한국통신학회 2018년도 동계종합학술발표회

LSTM 기반 학내 망 트래픽 발생량 예측

백의준, 이민섭, 구영훈, 김명섭

고려대학교

{pb1069, chenlima2, gyh0808, tmskim} @ korea.ac.kr

Prediction of Campus Network Traffic Usage Based on LSTM

Eui-Jun Baek, Min-Seob Lee, Young-Hoon Goo, Myung-Sup Kim

Korea Univ.

요 약

오늘날 네트워크 환경이 거대해지고 네트워크 기능을 사용하는 다양한 응용과 서비스가 증가함에 따라 네트워크 시스템의 트래픽 발생량이 나날이 증가하는 추세이다. 기존의 수동적인 대역폭 관리 시스템으로는 비정상으로 발생하는 트래픽에 대해 효율적인 대처를 하지 못한다. 따라서 효율적으로 트래픽 대역폭을 관리하기 위해서 네트워크 트래픽 발생량을 예측하는 새로운 방법이 요구된다. 본 논문에서는 LSTM 알고리즘을 이용한 네트워크 트래픽 발생량 예측 방법과 최적의 결과를 나타내는 네트워크 트래픽의 Feature-Set을 정의한다. 학내 망에서 발생하는 네트워크 트래픽을 수집하여 트래픽 패킷의 정보(프로토콜, 방향성) 조합으로 만들어낸 Feature-Set을 정의하고 여러 Feature-Set을 모델에 입력하여 발생될 트래픽의 양을 예측하고 비교 및 평가한다. 비교 및 평가 결과 전체 패킷 수와 방향성 및 프로토콜을 같이 고려했을 경우 전체 패킷 발생량만을 고려하였을 때 보다 오차가 현저히 떨어지는 것을 확인하였다.

Ⅰ. 서 론

오늘날 네트워크 환경이 거대해지고 네트워크 기능을 사용하는 다양한 응용과 서비스가 증가함에 따라 네트워크 시스템의 트래픽 발생량이 나날이 증가하는 추세이다. 기존의 수동적인 대역폭 관리 시스템으로는 갑작스러운 대량의 트래픽에 효율적이고 능동적인 대처를 할 수 없다. 또한 네트워크 관리자 입장에서 트래픽 모니터링 및 Qos(Quality of service) 정책 설정에 있어 네트워크 트래픽의 패킷 발생량 예측은 필수 불가결하며 이에 대한 연구는 활발히 이루어지고 있다.

이전의 연구에서 네트워크 트래픽 발생량 예측에 있어 기존에 사용하던 시계열 모형 예측 모델보다 신경 회로망(Neural Networks)을 이용한 모델이 높은 성능을 나타내는 것을 알 수 있다.[1,2] 하지만, 시계열 데이터를 분석하고 예측하는 것에 높은 정확도를 보이는 순환 신경망(Recurrent Neural Networks)모델을 이용한 네트워크 트래픽 발생량 예측 연구는 찾을 수 없었다. 또한 네트워크 트래픽은 비선형적이며 다양한 발생요인으로 많은 패턴이 나타난다. 이는 네트워크 트래픽 발생량을 예측할 때 많은 패킷들의 시퀀스 길이를 요구한다. 기존 순환신경망(RNNs)모델을 사용하였을 때 네트워크 트래픽 발생량 예측에 있어 모델 특성상 장기의존성 문제를 해결하지 못하고 기울기 소실(Gradient Vanishing)이 발생할 것으로 예상된다. 이에 본 논문은 순환 신경망 모델 중 기울기 소실 문제가 해결된 LSTM(Long Short- Term Memory Units)모델을 이용하여 실제 네트워크에서 발생하는 트래픽을 예측하는 방법을 제안한다. 다각도적인 실험결과를 위해 학습 비율, 학습되는 시퀀스의 길이를 변경하며 실험하였다. 실험에 사용된 네트워크 트래픽 데이터는 1분 간격으로 샘플링 하였고 이를 전체 패킷 수, 프로토콜 별 패킷 수, 방향성 별 패킷 수로 나눈 후 정규화를 진행하고 이를 다시 4가지의 조합으로 생성하여 실험하였다. 실험을 위해 학내 네트워크 에서 2016.11부터 2017.2월까지 발생한 트래픽을 수집하였으며 다량의 실제 데이터로 보다 자세하고 명확한 결과를 얻을 수 있었다.

본 논문은 1장 서론에 이어, 2장 본론에서 수집된 트래픽을 Feature set으로 가공하는 전 처리 과정과 입력데이터가 LSTM 모델에서 학습되고 결과를 도출하는 과정에 대해 서술한다. 3장 실험결과에서 앞서 정의된 Feature set과 LSTM 모델의 3가지 변수(학습 비율, 학습 횟수, Sequence Length)에 따른 결과를 표로 보이고 이에 대해 분석하고 설명한다. 마지막 4 장에서 결론 및 향후 연구에 대해 언급하고 본 논문을 마친다.

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원(No.2015R1D1A3A01018057) 및 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원(No. 2017-0-00513)을 받아 수행된 연구임.

Ⅱ. 본론

본 장에서는 수집된 트래픽을 Feature set으로 만들기 위한 전체 전처리 과정 및 Feature set의 구조와 입력데이터가 LSTM 모델에서 학습되고 결과를 도출하는 과정에 대해 서술한다.

트래픽 데이터는 수집 방법은 교내의 트래픽 데이터 수집 시스템을 사용한다. 외부 인터넷 망과 연결된 라우터에서 미러링을 통해 데이터 머신으로 트래픽 정보를 수집한다. 수집된 트래픽 데이터는 1분 단위 양방향 플로우 정보로 이루어져 있으며 이를 전 처리 과정을 통해 전체 패킷 발생량, 방향성 별 패킷 발생량(Forward, Backward), 프로토콜 별 패킷 발생량(TCP,UDP)으로 가공한 후 정규화를 진행한다. 정규화는 0~1 사이의 값을 가지도록 정규화 하였으며 계산은 수식 1 과 같다.

$$X = \begin{matrix} Traffic\ data\ sequence \\ Normalized\ traffic\ data\ sequence \end{matrix} \qquad A = \begin{matrix} Ai - Min(A) \\ Max(A) - Min(A) \end{matrix} \qquad (1)$$

이후 각각의 데이터를 조합하여 Feature-Set을 생성하고 이는 곧 각각의 Trace가 된다. 각 Trace가 포함하는 특징들은 표 1 과 같다.

표 1. 각 Trace가 포함하는 특징들

Trace	전체 트래픽	방향성	프로토콜
C1	O		
C2	O	O	
C3	O		O
C4	O	O	O

LSTM 모델은 순환 신경망의 일종으로 시계열 데이터의 분석 및 예측에 높은 정확도를 보이는 딥러닝 모델이며 기존의 순환 신경망(RNNs)의 기울기 소실로 인한 장기 의존성 문제를 해결한다. LSTM의 구조와 상세한 학습 과정은 [3]에서 설명하고 있다. 전 처리 과정을 마친 각각의 Trace는 LSTM 모델 내에서 학습된다. 학습 비율은 1e-4로 설정하였으며 다각도적인 실험 결과 분석을 위해 학습 횟수와 학습되는 시퀀스의 길이를 변경하며 실험하였고 그 내용은 표 2 와 같다.

표 2. 학습 변수

Learning_Ratio	Seq_Length	iterations
0.0001	60	500
0.0001	60	1000
0.0001	180	500
0.0001	180	1000
0.0001	180	2000

그림 1 은 모델의 입력과 출력 그리고 학습 과정에 대해서 간략하게 설명한다. 각각의 Trace가 가지는 Feature의 개수는 입력 X의 차원을 나타낸다. 시퀀스의 길이는 예측 값을 도출하기 위한 입력 X의 길이를 나타낸다. 입력 X는 1분간 발생한 패킷 개수이므로 시퀀스의 길이가 60이라고 가정한다면 이는 1시간 동안의 패킷 개수를 의미한다. 데이터의 전 처리와 학습변수 설정이 완료된 후 모델은 각 Step을 수행한다.

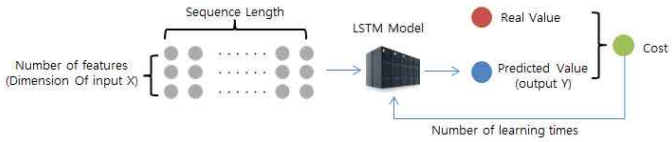

그림 1. LSTM 모델의 간략한 학습 과정

모델은 각 Step을 수행

Step 1. 모델이 입력시퀀스를 학습한다.

Step 2. 모델이 발생될 패킷 개수를 예측한다.

Step 3. 예측된 패킷 개수와 실제 발생 패킷 개수의 수식 2 와 같이 학습오차를 계산한다.

$$st = \sum_{=1}^{seq\ ength} y_{predicted} - y_{real} \quad (2)$$

Step 4. 정해진 학습 횟수만큼 Step 1 부터 Step 3 까지 반복하며 학습오차를 줄여나간다.

학습을 마친 모델에 Test 데이터가 입력되고 학습 단계와 같이 예측 된 값과 실제 값 사이의 오차를 계산한다. Test 단계의 오차는 평균 제곱근 오차(Root Mean Square Error)로 표현되며 계산은 수식 3 과 같다.

$$E = y_{predicted} - Y_{real}$$

$$RMSE = \quad (E_1 + E_2^2 + \cdots + E_n^2)/n \quad (3)$$

수식 2로 계산된 학습 단계의 Cost와 수식 3으로 계산된 테스트 단계의 RMSE로 각 Trace 별 결과를 분석하고 비교한다.

Ⅲ. 실험 결과

본 장에서는 앞서 정의된 Feature set과 LSTM 모델의 3가지 변수(학습비율, 학습 횟수, Sequence Length)에 따른 결과를 표로 보이고 이에 대해 분석하고 설명한다.

실험 결과는 표 3 과 같다. 결과를 통해 전체 패킷 발생량과 방향성에 따른 패킷 발생량 조합과 전체 패킷 발생량과 프로토콜에 따른 패킷 발생량 조합은 전체 패킷 발생량만을 고려하였을 때보다 학습 단계와 테스트 단계 모두 더 큰 오차가 발생하는 것을 확인하였다. 반면에 전체 패킷 발생량과 방향성, 프로토콜을 같이 고려했을 때 전체 패킷 발생량만을 고려했을 때보다 평균적으로 학습단계에서는 34% 테스트 단계에서는 12% 적게 발생하는 것을 확인할 수 있었다. 또한 학습 단계의 오차와 테스트 단계의 오차 모두 트래픽 시퀀스의 길이보다 학습 횟수에 큰 영향을 받는 것을 확인하였다. 그림 2는 Seq_len이 180, Iterations이 2000일 때, C1~C4 각 Trace의 학습 횟수에 따른 오차 감소 그래프이다. 그래프의 X축은 학습 횟수를 나타내고 Y축은 Cost의 값을 나타낸다.

표 3. Trace 별 실험 결과

Seq_len	Iterations	Trace	Cost	RMSE
60	500	C1	121.752	0.04168
		C2	127.816	0.04168
		C3	177.214	0.04736
		C4	76.115	0.03436
	1000	C1	111.169	0.03857
		C2	113.147	0.03967
		C3	109.481	0.03860
		C4	72.982	0.03401
180	500	C1	121.611	0.03987
		C2	127.878	0.04169
		C3	175.863	0.04727
		C4	76.141	0.03437
	1000	C1	111.116	0.03856
		C2	113.168	0.03968
		C3	109.745	0.03864
		C4	73.010	0.03401
	2000	C1	93.667	0.03631
		C2	90.068	0.03632
		C3	87.475	0.03562
		C4	67.536	0.03343

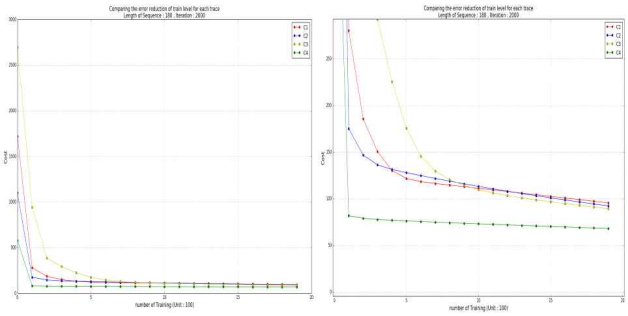

그림 2. 각 Trace의 학습 횟수에 따른 오차 감소 그래프

Ⅳ. 결론

본 논문은 LSTM 모델의 입력이 될 Feature-Set들을 제안하였고 이를 가지고 LSTM 모델을 통하여 네트워크 트래픽 발생량을 예측하는 방법을 제안하고 실험하였다. 실제 데이터로 실험한 결과, Feature-Set에 따른 결과 분석 및 비교를 통하여 더 정확한 예측 결과를 도출하는 Feature-Set을 찾을 수 있었으며 학습 횟수와 학습 시퀀스 길이가 예측 정확도에 미치는 영향의 차이를 알 수 있었다. 향후 연구로는 LSTM 모델의 예측 정확도 검증을 위한 기존 모델과의 비교 및 분석, 실험 결과와 Feature 간 상관관계를 검증을 할 예정이다.

참 고 문 헌

[1] Park, Dong-Chul. "Structure optimization of BiLinear Recurrent Neural Networks and its application to Ethernet network traffic prediction." Information Sciences 237 (2013): 18-28.

[2] Chen, Yuehui, Bin Yang, and Qingfang Meng. "Small-time scale network traffic prediction based on flexible neural tree." Applied Soft Computing 12.1 (2012): 274-279.

[3] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." Thirteenth Annual Conference of the International Speech Communication Association. 2012. p.2