Visual Deep Learning Models Analysis for Air Pollution Predictions

Hyesook Son, Seokyeon Kim, Hanbyul Yeon, Miyeon Lee, Yejin Kim, Yun Jang
Sejong University

Data Visualization Lab
Sejong University

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.

Air pollution data

- Air pollution 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.

Deep learning modeling based on the correlation

- We can see that different temporal autocorrelation patterns appear for each variable in figure 2.
- As shown in Figure 1, 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.
- When we reconsider the feature selection, we need to identify the problem with the selected features. 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}.
- Second, we fix the selected features and apply ConvLSTM.

Figure 1. Scatter plot

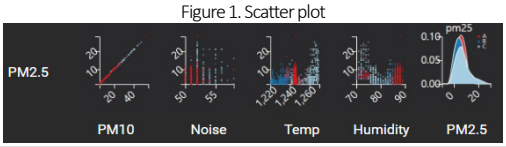
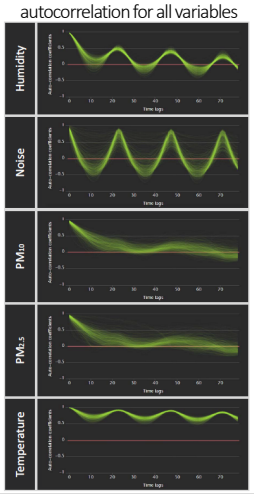
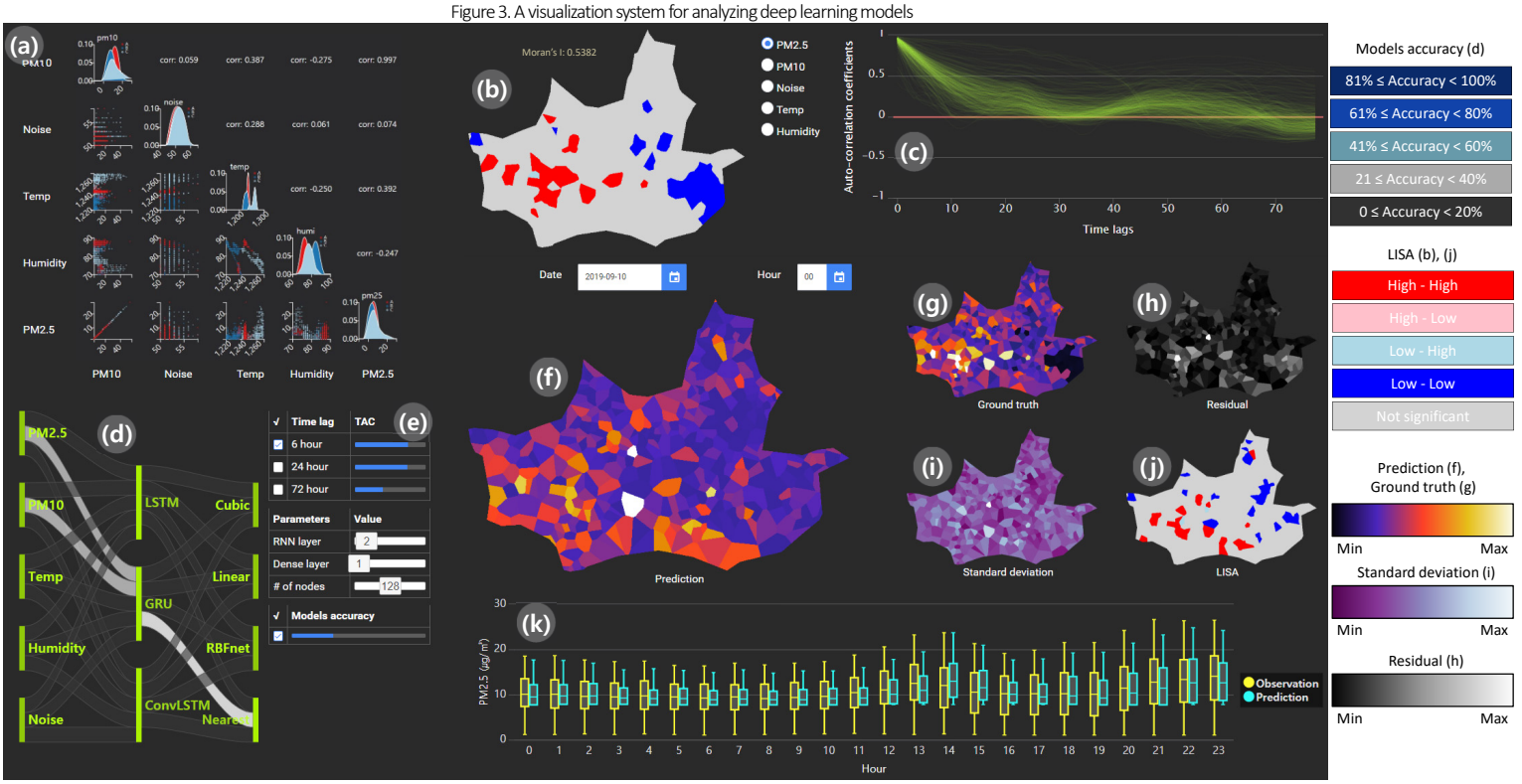


Figure 2. Temporal autocorrelation for all variables



Visualization system

Figure 3. A visualization system for analyzing deep learning models



- (a): The scatterplot shows the correlation between input variables and probability distribution
- (b), (j): Spatial autocorrelation (the LISA visualization)
- (c): Temporal autocorrelation
- (d): The Sankey diagram supports the modeling of the spatiotemporal prediction by combining features, deep learning models, and interpolation models
- (e): Our prediction modeling parameter settings
- (f): Interpolated predictions with the nearest neighbor algorithm
- (g): Ground truth data
- (h): The errors between the observed data and predictions
- (i): The standard deviation of prediction over time
- (k): The box plots show temporal predictions with the actual observed values

Air pollution predictions

- The results are summarized Table 1. The MAPE of the ConvLSTM with the three features and 6 hours time lag is lower than the ones of the GRU and LSTM networks.
- We can refer to Figure 3 (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 significant spatial autocorrelation, we can see that predictive performance is better when considering spatial information.

Table1. MAPEs of the deep learning models

Selected 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

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.
- During the modeling process, we can improve our understanding of the data and explore the deep learning models efficiently.

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).