의 Grand ICT 연구센터지원사업 (IITP-2019-2015-0-00742), 과학기술정보통신부 및 정보통신기획평가원의 글로벌핵심인재양성지원사업(2019-0-01579)과 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-00421, 인공지능대학원지원)의 연구결과로 수행되었음.

참고문헌

[1] M. Joshi, T. Hassn Hadi, "A Review of Network Traffic Analysis and Prediction Techniques,“ arXiv preprint arXiv:1507.05722, 2015

[2] Y Bengio, G Hinton and Y. LeCun, “Deep learning,” nature, 2015

[3] S. Rahman, T. Ahmed, M. Huynh, M. Tornatore, and B Mukherjee, “Auto-scaling VNFs using Machine Learning to Improve QoS and Reduce Cost,” IEEE International Conference on Communications (ICC), 2018