Fine Dust Predicting using Recurrent Neural Network with GRU

Thanongsak Xayasouk, Guang Yang, HwaMin Lee

Abstract*: The particulate matter especially PM2.5 can cause respiratory, cardiovascular and nervous system damage as many studies prove. The monitoring and forecasting system are highly required. This paper proposed a predicting model to forecast PM10 and PM2.5 concentrations in Seoul, South Korea. The proposed model combines the recurrent neural network with GRU. The proposed model can extract the hidden patterns in the long sequence data as RNN's feature. The proposed model proved they could make satisfying particulate matter concentration in the urban area. The prediction results are reliable even for future 20 days. Meteorological data also contribute to higher predicting results as auxiliary data for the neural network. In further work, we will try to evaluate the model's universality with more urban cities. Additionally, try to combine other deep learning methods to improve accuracy and reduce time-consuming for prediction.

***Index Terms*: Air pollution, Deep Learning, GRU, RNN.**

I. INTRODUCTION

In the present climate, Air quality is still problematic for human health in South Korea. South Korea got ranks quite high 173/180 countries in the problem of air pollution [1]. The ambient air pollution consists of Carbon Dioxide (CO_2), particulate matter (PM), Nitrogen dioxide (NO_2), Sulphur dioxide (SO_2), Ozone (O_3). In South Korea, the most problem of air quality is Particulate matter. Particulate matter includes PM10 and PM2.5 that harms the respiratory system [2] such as heart disease, lung cancer, the burden of disease from stroke, asthma.

From such a problem, we proposed a short-term prediction of fine dust Using Recurrent Neural Network with GRU model. We apply the Gated Recurrent Unit model can predict a concentration of PM10 and PM2.5 in the next seven days (one week), ten days, 15 days, and 20 days and This experimental also show the prediction accuracy by using RMSE and MAE as the express average of the model prediction error. The result shows the proposed model got the effectively predict the concentration of particulate matter value in next future.

This paper organized as follow contents. In Section 2, related works are proposed. The methodology such as RNN and GRU architecture model are introduced in Section 3. Detail of implementation is shown in Section 4 and Section 5 is the conclusion.

Revised Manuscript Received on May 22, 2019.

Thanongsak Xayasouk, Dept. of Computer Science, Soonchunhyang University, Asan, Chungcheongnam, Rep. of Korea.

Guang Yang, Dept. of Computer Science, Soonchunhyang University, Asan, Chungcheongnam, Rep. of Korea.

HwaMin Lee, Dept. of Computer Software and Engineering, Soonchunhyang University, Asan, Chungcheongnam, Rep. of Korea.

II. RELATED WORKS

As mention previously, a large number of researchers have proposed various approaches for air pollution prediction by using deep learning models. Fine particulate matter has been predicting in various of the research previously. In [3] used the multi-task learning (MTL) model to hourly forecast the concentration of air pollution by climatology data. They use the data from air pollution measurement stations and climatology data from 2006-2015 and mainly use three different models for forecasting.

A stacked autoencoder (SAE) [4] is a model that used to extract the feature from the air quality dataset. This paper has shown the air pollution prediction of each station that has superior performance. Using Recurrent Neural Network (RNN) with LSTM [4] to forecasting the air pollution, this experiment evaluates the forecasting the PM2.5 concentration value for the next 4 hours at 66 stations around Taiwan. The results of this proposed model can predict PM2.5 value. Long short - term memory - fully connected (LSTM-FC) method [5] use to simulate the modification of particulate matter concentration and using an LSTM-FC to capture spatial data between the PM2.5 in the main measurement station and the nearest stations. This research evaluates the model on dataset consists of 36 monitoring station in Beijing. The experimental result of the LSTM-FC model shows an effective performance.

CNN-LSTM (APNet) [6] is a hybrid method of the Convolution Neural Network (CNN) and Long-Short Term Memory (LSTM) to apply to the fine particulate matter prediction system. The result of this paper shows the prediction precision of the CNN-LSTM (APNet) method and can also apply to the prevention and control of Fine particulate matter.

III. METHODOLOGY

In this topic, we proposed the data preprocessing to process and details of the Recurrent Neural Network and Gated Recurrent Unit model. This proposed of our model is to get the accuracy of Particulate matter (PM10, PM2.5) prediction in the next seven days, ten days, 15 days, and 20 days in short term prediction.

A. Data Preprocessing

Seoul is a capital city of the Republic of Korea and also has an adverse effect of fine particulate matter effect to Korean people's health.

This research, we determine the Seoul city as our demonstrate area and the dataset consists of the Meteorological (Date time, Rain, Rain Condition, Wind speed, Wind direction, Temperature, Humidity and Sky condition) data and the concentration of fine particulate matter (PM10, PM2.5) from 2015-2017 as shown in Table 1.

The important thing about fine dust prediction is wind speed. It is a reliable point for the transport of fine particulate matter. Wind speed has a relation to Climatic distribution process and process of air quality level control [7]. The horizontal transport may be determined by wind speed [8], lower wind speed is related to higher PM10 concentration.

We collected the dataset the fine particulate matter measurement data offered by the Air Korea website [9] and weather information data provided by the Korea Meteorological Agency website [10]. This dataset, we collected a dataset of 1,024,921 data from 1st January 2015 to 31st December 2017, in the South City area.

Table 1: List of model parameters

Attribute	Unit of Measurement
DateTime	yyyy-mm-dd
PM10	ug/m^3
PM2.5	ug/m^3
Rain	mm
Rain condition	0 - 3
Wind speed	m/s
Wind direction	Degree
Temperature	^oC
Humidity	%
Sky condition	0 - 3

B. Proposed Model Architecture

- Recurrent Neural Network (RNN)

RNN is one part of an artificial neural network (ANN) and a popular model that has a great result in many Neural Language Processing work. RNN is popular research in Translation [11], Image classification[12], Voice recognition [13], Object tracking [14], etc. RNN includes Long short-term memory (LSTM) and Gated Recurrent Unit (GRU) [15], etc.

RNN is used to make sequential information. Figure 1 shows the RNN architecture has unrolled and unfolded in the whole network. For unrolled, it means it writes out the network for the complete sequence (one layer per word).

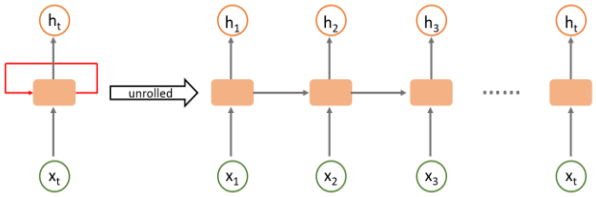

Figure 1. RNN Architecture.

(h_t) is the activation of the GRU at a time (t), The previous activation (h_{t-1}) and the candidate activation (h_t) is a linear reformation. The equation of RNN as the following equations:

$$h_t = \tanh(W_x h_t + W_h h_{t-1}) \quad (1)$$

Where h_t is current unit output, the w_x and w_h are weight for the inputs, x stands for input.

- Gated Recurrent Unit (GRU)

GRU [15] is included RNN and similar to a LSTM unit. The GRU unit consists of the reset gate and updates gate, in Figure 2 shown the GRU architecture. The reset gate designed to forget the previous state between the prior activation and the next candidate activation, the update gate used to select the number of the candidate activation that updates the cell state.

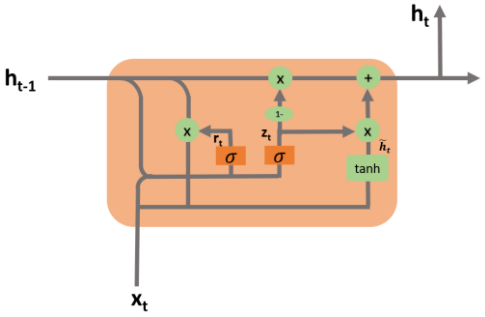

Figure 2. GRU Architecture.

GRU formulas are list below:

$$z_t = \sigma(W_z.[h_{t-1}, x_t]) \quad (2)$$

$$r_t = \sigma(W_r.[h_{t-1}, x_t]) \quad (3)$$

$$\tilde{h}_t = \tanh(W.[r_t * h_{t-1}, x_t]) \quad (4)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (5)$$

Where Z_t is an update gate and r_t is a reset gate.

The reset gate was introduced on combining the new input to the previous memory and the update gate determines the number

of previous memories that have around the system.

IV. EXPERIMENTS

A. Experiment Preparations

The training data consisted of two main parts, pollution data which consisted of PM10 and PM2.5 historical data. The datasets may contain missing data; we fill the mean value near the empty data value to make the original data more stable instead use zero.

All data in the dataset were normalized before a feed to the neural network. Consider the data range in different vectors is different; MaxMinScaler is applied to scale each data vector from 0 to 1 for training in the neural network. The batch generator is defined to feed data into the neural network as batches as the reason for the large dataset utilized in this paper. We separate the whole dataset into three parts. The training set used for model training, contains 92 % data, validation dataset, and test dataset are share the remain 8 % data. PM10 and PM2.5 are applied as target data, shift the concentration data in dataset one day ahead to prove the performance. We use Keras API running on top of Tensorflow. Training data generated into batches into GPU to improve the training speed. In the hope of improving the accuracy, the earlier 50 time-steps not contribute the accuracy in the lost function. We use early stopping to prevent the overfitting if the loss same for three epochs Adam optimizer as an optimizer for training. To evaluate results, mean absolute error (MAE) and root mean square error (RMSE) is practical to calculate the difference between predicted and observed data. RMSE as the equation 6, capable of evaluating the degree of change and data accuracy, the lower the RMSE generally stands for better performance, the MAE in equation 7 can reflect the actual situation of the prediction error.

B. Experiment Results

Table 2. Prediction performance for future days

Day	PM10		PM2.5		Accuracy
	RMSE	MAE	RMSE	MAE	
7	8.692	5.739	5.205	3.009	89.316%
10	8.906	5.851	5.366	3.155	91.474%
15	9.592	6.249	5.351	3.232	91.852%
20	10.059	6.493	5.559	3.323	91.651%

The proposed model applied to forecast future days particulate matter's concentrations. The predicting results as Table 2 show 15 days achieve best results with the highest accuracy. Noticeably, seven days predicting result seems not standing compared to another time scale, the reason probably the predicting result is the average result of the sub-areas in Seoul, part of the area's predicting result is better for a shorter time. This model is capable of predicting average concentration in next one month.

As Figure 3 and Figure 4 shown, the proposed model produced satisfying predicting results. Even with several days in time scale, predicted result still quite trustable. For PM10 prediction in Figure 3, overall predicted results are a little lower than the observed data. PM2.5 contributed better performance in general. Also, the proposed model can track some extreme high concentration days.

Figure 3 shows compassion between the PM10 prediction result and the real data in next 15 days using RNN with GRU model, the blue solid line stands for real data, and the dotted yellow line means the predicted result, vertical label represents the concentration of PM10 and the number in the horizontal label stand for predicting future hours. The model performed well, even in 15 days, and the extremely high concentrations days are predictable.

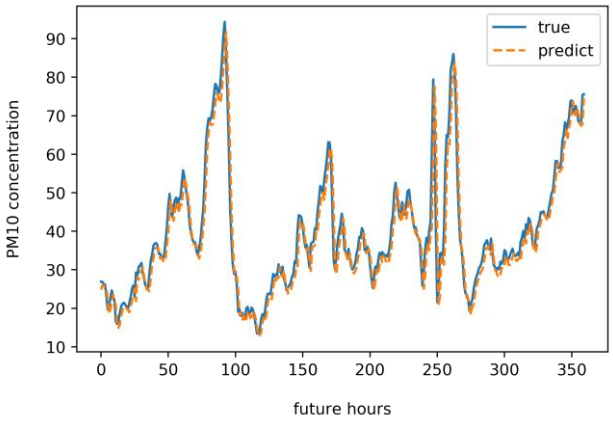

Figure 3. PM10 concentration in next 15 days (360 hours)

Figure 4 shows compassion between the PM2.5 prediction result and the real data in next 15 days using RNN with GRU model, the blue solid line stands for real data, and the dotted yellow line means the predicted result, vertical label represents the concentration of PM2.5 and the number in the horizontal label stand for predicting future hours. For PM2.5 prediction, the normal days predicted result more near to the real data, while may hard to predict the extremely high concentration in some days.

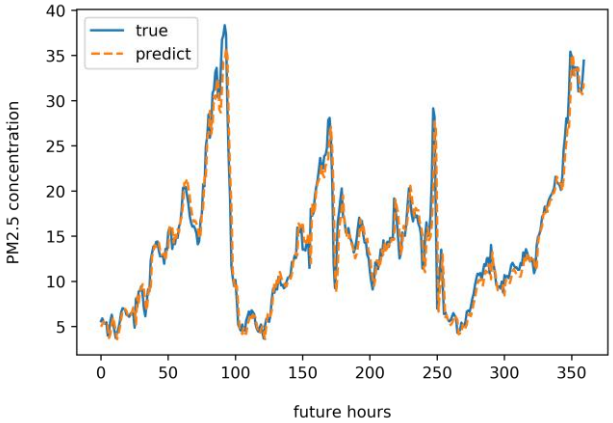

Figure 4. PM2.5 concentration in next 15 days (360 hours)

V. CONCLUSION

In this paper, an RNN with gated recurrent units experimented with predicting PM10 and PM2.5 concentrations. This model

can forecast future 20 days concentration in Seoul based on previous three years history data observed in sub-areas. Meteorological data as additional input data contributed higher accuracy in this model. The proposed model can give satisfying predicting results up to 20 days in the future. This model may contribute to improving the estimation of air pollution in urban cities. In future research, we will evaluate the model with more urban cities, predicting accuracy will be focused by combining the current model with other neural network or traditional models.

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-2017R1A2B4010570) and Soonchunhyang University Research Fund.

REFERENCES

1. Jung W South Korea's Air Pollution: Gasping for Solutions. In: Institute for Security and Development Policy. http://isdp.eu/publication/south-koreas-air-pollution-gasping-solutions/. Accessed 6 Apr 2019
2. Xing Y-F, Xu Y-H, Shi M-H, Lian Y-X (2016) The impact of PM2.5 on the human respiratory system. J Thorac Dis 8:E69–E74
3. Zhu D, Cai C, Yang T, Zhou X (2018) A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big Data and Cognitive Computing 2:5
4. Tsai Y, Zeng Y, Chang Y (2018) Air Pollution Forecasting Using RNN with LSTM. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). pp 1074–1079
5. Zhao J, Deng F, Cai Y, Chen J (2019) Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere 220:486–492
6. Huang C-J, Kuo P-H (2018) A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. Sensors 18:2220
7. Grivas G, Chaloulakou A (2006) Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment 40:1216–1229
8. Sayegh AS, Munir S, Habeebullah TM (2014) Comparing the Performance of Statistical Models for Predicting PM 10 Concentrations.
9. Air Korea. http://www.airkorea.or.kr/web. Accessed 6 Apr 2019
10. Korea Meteorological Agency. https://data.kma.go.kr/cmmn/main.do. Accessed 6 Apr 2019
11. Mahata SK, Das D, Bandyopadhyay S (2018) MTIL2017: Machine Translation Using Recurrent Neural Network on Statistical Machine Translation. Journal of Intelligent Systems. doi: 10.1515/jisys-2018-0016
12. Zhang D-Q (2018) Image Recognition Using Scale Recurrent Neural Networks. arXiv:1803.09218 [cs]
13. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp 6645–6649
14. Milan A, Rezatofighi SH, Dick A, Reid I, Schindler K (2016) Online Multi-Target Tracking Using Recurrent Neural Networks. arXiv:1604.03635 [cs]
15. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs]

AUTHORS PROFILE

Thanongsak Xayasouk is a M.S student in the Department of Computer Science at Soonchunhyang University. He received his B.S. degrees in Department of Computer Engineering and Information from the National University of Laos in 2016. His research interests include Deep learning, Big data.

Guang Yang is a M.S student in the Department of Computer Science at Soonchunhyang University. He received his B.S. in Deparment of Computer Software Engineering from Soonchunhyang University in 2018. His research interests include Deep learning, Big data.

HwaMin Lee is a professor in the Department of Computer Software Engineering at Soonchunhyang University. She received her B.S., the M.S. and the Ph.D. degrees in Computer Science Education from Korea University in Seoul, Korea in 2000, 2002, and 2006, respectively. Her research interests include Cloud computing, deep learning, IoT, mobile computing, wellness, and resource management and fault tolerant system for large-scale distributed systems.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication