

Fine Dust Forecast Based on Recurrent Neural Networks

Sunwon Kang, Namgi Kim and Byoung-Dai Lee*

Division of Computer Science and Engineering, Kyonggi University
Kyonggi, South Korea

sunwonkang97@gmail.com, ngkim.@kgu.ac.kr, blee@kgu.ac.kr

Abstract— In this paper, we propose a fine dust forecast model based on deep neural networks. The proposed model uses five kinds of air quality-related information as input variables and presents fine dust levels on an hourly basis. For training, we built training datasets by crawling air quality open data provided by the Korea Meteorological Administration and Seoul City. According to the experimental results, the proposed method achieved an RMSE of 8.966 for the prediction of fine dust levels after one hour.

Keywords— Deep learning, RNN, TensorFlow (2 개 더)

I. INTRODUCTION

Fine dusts are micro-sized dusts that cannot be seen with the naked eye; specifically, they refer to dusts with 10 μm or smaller diameters. Dusts are air pollutants and harmful substances. They not only induce corrosion by accumulating on constructions, but they can also accumulate in the body and cause respiratory, skin, and eye diseases, because unlike general dusts, they are not filtered by the nose, throat, or bronchus [1]. In South Korea, various industries have developed through the 1990s, and due to the high population density, the severity of fine dusts is increasing every year, unlike other countries and cities, where the fine dust levels are decreasing.

For this reason, South Korea urgently needs prevention of and countermeasures to fine dusts. To achieve this goal, we need to accurately forecast future fine dust levels and determine the factors that contribute to these levels. Statistical forecast models that have been used for this task include the multiple regression analysis model, which assumes and analyzes linear relationships between variables, the neural networks model, which uses artificial intelligence for nonlinear system modeling, and the decision-making model, which is a widely used inductive reasoning method for classification and forecasting. These statistical forecast models have been used to develop fine dust forecast systems that have in fact been applied in Seoul City. However, the accuracy drops for high-concentration forecasts for fine dusts above 100 $\mu\text{g}/\text{m}^3$ [2]. Therefore, this study proposes a fine dust forecast model using the hourly data of five forecast variables, including fine dust levels, using the long short-term memory (LSTM) model of recurrent neural networks (RNNs), which is appropriate for time-series data such as stocks and weather, among the deep learning models drawing attention as core technologies for AI. The results

showed that forecasts for the near future were highly accurate, and a future research plan has been established with the goal of performing accurate forecasts for a relatively distant future as well.

II. RELATED WORK

A. RNN and LSTM

The RNN model is used to deal with time-series data that are difficult to handle with general deep learning models. In order to introduce time concepts into a neural network to predict time-series data, past states must be stored in the model. In other words, we should understand the invisible effects of the past on the present. In the RNN theory, this understanding is defined as the “hidden layer of the past.” This concept can be represented in a simple graphic model, as shown in Fig. 1.

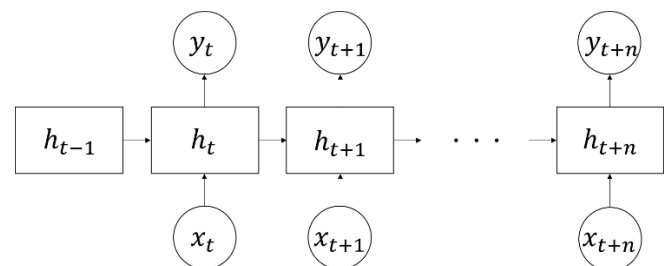


Figure 1. RNN graphic model

What should be noted here is that the value $h(t-1)$ of the hidden layer at time $t-1$ is stored together with the input $x(t)$ given at time t , and it is also delivered to the hidden layer at time t . This means that the state at time t is stored as the state at $t-1$, and all past states are recursively reflected in the value $h(t-1)$ of the hidden layer of the past [3].

However, one problem with respect to RNN is that the past learning results disappear if the chain becomes long. In other words, it cannot train time dependence over a long period. LSTM has been devised to solve this problem. It can train time dependence for both short and long terms. The LSTM blocks are illustrated in Fig. 2. LSTM can send information continuously to the next step with no large changes through cell states, and it can solve the long-term dependence issue with this structure. At first, LSTM selects which information to discard

*corresponding author

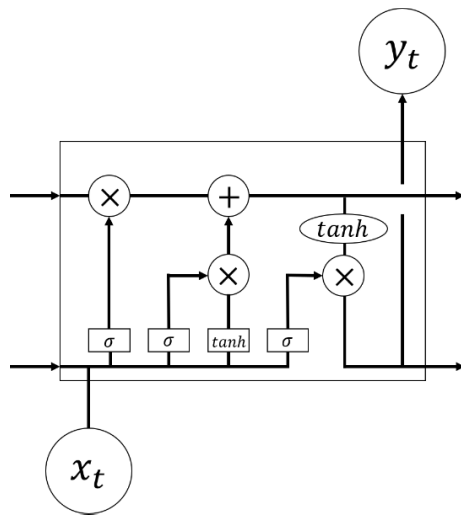


Figure 2. LSTM graphic model

through the forget gate layer. Then it creates new candidate values to be added to the cell states through the input gate layer and tanh layer. Next, it updates old cell states to new cell states using the values obtained in the previous step. Finally, it determines the final output value by multiplying the value obtained in the last sigmoid layer and the value output through the tanh function [4].

B. Existing Works

Until now, the multiple regression analysis model, the neural network model, and the decision-making model have been used to forecast fine dusts. The multiple regression analysis model is a statistical model that first sets a valid mathematical model representing the relationships among the variables, and it estimates the model using the measured values. Then it describes the relationships among variables through the estimated model or applies it for forecast or analysis [5]. The neural network model is a nonparametric model that generates output for an unknown input by training past experiences. In this model, if the causal relationships between input factors or between input and output are unclear, it can efficiently find their relationship [5]. The decision-making model draws decision-making rules in a tree structure to classify the target group into a few groups or to perform forecasts [5].

III. DATA

Fine dusts are largely classified by size into PM_{10} , $PM_{2.5}$, and $PM_{1.0}$. PM stands for “particulate matter” (fine particles in a solid or a liquid state, floating in air). PM_{10} refers to fine dusts with particle diameters smaller than 10 μm , $PM_{2.5}$ refers to fine dusts with particle diameters smaller than 2.5 μm , and $PM_{1.0}$ refers to fine dusts with particle diameters smaller than 1.0 μm [6]. In this study, PM_{10} was selected as the target for forecasting, considering that the available numerical change trends of $PM_{2.5}$ and $PM_{1.0}$ have a close relationship to PM_{10} , and their chemical compositions are similar [7]. The amount of available PM_{10} data is also much larger than that for other types [8].

TABLE 1. DESCRIPTION OF INPUT VARIABLES

Type	Description of input variables
PM_{10} (fine dust)	- Dusts of microparticles that are invisible to the naked eye
O_3 (ozone)	- An isomer of oxygen consisting of three oxygen atoms - Strong oxidizing power - Has toxicity for the human body; harms the respiratory organ if inhaled for too long
NO_2 (nitrogen dioxide)	- A nitrogen oxide generated when mixing carbon monoxide with oxygen - Has toxicity and strong oxidizing power - Classified as air pollutant - If inhaled, penetrates into the breathing airway or lung
CO (carbon monoxide)	- Generated by complete combustion when a fuel is burned with insufficient oxygen - Contained in large amounts in briquette gas or exhaust gas of automobiles - Combines with hemoglobin (Hb) in blood in lungs and interferes with the oxygen supply inside the body, which is the original function of Hb; as a result, the lack of oxygen causes poisoning symptoms
SO_2 (sulfur dioxide)	- Generated when sulfur burns as a compound of sulfur and oxygen - A colorless gas with an irritating odor and toxicity that penetrates into the mucous membranes of the human body - If inhaled in large amounts, causes runny nose, phlegm, and cough as well as pains in throat or chest and difficulty breathing

To select forecast variables, it is necessary to classify matters that have a large impact on fine dusts among the numerous air pollutants. For this purpose, we referred to a paper by the National Institute of Environmental Research, “Study on Meteorological Parameters Impacting on Fine Particle Concentration” [9]. According to this study, air pollutants such as NO_2 , CO , and SO_2 are closely related to PM_{10} . The official air pollution forecasts of the Korea Meteorological Administration include PM_{10} , $PM_{2.5}$, and O_3 forecasts. Based on this, we can estimate that O_3 could also be related to PM_{10} . For these reasons, five forecast variables were selected: O_3 , NO_2 , CO , SO_2 , and PM_{10} , which is a target for forecasting. The following table lists these forecasting variables.

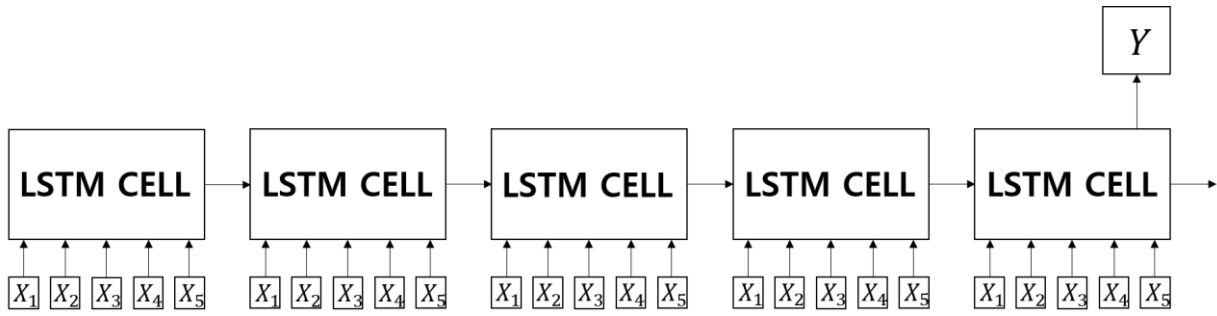


Figure 3. LSTM Model

IV. EXPERIMENTS

A. Experiment Environment

The prediction model of this experiment is shown in Fig. 3. For the experimental environment, this study used Python 3.6 based on Windows 10 and the TensorFlow CPU version. The input values were the five forecast variables, PM_{10} , O_3 , NO_2 , CO , and SO_2 . Five intermediate nodes were also defined, the same number as the input nodes. The final output was the fine dust forecast value after a certain time. The learning rate was set at 0.0005.

B. Dataset

We used the Excel format data in the “Final Confirmation Data” for Air Korea provided by the Korea Environment Corporation [10]. The period of the data was from January 1, 2016, to December 31, 2017, and the data was measured at the Seoul Gangnam-gu Measurement Station. Some data are missing due to defects, repairs, and inspections of the measuring instruments. In this case, if the data was empty in one cell, the arithmetic mean of the front and rear data was inserted. If two or more cells were empty, we inserted data from the Seongdong-gu Measurement Station. The closest station to the Gangnam-gu Measurement Station is the Gangnamdaero Measurement Station, but we used data from the Seongdong-gu Measurement Station, because the change characteristics of the street-side fine dust level were considered to be different from those of the urban fine dust level.

The forecast variables used in this study had different units and ranges, which meant that it might be difficult to analyze the data, because the comparisons between variables were not easy. Therefore, in this study, data normalization was performed to match the uniform training data range between 0 and 1. The data normalization formula used in this study is as follows:

$$\tilde{d}_i = \frac{d_i - \min\{d\}}{\max\{d\} - \min\{d\}}$$

C. Experiment and Results

The goal of this experiment was to forecast the fine dust level after one hour using the time-series data for three hours

as well as to forecast the fine dust level after 12 hours using the time-series data for 24 hours.

The root-mean-square error (RMSE) value was determined in order to measure the model's performance. The RMSE equation is as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (M_i - O_i)^2}$$

Where M denotes the model value, O denotes the observed value, and N, n indicates the amount of data. The number of iterations for this experiment was set to 300,000. The loss value graph of the fine dust forecast results for one hour after using three-hour time-series data is shown in Fig. 4.

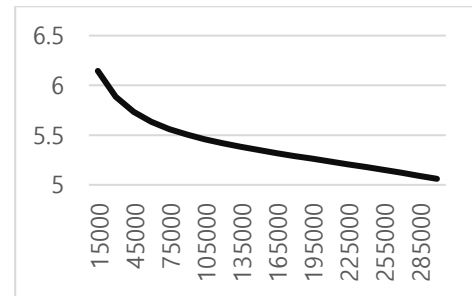


Figure 4. Training loss for on hour advance forecast using three-hour time-series data (3/1) Loss graph

The loss graph of the result of the fine dust forecast after 12 hours, using the 24-hour time-series data, is shown in Fig. 5.

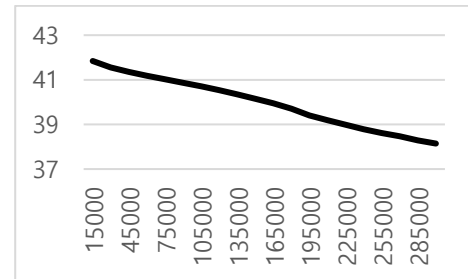


Figure 5. Training loss for twelve-hour advance forecast using 24-hour time-series data (24/12) 24/12 loss graph

The graph comparing RMSE values in each case is shown in Fig. 6.

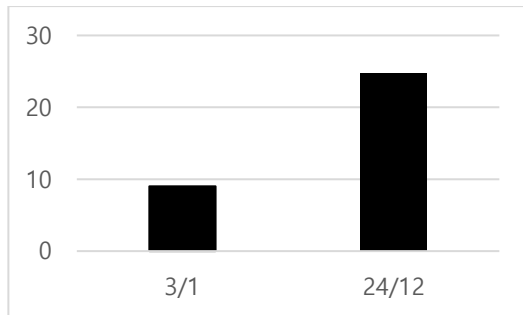


Figure 6. RMSE comparison graph

The standard deviation of the test data target for the fine dust after one hour using three-hour time-series data was 29.640, and the RMSE was 1/3 of the standard deviation of the target. These results show very high accuracy. However, the standard deviation of the test data target was 32.110, and the RMSE was a little over half, which suggests relatively low accuracy.

V. CONCLUSION & FUTURE RESEARCH

This study implemented the LSTM structure of RNN with TensorFlow for the hourly air pollution data of Seoul to forecast fine dust levels in the future. However, the fine dust level for the relatively far future was inaccurate. The forecast of fine dust that largely changes was also somewhat inaccurate, but the values of the forecast variables, O_3 , NO_2 , CO , and SO_2 , did not change much, except for PM_{10} . Based on this result, other forecast variables need to be added besides these five forecast variables. Therefore, in the future, we will implement a model by adding weather and climate data as forecast variables for predicting climate and meteorological data. The accuracy of the fine dust forecast for the distant future will also be increased through use of an LSTM model that we have improved, instead of the basic TensorFlow LSTM model used in the present study.

ACKNOWLEDGMENT

This work was supported by the GRRRC program of Gyeonggi province [2018-0262, Image/Network-based Intellectual Information Manufacturing Service Research]

REFERENCES

- [1] Doopedia, Retrieved from http://www.doopedia.co.kr/doopedia/master/master.do?_method=view&MAS_IDX=101013000716904
- [2] Wonjeong Yoon, "A Study on the Improvement of PM10 Statistical Forecasting Model," 2005.
- [3] Yusuke Sugomori, 2017. "SHOUKAI DEEP LEARNING TENSORFLOW, KERAS NI YORU JIKEIRETSU DATA SHORI," *wikibook*, 2017.
- [4] <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [5] Wonjeong Yoon, "A Study on the Improvement of PM10 Statistical Forecasting Model," 2005.
- [6] Doopedia, Retrieved from http://www.doopedia.co.kr/doopedia/master/master.do?_method=view&MAS_IDX=101013000716904
- [7] Eun-Kyung Cho, "Analysis of Chemical Compositions of PM10 and PM2.5 Fine Dust Particles Collected at Background Area of Korea in 2008," 2010.
- [8] <http://www.airkorea.or.kr/index>
- [9] National Institute of Environmental Research, "Study on meteorological parameters impacting on fine particle concentration," 2006.
- [10] <http://www.airkorea.or.kr/index>



Suwon Kang is an undergraduate student at the division of computer science and engineering, Kyonggi University, Korea. Her research interests include bigdata analysis and deep learning,



Namgi Kim is an associate professor at the department of computer science, Kyonggi University, Korea. He received the B.S. degree in Computer Science from Sogang University, Korea, in 1997, and the M.S. degree and the Ph.D. degree in Computer Science from KAIST in 2000 and 2005, respectively. From 2005 to 2007, he was a research staff of the Samsung Electronics. Since 2007, he has been a faculty of the Kyonggi University. His research interests include sensor system, wireless system, and mobile communication.



Byoung-Dai Lee is an associate professor at the department of computer science, Kyonggi University, Korea. He received his B.S. and M.S. degrees in Computer Science from Yonsei University, Korea in 1996 and 1998 respectively. He received his Ph.D. degree in Computer Science and Engineering from University of Minnesota, Minneapolis, U.S.A. in 2003. Before joining the Kyonggi University, he worked at Samsung Electronics, Co., Ltd as a senior engineer from 2003 to 2010. During the period, he has participated in many commercialization projects related to mobile multimedia systems. His research interests include cloud computing, deep learning, multimedia systems and medical image analysis.