Visual Deep Learning Models Analysis for Air Pollution Predictions

Hyesook Son* Seokyeon Kim† Hanbyul Yeon‡ Miyeon Lee§ Yejin Kim¶ Yun Jang‖

Sejong University, South Korea

ABSTRACT

The output of a deep learning model delivers different predictions depending on the input of the deep learning model. In particular, the input characteristics might affect the output of a deep learning model. In this paper, we propose a visualization system that can analyze deep learning model predictions according to the input characteristics with air pollution data. The input characteristics include space-time and data features, and we apply temporal prediction networks (LSTM, GRU), and spatiotemporal prediction networks (ConvLSTM) as deep learning models. We interpret the output according to the characteristics of input to show the effectiveness of the system.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Spatiotemporal data contain feature information, such as temporal and spatial information at the same time [3]. However, modeling a prediction model of spatiotemporal data is challenging because each field has a different degree and type of spatiotemporal correlation and complexity [2]. To solve this problem, a deep learning model has been studied that learns temporal and spatial patterns and makes predictions. Many studies have been conducted to predict spatiotemporal data with Gated Recurrent Units (GRU) networks and Long Short Term Memory (LSTM) networks [4–6]. GRU and LSTM are networks with the structure of recurrent neural network (RNN), and trained with historical sequence information. we propose a visualization system that can analyze deep learning models with air pollution data, one of the spatiotemporal data. The proposed system visualizes the predictions according to the input characteristics. The input characteristics include space-time and data features, and we apply temporal prediction networks (LSTM, GRU), and spatiotemporal prediction networks, Convolutional LSTM (ConvLSTM). We interpret the output according to the characteristics of input to show the effectiveness of the system.

2 SYSTEM EVALUATION

In this paper, we compare the performances of deep learning models to predict air pollutant data as spatiotemporal data. We utilize air pollutant data provided by kweather [1]. Data was collected from 413 discrete stations in Seoul. The collected data include $PM_{2.5}$, PM_{10}, noise, temperature, and humidity, and we utilize data measured every hour for 75 days from September 5, 2019, to November 18, 2019. We separated the datasets into the training dataset, validate dataset and test dataset at the ratio of 7: 1.5: 1.5. In this paper, we design $PM_{2.5}$ prediction models using these datasets and compare the $PM_{2.5}$ prediction performances depending on th data feature selection and model selection.

Deep learning for temporal forecast has been examined focusing on RNN, and representative algorithms are LSTM and GRU. ConvLSTM is a network structure that can be employed to predict spatiotemporal data by applying convolution LSTM structure. In this paper, we choose LSTM and GRU as temporal prediction algorithms and ConvLSTM as a spatiotemporal prediction algorithm. We predict $PM_{2.5}$ with temporal and spatiotemporal deep learning models. Then, we calculate the mean absolute percentage error (MAPE) from the test dataset as a measure of the performance of the model.

Our spatiotemporal data prediction modeling system, as shown in Figure 1 is a web-based application developed under the Flask framework and D3.js. Figure 1 presents our air pollutant prediction modeling system that enables us to compare spatiotemporal data prediction models and investigate the prediction performances. In Figure 1 (a), the scatterplot shows the correlation between input variables and probability distribution. The system visualizes spatial autocorrelation in (b), and temporal autocorrelation in (c). The Sankey diagram supports the modeling of the spatiotemporal prediction by combining features, deep learning models, and interpolation models, as shown in Figure 1 (d). (e) presents our prediction modeling parameter settings. Our system supports three time lags as an input time range, including 6, 24, and 72 hours. (f) presents interpolated predictions with the nearest neighbor algorithm. (g) shows the observed data. (h) presents the errors between the observed data and predictions. (i) shows the standard deviation of prediction over time. (j) presents the LISA visualization. The box plots in Figure 1 (k) show temporal predictions with the actual observed values.

As shown in Figure 1 (a), the Pearson correlation coefficient between $PM_{2.5}$ and PM_{10} is close to 1, and the scatter plot shows the strong linear correlation, which confirms that PM_{10} has the highest correlation with $PM_{2.5}$. Therefore, we can attempt to predict $PM_{2.5}$ by inserting $PM_{2.5}$ and PM_{10} features together in the GRU network and the LSTM network. The results are summarized in first row of Table 1.

We can try two things to improve the performance of the GRU and LSTM. First, the models are fixed with GRU and LSTM and reselect features for the training. Second, we fix the selected features and apply another model, such as the ConvLSTM. When we reconsider the feature selection, we need to identify the problem with the selected features. The selected features, PM_{10} and $PM_{2.5}$, have a strong linear relationship. If duplicate or nearly similar information is included in the input, the information may be insignificant in the prediction. Therefore, we train $PM_{2.5}$ again with temperature and humidity features, which have high linear coefficients next to PM_{10}. The results are summarized in second row of Table 1.

The MAPE of the ConvLSTM is lower than the ones of the GRU and LSTM networks. We can refer to Figure 1 (b) to see why the predictive performance is better when using a model reflecting the spatial information. In (b), Moran's I for $PM_{2.5}$ is 0.5382, which shows a relatively significant spatial correlation. Since $PM_{2.5}$ has high spatial autocorrelation, we can see that predictive performance is better when considering spatial information. We train the ConvLSTM with three features, including temperature, humidity, and $PM_{2.5}$. The result of ConvLSTM with the three features and 6 hours

*e-mail: atieer@naver.com
†e-mail: ksy0586@gmail.com
‡e-mail: hbyeon109@gmail.com
§e-mail: mioriii@naver.com
¶e-mail: kimyejin.kr@gmail.com
‖Corresponding author, e-mail: jangy@sejong.edu

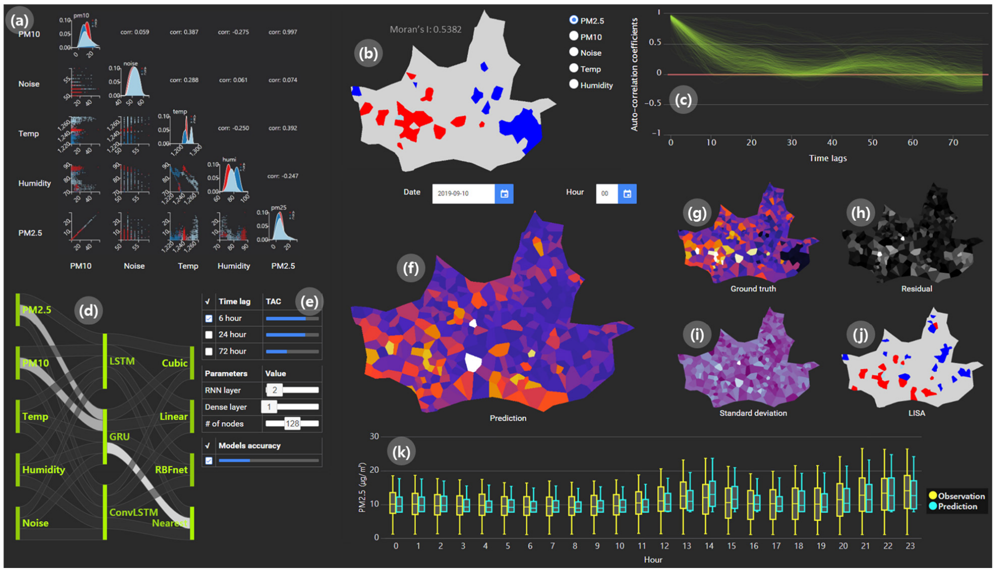

Figure 1: A visualization system for analyzing deep learning models.

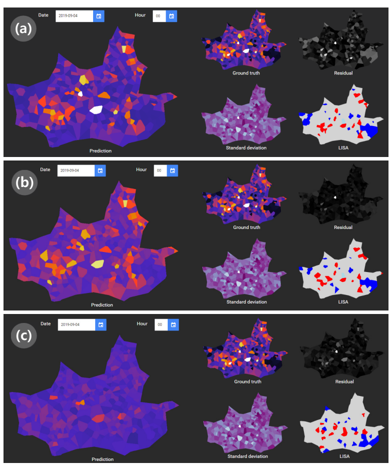

Figure 2: The visualizations of $PM_{2.5}$ predictions trained with lstm, including humidity, temperature, and $PM_{2.5}$. (a), (b), and (c) present the result of time lags 6 hours, 24 hours, and 72 hours, respectively.

time lag is 78.084%. After reselecting the features, we can see that predictive performance becomes better. We then explore how significant the spatial information of a feature can affect the prediction in Figure 2.

3 Conclusion

In this paper, we proposed a visualization system that can analyze deep learning models. We proposed an approach to select the appropriate features and deep learning model by analyzing correlations, spatial correlations, and temporal correlations for spatiotemporal data prediction with air pollutant dataset. We have attempted to interpret the prediction results for each case as we have stepped through the changes of features, time lags, and deep learning models. The proposed system enables deep learning modeling with spatiotemporal data and supports to interpret the causes for the results. During the modeling process, we can improve our understanding of the data and explore the deep learning models efficiently.

Table 1: Prediction accuracy of different time lags and models with $PM_{2.5}$, PM_{10}, temperature, and humidity.

Seleted features	Model	Time lag(hrs)	Accuracy(%)
$PM_{2.5}$ PM_{10}	GRU	6	30.585
		24	10.411
		72	27.137
	LSTM	6	29.479
		24	17.04
		72	27.777
$PM_{2.5}$ Temperature Humidity	GRU	6	54.648
		24	57.049
		72	50.102
	LSTM	6	50.9
		24	40.436
		72	17.114
	ConvLSTM	6	78.084

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2019-0-00795, Development of integrated cross-model data processing platform supporting a unified analysis of various big data models) and (2019-0-00374, Development of Big data and AI based Energy New Industry type Distributed resource Brokerage System).

References

[1] *www.kweather.co.kr*, 2020 (accessed January 23, 2020).

[2] X. Cheng, R. Zhang, J. Zhou, and W. Xu. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. In *International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2018.

[3] N. Cressie, T. Shi, and E. L. Kang. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745, 2010.

[4] C.-J. Huang and P.-H. Kuo. A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. *Sensors*, 18(7):2220, 2018.

[5] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, 231:997–1004, 2017.

[6] Q. Tao, F. Liu, Y. Li, and D. Sidorov. Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. *IEEE Access*, 7:76690–76698, 2019.