LSTM 기반의 네트워크 트래픽 용량 예측

뉘엔양쯔엉, 뉘엔반퀴엣, 뉘엔휴쥐, 김경백
전남대학교 전자컴퓨터공학부
e-mail: truongnguyengiang.bk@gmail.com, quyetict@utehy.edu.vn, huuduyce@gmail.com, kyungbaekim@.jnu.ac.kr

LSTM based Network Traffic Volume Prediction

Giang-Truong Nguyen , Van-Quyet Nguyen, Huu-Duy Nguyen, Kyungbaek Kim
Department of Electronics and Computer Engineering, Chonnam National University

Abstract

Predicting network traffic volume has become a popular topic recently due to its support in many situations such as detecting abnormal network activities and provisioning network services. Especially, predicting the volume of the next upcoming traffic from the series of observed recent traffic volume is an interesting and challenging problem. In past, various techniques are researched by using time series forecasting methods such as moving averaging and exponential smoothing. In this paper, we propose a long short-term memory neural network (LSTM) based network traffic volume prediction method. The proposed method employs the changing rate of observed traffic volume, the corresponding time window index, and a seasonality factor indicating the changing trend as input features, and predicts the upcoming network traffic. The experiment results with real datasets proves that our proposed method works better than other time series forecasting methods in predicting upcoming network traffic.

1. Introduction

Recently, network traffic volume analysis [1] has attracted a lot of attention from researchers in many fields such as data centers and huge companies. Through the analysis in network traffic volume, the network administrators could know the possibility of a bottleneck situation and it enables them to prepare for the possible solutions ahead. Moreover, this analysis provides us the chance to detect network traffic anomalies which occur inside the network. Among various analysis techniques, estimating network traffic volume is one of the most interesting techniques. The network traffic would be sampled into a measurable unit during continuous equal time durations, known as time windows, and the next coming traffic of the next time window can be predicted based on the previous observed traffic values.

In past, many methods for predicting network traffic volume have been proposed [2][3]. However, they only use the previous observed traffic to predict the next traffic, and these methods may lose the accuracy of estimating the upcoming traffic under highly fluctuated traffic. In order to mitigate the impact of high fluctuation of traffic to the traffic estimation, a seasonality behavior of traffic can be considered. For example, the observed traffic increases very much in the morning of a day and it decreases in the evening. Sometimes, the observed traffic is fluctuated within a limited range. As a result, when not only the raw observation but also the seasonality of traffic volume is used, the accuracy for estimating the upcoming traffic volume increases.

Usually, this kind of estimating upcoming value based on the previous values can be resolved by using a time-series based methods such as moving average and exponential smoothing. However, recently, deep neural network based approaches have been proposed to estimate a time series value such as RNN (Recurrent Neural Network) and LSTM (Long Short Term Memory) neural network. Especially, LSTM neural network [6] has considered one of the most powerful method. It could work with not only the image, but also the multivariate data like natural language and also network traffic volume data, which includes the raw observation and our defined seasonality feature.

In this paper, we propose a new method for predicting upcoming network traffic volume by using LSTM neural network. This method does not predict the traffic volume directly, but estimate the traffic changes on a given time duration of a day. Then, the upcoming traffic is calculated by using the estimated traffic change rate and the previous traffic volume. Our work makes the following contributions:

- Firstly, we define some features which could be used for predicting the upcoming observed traffic. These features are the changing rate of observed traffic, the time window index and the changing trend, which is encoded into a one-hot vector.

- Secondly, we propose an LSTM neural network model which is composed of two layers of LSTM nodes and a dense network to extract the predicted upcoming changing rate of observed traffic, which is used for calculating the predicted upcoming traffic. This model is trained with 30 days of traffic logs and the trained model is used for estimating the traffic.

2. Network Traffic Prediction with LSTM network

When a gateway of a network captures all outgoing traffic, it can take an observation of traffic during a given time duration. We define this given time duration as a time window and the duration of the time window is set to t_w. In here, we can observe only interesting traffic such as DNS requests and SNMP requests during every time window. We define the observed traffic at the z^{th} time window as s_z.

In the proposed method, we manage the observed traffic in daily manner. That is, the observed traffic of a day becomes one collection. If there are k time windows in a day, the collection of the observed traffic of day y is denoted as $d_y = \{s_1, s_2, \ldots, s_k\}$. If there are n day traffic logs, we may have a day collection as $d=\{d_1,d_2,\ldots,d_n\}$.

With this basic setting, we consider three input features of LSTM neural network: 1) the changing rate of observed traffic, 2) the time window index, 3) the changing trend. As the first feature, the changing rate of observed traffic is the ratio between two continuous observed traffic, and the changing rate of the z^{th} time window is denoted as $ch_z = s_z/s_{z-1}$. Similar to the observed traffic, the changing rate of observed traffic is managed in daily manner and a collection of changing rate for a day y can be denoted as $c_y=\{ch_1, ch_2, \ldots, ch_k\}$. Accordingly, the changing rate of traffic in the z_{th} time window of y_{th} day can be calculated as the follow equation.

$$c_y[ch_z] = \frac{d_y[s_z]}{d_y[s_{z-1}]}$$

One exception of calculating the changing rate of observed traffic is the case of ch_1. In this case, we need to consider the traffic change between the last time window of the previous day and the first time window of the current day like the follow equation.

$$c_y[ch_1] = \frac{d_y[s_1]}{d_{y-1}[s_k]}$$

The second feature is the time window index of the given changing rate of the observed traffic. In this paper, we set the duration of time window, t_w, as 1 minutes, and the size of one day collection is 1440. Accordingly, in this paper, the time window index, i, lies within 1 and 1440.

The third feature is the changing trend of the traffic, and it is a kind of a representation of seasonality of traffic. We consider three changing trend: increasing trend, decreasing trend, fluctuating trend. Usually, these changing trends are observed in every day with similar shape and we use the same changing trend on the same time window of different days. The changing trend is represented with a vector where each element indicates each trends. That is, the change trend of the z^{th} time window of a day is denoted as $p_z=\{v_1,v_2,v_3\}$, where v_1,v_2, and v_3 represents increasing trend, decreasing trend and fluctuating trend, respectively and only one element has 1 value to indicate the trend. For example, {0,0,1} denotes the fluctuating trend and {1,0,0} denotes the increasing trend.

These three features are used as an input tuple $I_i = (ch_i, i, p_i)$ for the LSTM network whose architecture is depicted in Fig. 1. Specifically, we employ two layers of LSTM network and each of layer contains 20 LSTM nodes which generate output with 20 dimensions. Accordingly, as the final step, a node with linear activation function over the dense network with 400 dimension inputs is used to predict the changing rate of the upcoming observed traffic.

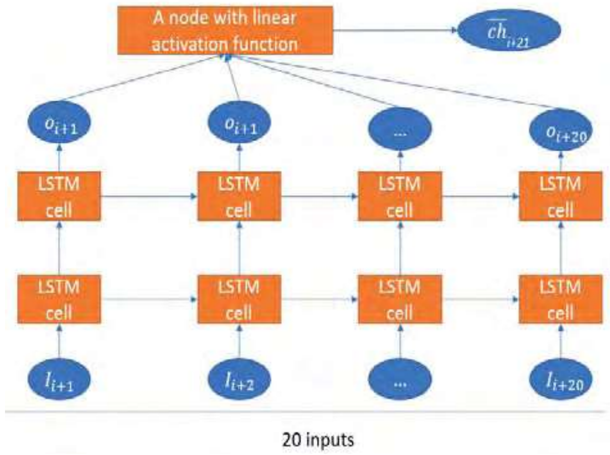

Figure 1. Overview of the proposed LSTM architecture

Consequently, in order to predict the traffic volume of the z^{th} time window of a day, we need to prepare the inputs as $I_{z-21}, I_{z-20}, \ldots, I_{z-1}$ for the proposed LSTM network, and the LSTM network provides the output of ch_z. Then, the predicted traffic volume of the z_{th} time window is calculated as $s_{z-1} \cdot ch_z$.

3. Evaluation

We conduct our evaluation based on the observed traffic of DNS query obtained from DNS-STAT:Hedgehog [4]. In which, the first 30 days of 2018 are used to train, while the next 10 days are used for testing. Regarding the training phase, we use 4000 training epochs, with batch size is 100 observation. The optimization used for training is Gradient Descent [5].

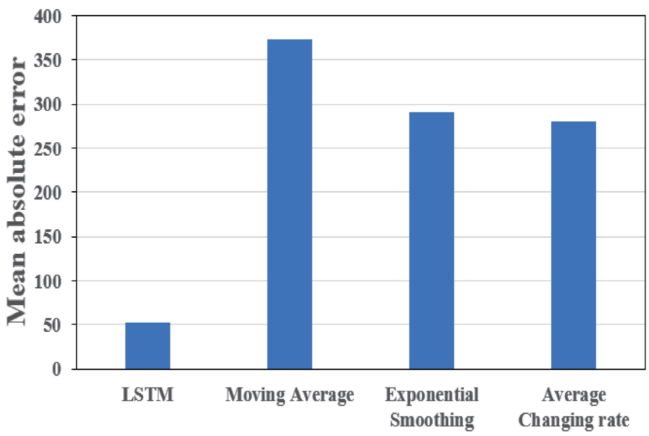

Figure 2. Evaluation comparison between LSTM based estimation and other methods

We use the mean absolute error [7] to compare the performance of the proposed method with other similar method, which are Moving Average estimation [2], Exponential Smoothing estimation [3], and the Average changing rate estimation. The result is shown in Figure 2 and it illustrates that that proposed LSTM based method outperforms other compared methods with much lower mean absolute error value.

4. Conclusion

In this paper, we have proposed a model based on LSTM neural network, which estimates the observed outgoing traffic of a network. Instead of using only raw observation as feature like other methods, we employ 3 features, which are the changing rate, current time in minutes and the changing trend. The result proves that our estimation is more effective than other methods for estimating the observed traffic of the next time window.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2017R1A2B4012559). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

[1] Marchetti, Mirco, et al. "Analysis of high volumes of network traffic for Advanced Persistent Threat detection." Computer Networks 109 (2016): 127-141.
[2] Sandgren, Niclas, Petre Stoica, and Prabhu Babu. "On moving average parameter estimation." Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.
[3] Gardner Jr, Everette S. "Exponential smoothing: The state of the art." Journal of forecasting 4.1 (1985): 1-28.
[4] http://stats.dns.icann.org/hedgehog/
[5] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).
[6] Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. "Recurrent neural network regularization." arXiv preprint arXiv:1409.2329 (2014).
[7] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.