Indonesian Journal of Electrical Engineering and Computer Science
Vol. 20, No. 1, October 2020, pp. 306~312
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v20.i1.pp306-312

Improving data quality using a deep learning network

Chulhyun Hwang[1], Kyouhwan Lee[2], Hoekyung Jung[3]
[1]Department of Smart IT Software, Kyungbok University, South Korea
[2,3]Department of Computer Engineering, PaiChai University, South Korea

Article Info

Article history:

Received Feb 10, 2020
Revised Apr 9, 2020
Accepted May 11, 2020

Keywords:

Data quality
Deep learning
IoT
LSTM
Recurrent neural networkl

ABSTRACT

IoT data is collected in real time and is treated as highly reliable data because of its high precision. However, it often exhibits incomplete values for reasons such as sensor aging and failure, poor operating environment, and communication problems. The characteristics of IoT data transmitted with high precision and time series are suitable to use LSTM, which is one kind of RNN. In this paper, when applying LSTM to data quality improvement in IoT environment where data are collected simultaneously from several sensors, it is suggested that it is effective to construct LSTM individually for each sensor accuracy.

Corresponding Author:

Hoekyung Jung,
Department of Computer Engineering,
Paichai University,
155-40 Baejae-ro, Seogu, DaeJeon, South Korea.
Email: hkjung@pcu.ac.kr

1. INTRODUCTION

IoT technology helps to create a more sophisticated virtual world by recording the real world more closely. Therefore, IoT technology is the next generation tool that transforms most of our everyday life and industry [1-6]. IoT is defined as a global network with an infrastructure that has self-configuring capabilities [7].

Since the sensors connected to each other using various communication technologies form a network while interacting with each other, the data transmitted from each object must be reliable. However, the level of quality of IoT data is threatened due to external exposure or moving objects, the physically unprotected networks or local area networks, and the aging of the natural environment or objects.

Due to the importance of IoT data, efforts to increase the reliability of the IoT sensor and the communication environment itself are continuing. In addition, it is recognized that the IoT environment is inferior, and efforts are being made to verify reliability or to replace it with the correct value while collecting data.

For example, statistical models have been developed for a long time as a way to improve quality problems by replacing missing data with predicted values [8-15]. IoT data has time-series characteristics because it periodically collects data from sensors. Using this characteristic, Recurrent Neural Network such as LSTM is used [16-28].

In this paper, when LSTM is applied to quality problems such as missing data generation in IoT environment, the accuracy of prediction depends on the dimensionality of the input data. In the IoT environment, multiple data are collected at the same time, so it is possible to construct an individual LSTM network for each sensor or to integrate a large number of data into one LSTM network. In this paper, we try to show how the difference between the two methods affects the quality of IoT data.

Our approach implements multiple LSTM networks to individually process the data collected from the sensors and a single LSTM network that batches the input data into an array. We compare how the two network configurations show performance differences when multiple data is inputted simultaneously. In this paper, we propose an efficient method for constructing LSTM in IoT environment.

The remainder of this paper is organized as follows. Section 2 describes the related research. Section 3 provides a detailed description of the proposed verification method. In Section 4, simulation results and analysis are performed. Section 5 presents conclusions.

2. EXISTING RESEARCHES ABOUT RESEARCH DATA

The direction of data quality improvement in IoT environment has mainly been focused on predicting missing data. Missing data prediction is a process of predicting and correcting a normal data value when data cannot be collected from the sensor due to various reasons. This process ensures the quality of the underlying data required to process or analyzes data after data collection.

Clinical data is a representative example of applying LSTM to the quality problem of IoT data. Since the clinical data consists of multivariate time series of observations, it is easy to apply in LSTM. As a result of applying the LSTM model, it proved that the performance is superior to that of the conventional hand-designed model and multi-layer perceptron [3].

Despite various research results, there is a lack of research on how to deal with sensor data input. For example, it is determined whether plurality of data to be simultaneously input is treated as one data set or individual input data.

3. DEEP LEARNING FOR DATA QUALITY IN IoT DATA

3.1. LSTM model design for data quality

LSTM learns data input in time series (tn-1… t0) and predicts data of next time(t1). Assuming that the predicted value provides an accurate value above a certain level, the difference between the predicted value and the input data indicates the possibility of error data. In particular, if the input data has a missing value, it can be corrected to the predicted value calculated by the LSTM.

Figure 1 is a general model that uses LSTM to improve the quality of time series data. In Figure 1, t + 1 is the current time at which the missing value occurred and t2 ... t0 is the value of the previous data of the sensor.

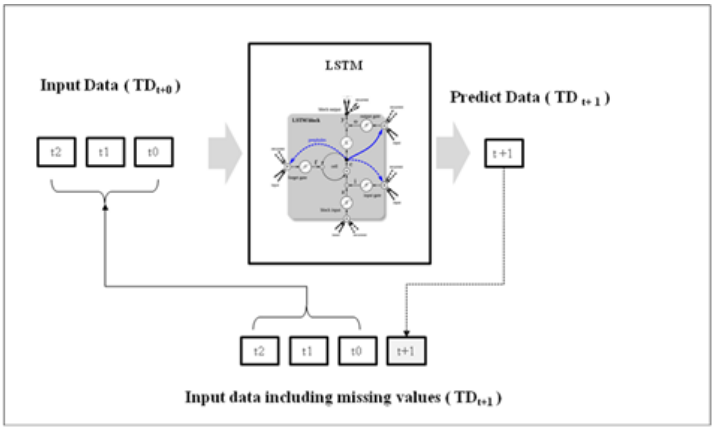

Figure 1. Missing value prediction model using LSTM

3.2. Two methods for LSTM input layer design in IoT data

We looked at the most common model of using LSTM for data quality problems. However, in actual IoT environment, it is rare to collect only one sensor data. Most of the time, data is collected from many sensors at the same time.

In this case, the network is designed by determining whether the data input at the same time is processed as one LSTM input or each independent LSTM. In this process, it should not be chosen as expectated that the input of data at the same time without experiment or verification of the data environment

will better describe the situation in which the data is generated. Therefore, in this paper, the difference of the prediction rate is verified through the experiment when the two methods are applied.

The verification method presented in this paper is shown in Figure 2 and Figure 3. First, Figure 2 is a method of integrating and predicting input data in one LSTM simultaneously. Figure 3 shows how LSTMs are individually constructed and predicted for each sensor. The error rate (RMSE) between the predicted result and the actual value of the test interval is calculated and compared, and then the network configuration method is selected.

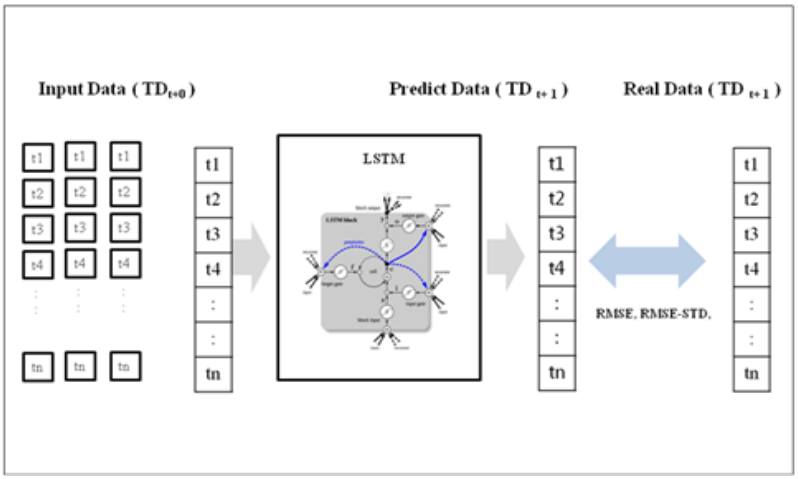

Figure 2. Multi dimensional input layer design

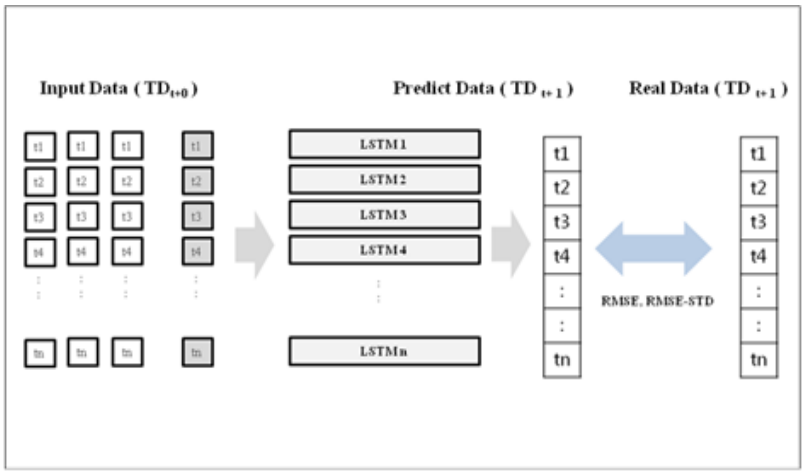

Figure 3. Single dimensional input layer design

4. EXPERIMENTS

4.1. Evaluation methodology

In order to calculate the error rate according to the above two methods, the experimental environment in which data is inputted simultaneously by a plurality of sensors is constructed. The LSTM environment applied to each experimental case is the same. In our implementation, our deep learning platforms, tensorflow and Keras, were used, and gradient algorithms were used to perform five epochs in each model. In addition, normalization and drop-out are applied to prevent over-fitting.

Some of the data used in the experiments are periodic and partly irregular, but all have time-series characteristics. Only the test results were analyzed without performing the verification process. The following Figure 4 presents the data pattern of the experimental data.

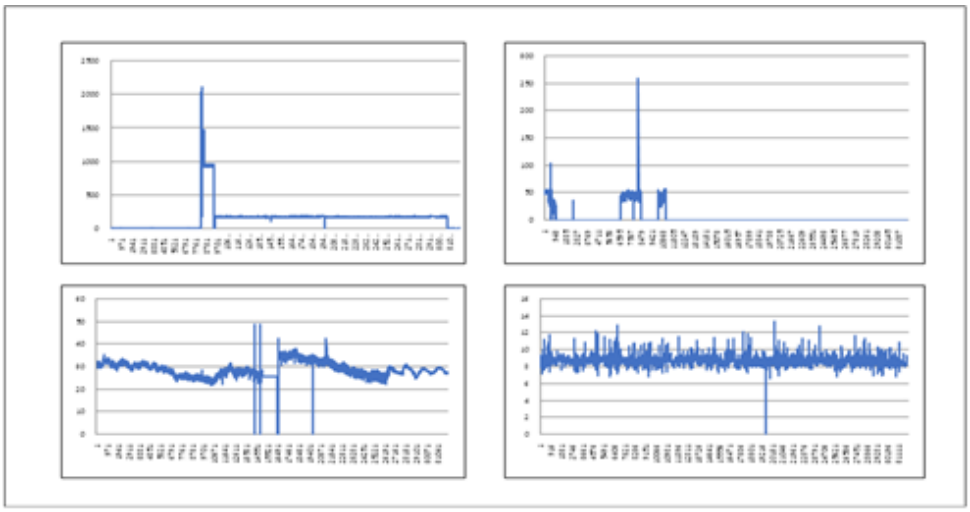

Figure 4. Patterns of experimental data

Data collected from 100 sensors were used in all experiments. To estimate the error rate according to long - term dependence, we use two kinds of data, total of 540,000 and the latest 3,600 data. A total of four experiments were performed using a single, multi-dimension network according to the input type of the network.

4.2. Results

Table 1 and Figure 5 show the average error rates of the sensors for each test method. RMSE (Root Mean Square Error) was used for each sensor. Experimental results show that when the LSTM is individually configured for each sensor, the error rate is low.

Table 1. As a result of calculating the error rate

Single Dimensional LSTM		Multi Dimensional LSTM	
54 million	3600	54 million	3600
4.802	17.593	47.191	28.500

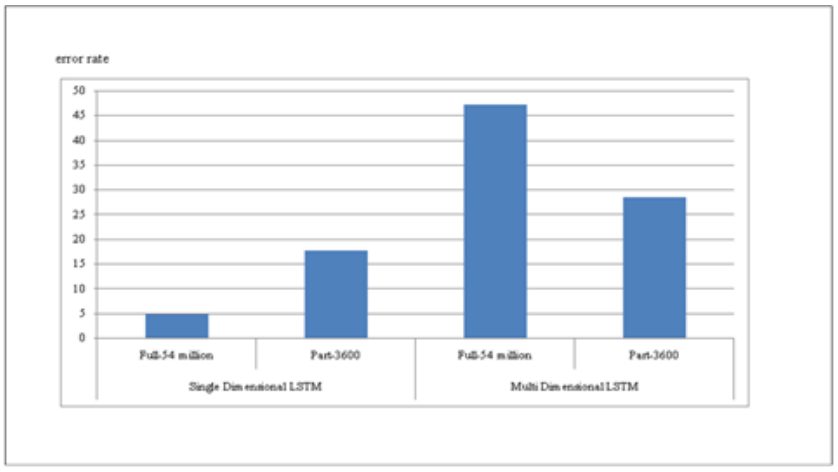

Figure 5. As a result of calculating the error rate

In addition to the average value of RMSE, the RMSE calculation result of each sensor also shows that the error rate is low by implementing LSTM individually in all cases. Table 2 shows representative values of the error rates of 100 sensors and shows the individual error rates in Figure 6.

Table 1. Representative value of error rate

Representative Value	Single Dimensional LSTM		Multi Dimensional LSTM	
	54 million	3600	54 million	3600
Average	4.802	17.593	47.191	28.500
Standard Deviation	4.684	22.221	104.321	41.161
Variance	21.941	493.770	10882.961	1694.192
Min Value	0.044	0.156	2.338	0.280
Max Value	18.472	123.772	965.859	248.772

Figure 6 shows that the error rate is limited to 100 in order to facilitate the comparison and is presented in order of error rate of 'all data - individual LSTM'.

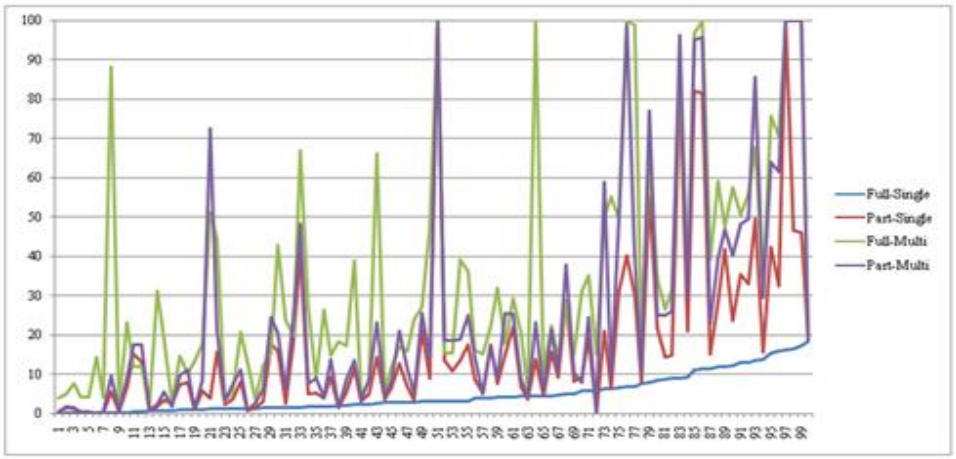

Figure 6. Error rate of individual sensor

4.4. Discussion

In the experiment using whole data, the individual LSTM construction method showed low error rate in all sensors. In some data experiments, 95 LSTM construction methods showed low error rate in 95 sensors. In both cases, it is suggested that the construction method of individual LSTM has higher predictive power than the method of inputting data at once. In particular, the error rate increases from 29% to 42% depending on the input method. This suggests that constructing and using LSTM by inputting collected data separately has better results in terms of long-term dependence.

5. CONCLUSION

The input data to be processed in one LSTM network is not only due to its ease of construction, but also to consider the effect of the data appearing at the same time. Experiments have shown that this method, however, reduces prediction accuracy compared to individual network conception methods. Therefore, we conclude that LSTM should be constructed separately for each number of time series data even in the environment where a large number of data is collected at the same time. Future research should include additional methods to consider the association between data collected at the same time and criteria to identify error data.

ACKNOWLEDGEMENTS

This study was supported by the research grant of Pai Chai University in 2020.

REFERENCES

[1] B. Fekade., et al., "Probabilistic Recovery of Incomplete Sensed Data in IoT," *IEEE Internet of Thing*, issue. 99, Jul 2017.

[2] H. B. Kim, J. B. Othman, L. Mokdad, S. Cho, P. Bellavista, "On collision-free reinforced barriers for multi domain IoT with heterogeneous UAVs," *Ubiquitous Computing Electronics and Mobile Communication Conference (UEMCON) 2017 IEEE 8th Annual*, pp. 466-471, 2017.

[3] V. Andrushchak, T. Maksymyuk, S. Dumych, M. Kaidan, O. Urikova, "Intelligent data flows management for performance improvement of optical label switched network," *Advanced Trends in Radioelecrtronics Telecommunications and Computer Engineering (TCSET) 2018 14th International Conference on*, pp. 1143-1146, 2018.
[4] O. Krasko, H. Al-Zayadi, V. Pashkevych, H. Kopets, B. Humeniuk, "Network functions virtualization for flexible deployment of converged optical-wireless access infrastructure," *Advanced Trends in Radioelecrtronics Telecommunications and Computer Engineering (TCSET) 2018 14th International Conference on*, pp. 1135-1138, 2018.
[5] V. Andrushchak, T. Maksymyuk, M. Klymash, D. Ageyev, "Development of the iBeacon's Positioning Algorithm for Indoor Scenarios," *Problems of Infocommunications. Science and Technology (PIC S&T) 2018 International Scientific-Practical Conference*, pp. 741-744, 2018.
[6] H. E. Ko, S. H. Pack, Victor C. M. Leung, "Spatiotemporal Correlation-Based Environmental Monitoring System in Energy Harvesting Internet of Things (IoT)," *Industrial Informatics IEEE Transactions on*, vol. 15, no. 5, pp. 2958-2968, 2019.
[7] R. Minerva, A. Biru, D. Rotondi, "Towards a Definition of the Internet of Things (IoT)," *IEEE Internet Initiative*, Torino, Italy, 2015.
[8] R. L. Prentice, B. J. WILLIAMS, A. V. PETERSON., "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika*, vol. 68, no. 2, pp. 373-379, Aug 1981.
[9] Z. C. Lipton., et al., "Modeling Missing Data in Clinical Time Series with RNNs," *Proceedings of Machine Learning for Healthcare 2016 JMLR W&C Track*, vol. 56, no. 21, Mar 2017.
[10] Y. Tiana, K. Zhanga, J. Lib, X. Lina, B. Yanga, "LSTM-based traffic flow prediction with missing data," *Neurocomputing*, vol. 318, pp. 297-305, Nov 2018.
[11] X. Ma, Z. Tao, Y. Wang, H. Yu, Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp 187-197, May 2015.
[12] R. Fu, Z. Zhang, L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 11-13 Nov 2016.
[13] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast., IET Intelligent Transport Systems," vol. 11, no. 2, pp. 68-75, Mar 2017.
[14] M. Liang, R. Wen Liu, Q. Zhong, J. Liu, J. Zhang, "Neural Network-Based Automatic Reconstruction of Missing Vessel Trajectory Data," *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, Mar 2019.
[15] Victor O. K. Li, Jacqueline C. K. Lam, Y. Chen, J. Gu, "Deep Learning Model to Estimate Air Pollution Using M-BP to Fill in Missing Proxy Urban Data," *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017.
[16] T. Shumway, "Forecasting Bankruptcy More Accurately: A Simple Hazard Model," *The Journal of Business*, vol. 74, no. 1, pp. 101-124, Jan 2001.
[17] D. H. Choi, J. O. Park., "The Application Method of Machine Learning for Analyzing User Transaction Tendency in Big Data environments," *The Journal of the Korea Institute of Information and Communication Engineering*, vol. 19, no. 10, pp. 2232-2240. Oct 2005.
[18] K. B. Kim, "Hybrid Neural Networks for Pattern Recognition," *Journal of Information and Communication Convergence Engineering*, vol. 9, no. 6, pp. 637-640, Dec 2011.
[19] Y. H. Kim, "Distributed Estimation Using Non-regular Quantized," *Journal of Information and Communication Convergence Engineering*, vol. 15, no. 1, pp. 7-13, Mar 2017.
[20] C. H. Hwang, H. S. Kim, H. K. Jung, "Detection and Correction Method of Erroneous Data Using Quantile Pattern and LSTM," *Journal of Information and Communication Convergence Engineering*, vol. 16, no. 4, pp. 242-247, Dec 2018.
[21] K. Cao, H. Y. Kim, C. H. Hwang, H. K. Jung, "CNN-LSTM Coupled Model for Prediction of Waterworks Operation Data," *Journal of Information Processing Systems*, vol. 14, no. 6, pp. 1508-1520, Dec 2018.
[22] İbrahim Kök, Mehmet Ulvi Şimşek, Suat Özdemir, "A deep learning model for air quality prediction in smart cities," *2017 IEEE International Conference on Big Data (Big Data)*, pp. 11-14, Dec 2017.
[23] H. Li, K. Ota, M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," *IEEE Network*, vol. 32, no. 1, pp. 4671-4679, Feb 2018.
[24] F. Tang, B. Mao, Z. Md. Fadlullah, N. Kato, "On a Novel Deep-Learning-Based Intelligent Partially Overlapping Channel Assignment in SDN-IoT," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 80-86, Sep 2018.
[25] M. Mohammadi, A. Al-Fuqaha, S. Sorour, M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923-2960, 2018.
[26] U. Narayanan, V. Paul, S. Joseph, "A novel approach to big data analysis using deep belief network for the detection of android malware," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1447-1454, Dec 2019.
[27] M. Akour, H. A. Sghaier, O. A. Qasem, "A comparative review on deep learning models for text classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 325-335, Jul 2020.
[28] N. A. Rahmad, N. A. J. Sufri, N. H. Muzamil, M. A. As'ari, "Badminton player detection using faster region convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1330-1335, Jun 2019.

BIOGRAPHIES OF AUTHORS

Chulhyun Hwang received the M.S. degree in 1995 and Ph. D. degree in 2017 from Department of Computer Engineering of Pai Chai University, Korea. From 1991 to 2000. He worked for Korea Navy as a Computer Officer. Since 2019, he has worked in the Department of Smart IT Software at Kyung Bok University, where he works as a professor. His current research interests include deep learning, machine learning, IoT, big data and artificial intelligence.

Kyouhwan Lee received the M.S. degree in 2019 from the Department of Computer Engineering of Paichai University, Korea. Since 1990, he has been working in Korea Water Resources Corporation. His current research interests include bigdata, information retrieval, ERP and IoT.

Hoekyung Jung received the M.S. degree in 1987 and Ph. D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked in the Department of Computer Engineering at Paichai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, information processing, information retrieval, machine learning, bigdata, and IoT.