Master's Thesis

Air Quality Prediction with Multi-model Features on 1-Dimensional Convolution

멀티 모달 정보와 1차원 컨볼루션을 이용한 미세먼지 농도 예측

February 2020

Graduate School of Engineering
Seoul National University
Electrical and Computer engineering major

JunYoung Choi

Air Quality Prediction with Multi-model Features on 1-Dimensional Convolution

지도 교수 정교민

Submitting a master's thesis of Public Administration

February 2020

Graduate School of Engineering
Seoul National University
Electrical and Computer engineering major

JunYoung Choi

Confirming the master's thesis written by JunYoung Choi

February 2020

Chair	이경무	(Seal)
Vice Chair	양인순	(Seal)
Examiner	정교민	(Seal)

Abstract

Air pollution, especially from particulate matters, has become serious problem in many countries. To cope with these abrupt pollution, there has been several studies to predict the temporal concentration of air pollution using deep neural networks. However, these studies have difficulties in predicting accurately since the air quality is complexly correlated with various types of multi-modal features over a long time. In this paper, we propose a new architecture to predict air qualities of particulate matters incorporating deeply stacked 1-dimensional CNN with residual connection and attention mechanism. Specifically, 1-dimensional CNN extracts high-level features with large receptive fields and attention mechanism captures complex correlation among these features. Through extensive experiments with Seoul air pollution data and public benchmarks, we verify our architecture achieves state-of-the-art result in PM2.5 and PM10 prediction.

Keyword : Air pollution, 1DCNN, Attention
Student Number : 2018-27571

Table of Contents

Chapter 1. Introduction

The problem of air pollution has threatened human health since the industrial revolution. The air pollution problem can causes difficulty in breathing, coughing, asthma, and deterioration of existing respiratory and heart diseases in humans. Such adverse effects can lead to huge social costs, such as reduced productivity and reduced quality of life.

Recently, the issue of particulate matters in the air has emerged as the most important problem. As a result, there is a growing demand to prevent the impact of particulate matters in advance by predicting the densities of them. However, there is a fundamental limitation in predicting particulate matters' densities: Various factors such as precipitation, wind, traffic volume and industrial activity affect particulate matters' densities through complex interactions and vary dynamically with time and space [13]. Thus, it is difficult to develop model the effects of these factors properly. For example, interaction between particulate matters and wind speed can be affected by precipitation, humidity and whether it is day or night.

Therefore, considering multi-model factors and spatio-temporal relations among factors is necessary when modeling interactions. To tackle these issues, several studies have proposed deep learning based model. These works generally used LSTM [4] based models for a temporal prediction of the air quality [18]. However, LSTM is hard to capture long-term dependency over time and have not enough capacity to learn complex correlation among features. These limitations cause performance degradation on long-term prediction. In this paper, we focus on resolving these problems using 1DCNN (1 Dimensional Convolution Neural Network) and an attention mechanism [11]. To capture complex correlations among multiple features, we suggest 2-stage approach; 1DCNN module for intra-

correlation within data source and attention module for inter-correlation among data sources. First, our model uses 1DCNN layers to capture Intra-correlations among features within data source. Deeply stacked 1DCNN has shown a great success to extract high-level features from raw data and resolving long term dependency problem at various areas such as speech synthesis [17], machine translation [2], etc. Therefore, our approach is to apply Deeply stacked 1DCNN in order to capture long term dependencies on a target air pollution feature. Second, our model takes advantage of the attention mechanism to learn the complex correlation among different sources. Specifically we discriminate types of data according to data sources and extract features from each data source using Deeply stacked 1DCNN. Then, we apply the attention mechanism between particulate matters and other features. By doing so, our model can learn the influence on the particulate matters of other features while considering the distribution of data sources. Lastly, the final LSTM layer gets output of attention mechanism as an input and predicts the densities of particulate matters.

We conduct experiments on predicting PM10 (particulate matters < 10μm) and PM2.5 (particulate matters < 2.5μm) with Seoul air pollution dataset (which is recently published by Seoul). Through these experiment, we verify that our model outperforms all other state-of-the-art methods, especially on long term prediction. Furthermore, we explore an ablation study for identifying where the benefits come from including 1DCNN and attention mechanisms. Also additional experiments provide analytical explanation for different combination of data sources.

Chapter 2. Related Works

2.1. Time Series

Traditionally, time series analysis has long been studied in economics, engineering, etc. Several models were presented for the time series prediction, such as ARIMA(Autoregressive Integrated Moving Average), and VAR(Vector Auto Regresion). The ARIMA model is a combination of AR (auto Regressive) and MA (Moving Average) models, and takes into account the difference between current and past data. The VAR model is a generalized model for AR model to capture relationships between different time series. The above models have been successful for a long time, but have had a lot of difficulties predicting in environments where complex and dynamic changes and the correlation between factors is not constant. Recently, there have been papers that attempted time series prediction using Neural network. Yao Qin et al [2] used Dual-Stage Attention-Based Recurrent Neural Network to predict stock prices. They use first-stage attention to extract features of different time series. After first-stage attention, they use second stage attention to capture relevant hidden states across time steps. Li et al [3] used graph to express the city's traffic speed and then used GCN (Graph Convolution network) to predict the traffic speed. They use graph to capture the spatial dependency and capture temporal dependency using encoder-decoder architecture with scheduled sampling. Zhang et al [19] proposed a spatio-temporal model to predict the flow of the city by receiving traffic each hour as input by building the CNN(convolutional natural network) deep using the response

connection. Their model is composed with 4 components, each modeling temporal closeness, period, trend and external factors. They use deep convolutional neural architecture with residual connection for temporal closeness, period and trend.

2.2. Air pollution

There were several studies that applied the deep learning method to air pollution prediction. Cheng et al [20] use feed forward network and recurrent neural network to capture relation between spatial features and sequential features. Liang et al [18] use multi-level attention and recurrent neural network for air quality prediction and water quality prediction. They use local spatial attention and global spatial attention for spatial dependences and temporal attention for temporal dependences. Yi et al [1] proposed feed forward neural network based spatio-temporal model for air pollution prediction. They convert data using spatial transformation component for consistent input and fuse heterogeneous data for prediction.

Chapter 3. Model Architecture

3.1. Problem Definition

For predicting air quality, our data consists of three types of data: 1) air pollution in the target city, 2) weather around the target city, and 3) air pollution data in neighboring cities. Each type of data has its own station for observation and we denote station for the target city, weather and nearby cities by a_i, w_j, c_k respectively. For fixed time interval T, each feature $\mathrm{x} \in \mathrm{R}^T$ observes discrete/continuous values every hour and each station a, w, c observes feature values $\mathrm{X}^a = \left(x_1^a,\ x_2^a, \dots\dots, x_{\mathrm{N}_a}^a\right) \in \mathrm{R}^{\mathrm{N}_a * T}$, $\mathrm{X}^w = (x_1^w,\ x_2^w, \dots\dots, x_{\mathrm{N}_w}^w) \in \mathrm{R}^{\mathrm{N}_a * T}$, $\mathrm{X}^c = (x_1^c,\ x_2^c, \dots\dots, x_{\mathrm{N}_c}^c) \in \mathrm{R}^{\mathrm{N}_a * T}$. Our goal is to predict the concentrations of PM10 and PM2.5, ($x_{pm}^{\mathrm{a}_i}$over the next H hours for each target city at station a_i) using features from station a_i, w_j, c_k past T hours:

$$\mathrm{f}(\mathrm{X}^a, \mathrm{X}^w, \mathrm{X}^c) = \{x_{pm}^{\mathrm{a}_i}\}$$

In order to build prediction model, we train fusing collected data with L1 loss between $\mathrm{f}(\mathrm{X}^a, \mathrm{X}^w, \mathrm{X}^c)$ and real data $x_{pm}^{\mathrm{a}_i}$.

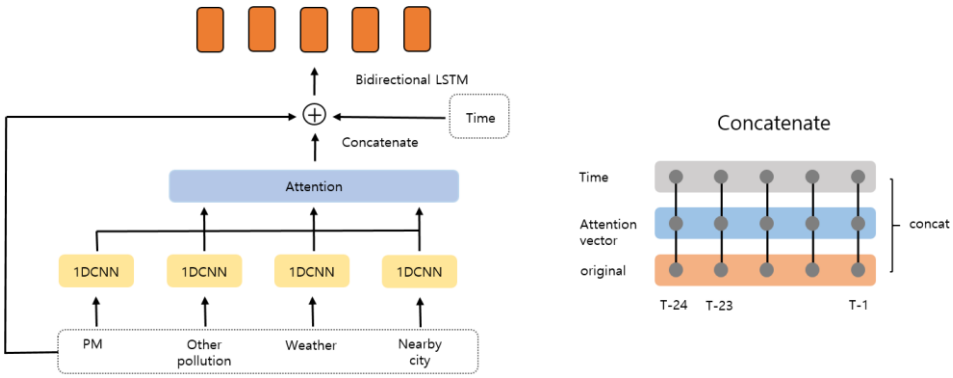

Fig. 1: Overall structure of our model

+

3.2. Overview

In our model, we preprocess X^a, X^w, X^cdata into 4 groups; particulate matters D^{pm}, and other pollutants D^{op} (air pollution except particulate matters), weather D^w, nearby city D^c. Our model has 3-stage modules including 1DCNN(1-Dimensional Convolution), the attention mechanism and LSTM for temporal decoding. The 1DCNN module extracts embedding (E^{pm}, E^{op}...) from each data groups(D^{pm}, D^{op}...) using multiple 1DCNN layers with residual connections. By doing so, the 1DCNN module captures intra-correlations among features in each data source. After that, the attention module gets embeddings of data groups from the 1DCNN module and applies attention between the embeddings E^{pm} (PM data) and other data. In contrast to the 1DCNN module, attention module captures inter-correlation among features that belong to other data sources. Lastly, our model uses Bidirectional LSTM that take output of attention module as input. Bidirectional LSTM is used to capture temporal correlation between all features and output prediction.

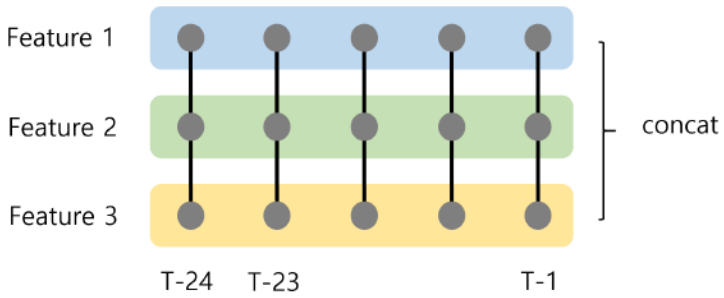

Fig. 2: Preprocessed data structure of our model

3.3. Preprocessing

As shown in figure 2, every data sources (particulate matters, other pollution, weather, nearby city) has multiple features over time. For pm(particulate matters), we use pm data from target station and neighboring stations. For brevity, we select three neighbor stations close to the target station. As for weather data, we select the nearest weather station according to each target station.

3.4. 1-Dimensional Convolution

We use 1-Dimensional Convolution layers to produce embeddings that capture the spatio-temporal relationships within each data source. Since each data source has different features and distributions, we learn a different 1DCNN layer for each data source. In this way, the model captures intra-correlation among features while considering distribution of each data source. 1D (1 Dimensional) CNN [4] is used because time series data are one-dimensional rather than two-dimensional data such as

image. In addition, deeply stacked 1d CNN layers have large receptive fields, which can consider long range of data.

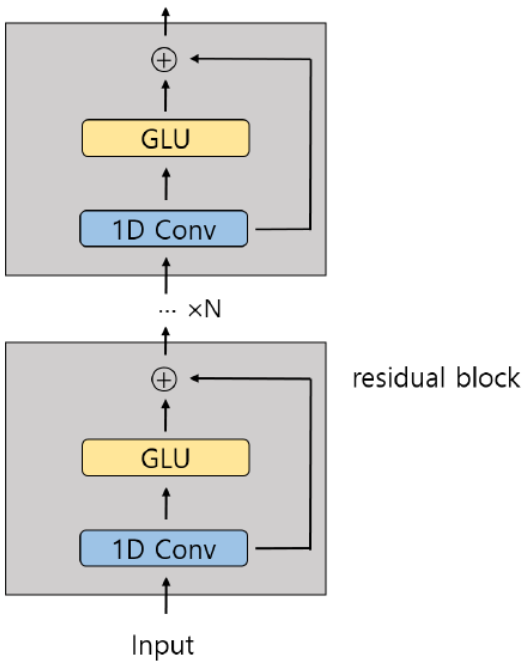

Fig. 3: 1DCNN structure of our model

As shown in Figure 3, the 1DCNN module consists of a stack of residual block. Each residual block consists of 1D convolution layer and GLU(Gated Linear Unit) [14] with residual connection. In general, GLU computes pointwise product with original inputs and inputs after applying sigmoid units. For an input x and a given 1DCNN operation g, the residual block's output is $\mathrm{GLU}(\mathrm{g}(\mathrm{x})) + x$. Total number of residual block is 6. The kernel size of 1D convolution layer is 3 and the number of filter is 128.

3.5. Co-Attention among different data sources

In our model, the attention mechanism is used to capture inter-correlation among different data sources. Using 1DCNN layers, our model gets embeddings E^i from each data source. The attention module generates two types of output embeddings: E^{pm} from particulate matters and E^i from other data sources. We use Loung [10] attention for our attention module.

At time t, alignment vector a_t^i is derived by the embedding of particulate matters E_t^{pm} and the embedding of source i E_t^i. We used dot product for score function.

$$a_t^i = \frac{\exp(score(\mathrm{E}_t^i, \mathrm{E}_t^{pm}))}{\sum_t \exp(score(\mathrm{E}_t^i, \mathrm{E}_t^{pm}))}$$

Alignment vector represents relative importance of E_t^i according to E_t^{pm}. After get alignment vector from data source i at each time t, we get context vector by product of E_t^{pm} and a^i

$$Attention(E_t^i, E_t^{pm}) = a_t^i * E_t^i$$

In this way, the attention vector can understand interaction among particulate matters and features from other data sources.

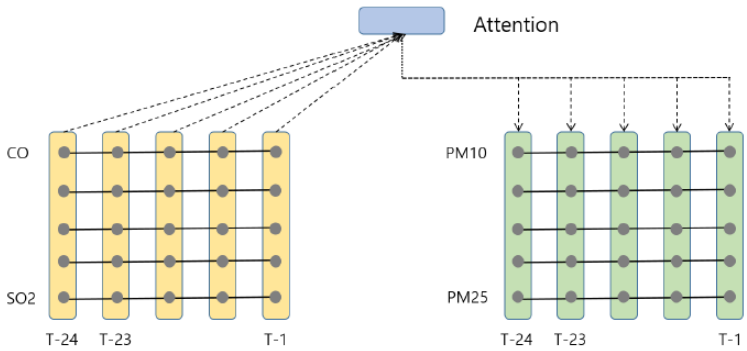

Fig. 4: Attention of our model

3.6. Output Layer

As shown in Figure 1, our model concatenates outputs of attention module and passes them to bi-directional LSTM (Long Short Term Memory) [9]. The attention module outputs three attention vectors, each vector represents the attention between particulate matters and other embeddings. Our model also takes advantage of time information (day of week, hour) with other data. Time information are concatenated with original data and three attention vectors before LSTM. After LSTM module, our model passes the output of the LSTM to a fully connected neural network. The fully connected neural network has T units and its output becomes our prediction for 1~T hours from the start time.

Chapter 4. Experiment

4.1. Data

We use Seoul air pollution data, Seoul meteorological data, China data, Time data for training and testing.

Seoul Air Pollution Data

Seoul air pollution data are collected from National Institute of Environmental Research, department of South Korea Government. Seoul air pollution data were measured every hour at 38 monitoring stations, including 25 distinct measuring stations and 13 road measuring stations in Seoul, from 2015 to 2017. Pollutants observed at the monitoring stations are PM10, PM2.5, O3, SO2, NO and CO. In the experiment, we used data from 25 distinct stations except 13 road stations. Seoul air pollution data have information of each station, including longitude and latitude.

Seoul Meteorological Data

Seoul air pollution data are collected from Korea Meteorological Administration. Seoul meteorological data consist of temperature, precipitation, wind speed, wind direction and humidity observed every hour at 29 monitoring stations in Seoul from 2015 to 2017. Like air pollution data, meteorological data also have information of each station, including longitude and latitude.

Chinese Cities' Air Pollution Data

We also use China air pollution data to predict Seoul particulate matters density. We gathered China air pollution data at Beijing, Shandong, Shenyang from US Embassy Website. Each city have only PM2.5 data every 1 hour at one monitoring station from 2015 to 2017. We use China data as the nearby city data because China's air pollution affects Seoul air pollution through westerlies.

Type	Feature	Type
Air Pollution Data	PM10 PM2.5 O3 SO2 NO CO	Float
Meteorological Data	Temperature Precipitation Wind Speed Wind Direction	Float Float Float Discrete
China Data	PM2.5 Beijing PM2.5 Shandog PM2.5 Shenyang	Float
Time Data	Day of Week Hour	Discrete Discrete

Table 1: Data Description

We use hour and day of week for time data. We set the threshold for each feature and set the value above the threshold value to the threshold value. For all features, the value below 0 is set to 0. We use one-hot encoding to transform discrete feature. Then we normalize data using min-max scale.

4.2. Details and Evaluation Metric

We use hour and day of week for time data. We set the threshold for each feature and set the value above the threshold value to the threshold value. For all features, the value below 0 is set to 0. We use one-hot encoding to transform discrete feature. Then we normalize data using min-max scale.

We compare our model with baselines using MAE(Mean Absolute Error). MAE is widely used for Evaluation Metric in air pollution prediction. MAE is defined as follows

$$MAE = \sum_t |y_t - \hat{y}_t|$$

Where y_t and $\hat{y}_t$ mean the prediction value and real value of t timestamp.

4.3. Comparison with previous approaches

In this section, we evaluate our model with previous methods to show effectiveness of our model. We compare our model with four models as follows

Seq2Seq[11] is an Encoder-Decoder model that uses two LSTMs to act as encoder and decoder.

DARNN[15] is a time series model using a dual stage attention that shows good performance on stock price prediction.

DCRNN[8] is a model that captures spatial dependency and temporal de-

pendency by using the Graph convolutional natural network and sequence to sequence model.

Deepair[13] is a model that considers multiple domain data by fusing the acquired data through DNN.

PM10 (MAE)	1∼3h	4∼6h	7∼9h	10∼12h	13∼15h	16∼18h	19∼21h	22∼24h
DCRNN	7.97	11.21	13.33	14.76	15.69	16.42	17.00	17.49
DARNN	7.53	11.38	13.44	14.78	15.67	16.35	16.89	17.26
Seq2Seq	**6.17**	10.17	12.55	14.06	15.13	15.91	16.55	16.99
Deepair	6.85	10.29	12.32	13.51	14.33	14.89	15.18	15.46
our model	6.18	**9.58**	**11.54**	**12.44**	**13.05**	**13.51**	**13.90**	**14.20**

Table 2: Performance comparison on PM10 with other models and our model

PM2.5 (MAE)	1∼3h	4∼6h	7∼9h	10∼12h	13∼15h	16∼18h	19∼21h	22∼24h
DCRNN	4.38	6.57	8.02	8.95	9.66	10.24	10.60	10.83
DARNN	4.85	6.94	8.09	8.82	9.37	9.80	10.08	**10.30**
Seq2Seq	**4.11**	6.65	8.16	9.20	9.91	10.52	10.96	11.28
Deepair	4.16	6.61	7.99	8.92	9.59	9.96	10.18	10.52
our model	4.43	**6.58**	**7.83**	**8.59**	**9.25**	**9.66**	**9.98**	10.35

Table 3: Performance comparison on PM2.5 with other models and our model

Table 2 and Table 3 shows the result of previous approaches and our model at predicting concentration of PM10, PM2.5. We predicted concentrations from 1 to 24 hours and averaged every 3 hours to compare models. Our model shows better result than baselines at most hour. At PM10, our model shows better performance at every time period except 1~3 hour, and at PM2.5 our model shows better performance except 1~3, 22~24 hour. Especially after 3 hour, our model outperforms all previous approaches except 22~24 hour, which shows that our model has strength on long term prediction.

4.4. Ablation study

We conduct ablation study on modules to investigate effects of 1DCNN and attention. First, we compare performance of our model with our model without 1DCNN and attention modules. Then, we conduct ablation study while changing LSTM modules to 1DCNN and FNN(Fully Connected Neural Network). In this way, we verify effects of modules while removing impact of LSTM module. When implementing CNN with modules, we pass the output of modules to 1DCNN. Then, final fully connected layer gets the output of 1DCNN and makes prediction. When implementing FNN with modules, we just pass the output of modules to fully connected layer.

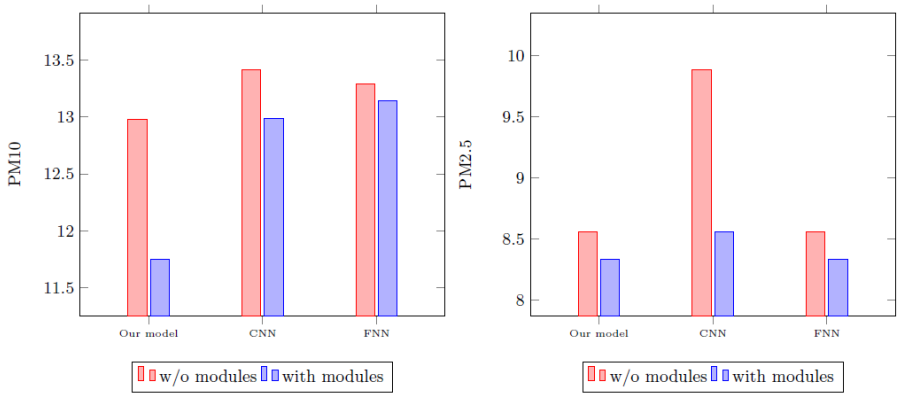

Fig. 5: Performance comparison w/o modules on PM10 and PM25

Figure 5 shows performance comparison w/o modules at our model, CNN and FNN. Using modules shows better performance in our model, CNN, FNN with 9.5%, 3%, 1% relatively low MAE showing that using modules contribute towards the model's performance regardless of LSTM modules

4.5. Performance on data selection

To evaluate our model's capability of capturing interaction among multi model features, we explore the performance of our model with different data combination. We remove nearby city and weather data step by step from our model and also remove 1DCNN layers and attention module that handling nearby city and weather data.

PM10	Average MAE
our model	11.79984
our model w/o nearby city data	11.96841
our model w/o nearby city, weather data	12.75586
PM2.5	Average MAE
our model	8.334361
our model w/o nearby city data	8.437566
our model w/o nearby city, weather data	8.779258

Table 4: Performance comparison of data selection

Table 4 shows the result of excluding nearby city and weather data on PM10 and PM2.5. We compare average MAE of each data combination from 1 to 24 hours. MAE of our model gradually increased as nearby city and weather data removed. This result shows that multi model features improve performance and our architecture has capability to learn these features.

4.6. Performance on residual connection

In this section, we compare our results while changing the number of residual block. We change the number of residual block from 2 to 5. Our assumption is that more residual blocks would improve the performance by making the receptive field larger.

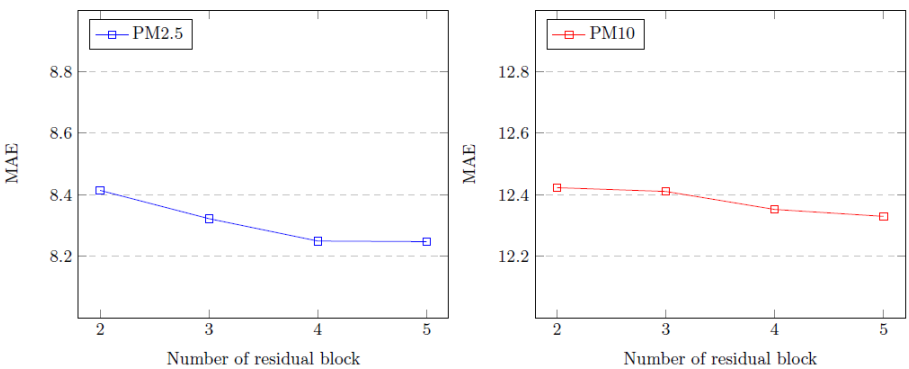

Fig. 6: Performance with varying the number of residual block

Fig. 6 shows average MAE from 1 to 24 hour of our model according to the number of residual block. As the residual number increases, average MAE value decreases. This result suggests that larger receptive fields helps model to learn long time data effectively.

Chapter 5. Conclusion

In this paper, we proposed new deep learning model for predicting air pollution density. Our model use 3 modules including 1DCNN, Attention and LSTM. First, our model use 1DCNN module to capture inter correlation within data. And then, our model use Attention module to capture inter correlation within data from different data sources. By doing so, our model can learn high level correlation among data from different data sources. In addition, with 1DCNN module and Attention module, our model can capture spatial temporal correlation among features from different stations.

In experiment, we train our model using Seoul air pollution data, Seoul weather data and China air pollution data. We compare our model with previous state-of-the-art method. Experiments show that our model outperforms all previous state-of-the-art methods. Furthermore, ablation study shows that our 1DCNN module and Attention module contribute the model's performance. In addition, experiment with data selection shows that our model has capability to learn multi model features. Finally, experiment with residual connection shows that increasing residual connection improves performance.

Bibliography

1. Abadi, M.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Savannah, USA (2016)

2. Felix, W., Angela, F., Alexei, B., Yann, D., Michael, A.: Pay less attention with lightweight and dynamic convolutions. In: International Conference for Learning Representations. New Orleans, USA (2019)

3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

4. Hochreiter, S.: Long short-term memory. Neural Computation pp. 1735–1780 (1997)

5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. Lille, France (2015)

6. Junbo, Z., Yu, Z., Dekang, Q.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2017)

7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations. San Diego, USA (2015)

8. Li, Y.: Graph convolutional recurrent neural network: Data-driven traffic forecasting. In: International Conference for Learning Representations. Vancouver, Canada (2018)

9. Mike, S., Kuldip, P.: Bidirectional recurrent neural networks. IEEE TRANSACTIONS ON SIGNAL PROCESSING pp. 2673–2681 (1997)
10. Minh-Thang, L.: Effective approaches to attention-based neural machine translation. In: EMNLP. Lisbon, Portugal (2015)
11. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing System. MIT Press, Montreal, Canada (2014)
12. Weiyu, C., Yanyan, S., Yanmin, Z., Linpeng, H.: A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2018)
13. Xiuwen, Y.: Deep distributed fusion network for air quality prediction. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 19–23. ACM Press, London, United Kingdom (2018)
14. Yann, D.: Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia (2017)
15. Yao, Q.: A dual-stage attention-based recurrent neural network for time series prediction. In: Proceedings of International Joint Conference on Artificial Intelligence. pp. 2627–2633. AAAI Press, Melbourne, Australia (2017)
16. Yuxuan, L., Songyu, K., Junbo, Z., Xiuwen, Y., Yu, Z.: Deep residual learning for image recognition. In: Proceedings of the International Joint Conference on Artificial Intelligence (2018)

17. Yuxuan, W.: Tacotron: Towards end-to-end speech synthesis. In: Interspeech. Stockholm, Sweden (2017)
18. Zheng, Y.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2267–2276. ACM Press, Sydney, Australia (2015)

초 록

대기 오염, 특히 미세먼지는 많은 국가에서 심각한 문제가 되었다. 이러한 급격한 오염에 대처하기 위해 딥러닝을 사용하여 대기 오염 물질의 농도를 예측하는 몇 가지 연구가 진행되어왔다. 그러나 미세먼지 농도는 다양한 유형의 멀티 모달 정보가 오랜 시간 동안 복잡하게 연관되어 있기 때문에 딥러닝을 이용하여 미세먼지 농도를 정확하게 예측하는 데 어려움이 존재했다. 본 논문에서는 잔존 연결 및 어탠션 메커니즘과 함께 깊이 쌓인 1 차원 컨볼루션을 이용하여 미세먼지 농도를 예측하는 새로운 아키텍처를 제안한다. 1 차원 컨볼루션은 고차원의 특징 벡터를 추출하며 어탠션 메커니즘은 추출한 고차원의 특징 벡터들간의 상관 관계를 포착한다. 서울 미세먼지 농도 데이터를 이용한 실험을 통해 PM2.5 및 PM10 예측에서 본 논문에서 제시한 모델이 기존 모델을 뛰어넘었다는 것을 확인했다.

주요어 : 미세먼지, 컨볼루션, 어탠션

학 번 : 2018-25751