Multimodal Deep Learning for Transboundary Haze Prediction

Phuc-Thinh Nguyen[1], Nazmudeen Mohamed Saleem[2]
[1]University of Information Technology, Vietnam
[2] University Teknologi Brunei

ABSTRACT

Environmental pollution, particularly air pollution, has long been an issue in every major city on the planet. For many years, accurate estimation of PM2.5 and PM10[8] [12] fine dust concentration values has been a fascinating study area. This study focused on 3-Day Transboundary Air Pollution Prediction, which proposed to merge many deep learning models and pick appropriate properties for the PM10 index prediction problem by utilizing various features such as timestamps, geographical features, and public weather data. Using the dataset provided by MediaEval, we examined the performance of several learning models and features in order to investigate the problem. Experimental results show that combining multiple deep learning models together gives a higher overall performance than other techniques and features in RMSE, MAE, SMAPE.

Keywords—PM10 prediction, LSTM, BiLSTM, multimodal

1 INTRODUCTION

Haze air pollution is defined as the presence in the air of particulate matter such as smoke, dust, and other vapours that arise from the large-scale forest and land fires, factories, and automobiles. When the concentration of airborne pollutants reaches dangerous levels, it causes respiratory issues and has significant consequences for visibility, economic productivity, transportation, and tourism.[7]

Transboundary haze is a recurring problem in many parts of the world, particularly in Southeast Asia, where haze pollution sources differ from nation to country, with varying percentages coming from localized or transboundary sources.[7]

The goal of this article is to address the sub-task 2 of the competition which is to examine the transnational PM10 estimate problem using timestamp information, location data, and weather data using the technique of mixing multiple deep learning models.[1][11]

2 METHODOLOGY

We'll start with a high-level overview of the topic and then go over to our specific approach that we have developed for a PM10 prediction algorithm.

2.1 Pre-processing data

At this stage the data for each station is separated from the original data (train air quality and train weather) based on the station's ID. Then we will combine the data by province as the data on air quality and weather can be found in the same province.

We'll utilize the interpolation method[6] for missing values in the data. In addition, we fill in the mean for the variables that cannot be interpolated.

With testing data (from 2018-2019), we'll utilize the zero-fill method to generate a mask to help the model work with missing values of data that is used for prediction.

2.2 Features Extraction

We model the PM10 index prediction problem as a regression problem with the following features to estimate the PM10 index in the near future from a list of specified attributes.

2.2.1 Timestamp features. Because outdoor air quality varies greatly depending on the time of day, timestamp information can be valuable for PM10 estimation difficulties. In particular, each country's PM10 index is unique, as are the provinces within the same country. As a result, we created a PM10 index for a country by averaging the PM10 scores of the provinces. We take the daily average and then average it across provinces in provinces where PM10 is reported hourly.

2.2.2 Location features. We choose one province as a landmark, and then we evaluate the closest distances between provinces in the vicinity of the landmark as it can provide useful information for plot analysis and reduce noise due to air pollution in the data.

We calculated the distance using the Haversine formula[9], which is an integral equation for navigation that yields precise results for calculating the great circle distance between two points on the Earth's surface based on their latitudes. The Haversine formula can be calculated using two positions, A and B:

$$d(A,B) = 2.r.arcsin\Big(sin^2\Big(\frac{\varphi_B - \varphi_A}{2}\Big)+ cos(\varphi_A).cos(\varphi_B).sin^2\Big(\frac{\lambda_B - \lambda_A}{2}\Big)\Big)^{\frac{1}{2}}, \quad (1)$$

where r is the Earth's radius, and $\varphi_A, \lambda_A, \varphi_B, \lambda_B$ are the latitudes and longitudes of two points A and B, respectively.

2.2.3 Weather features. Public weather features include information on weather, such as "temperature," "precipitation," "humidity," "wind direction," and "wind speed," obtained from local stations. These characteristics can be thought of as supplementary data that can help machine learning models become more robust and reliable.

In order to provide the best forecast results, we analyze the correlation between these weather features and the PM10 index in each country and select the features with the strongest correlation. The wind direction had a strong correlation with PM10 in the Indonesia dataset; the temperature, rain, and humidity in the Brunei dataset; the rain, humidity, and wind speed in the Thailand dataset; and the temperature in the Singapore dataset.

If some of the provinces are lacking "Temperature" data We utilize a basic LSTM[2] approach to predict the "Temperature" using weather data from 2018-2019 (testing data). .

MediaEval'21, December 13-15 2021, Online

2.3 Training and Testing Setup

To train a model that is appropriate for the problem, we split the raw training dataset (2010-2017) into training and test datasets, using 80% of them for training and the remaining 20% for model testing. After that, we will predict PM10 on testing dataset (2018-2019). This study uses a regression model to forecast PM10 for the next three days.

2.4 Multimodal Method

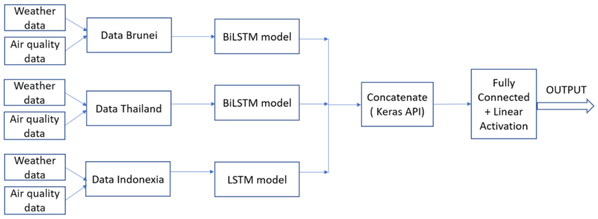


Figure 2: Diagram of combining multiple deep learning models

Process the data for each station in section 2.1, then build a dataset for each country by averaging the weather and air quality features of the stations in that country; for example, data of Brunei would be obtained by averaging the weather and air quality features of the stations in Brunei.

For each country, we use data from the previous three days to forecast the future three days.

We employ the concept of merging different models[5][3] to increase the performance of the PM10 estimator. The fundamental idea is to connect the outputs of each country's PM10 prediction deep learning model, then feed that information into a final deep learning model to get the final PM10 result.

We'll need three branches to construct our multi-input network: The first two forks will be a simple BiLSTM[2] that will handle Brunei and Thailand data repectively. A simple LSTM will handle Indonesian data inputs in the last fork. Finally, concatenate these branches to produce the final multi-input deep learning model. It's random; you're free to use Indonesian data in the BiLSTM model. We replaced it throughout the experiment, and the MAE and RMSE results are nearly similar.

We must fill in the mean from 2010 to 2015 because Singapore weather data is only available from 2016 to 2017, hence we do not recommend using Singapore data to train the multimodal model to avoid overfitting the model. We'll run the BiLSTM model for Singapore data separately.

2.5 Performance metrics

We employ root mean square error (RMSE), mean absolute error (MAE)[10], and Symmetric mean absolute percentage error (SMAPE)[4] to evaluate the performance of the proposed approaches.

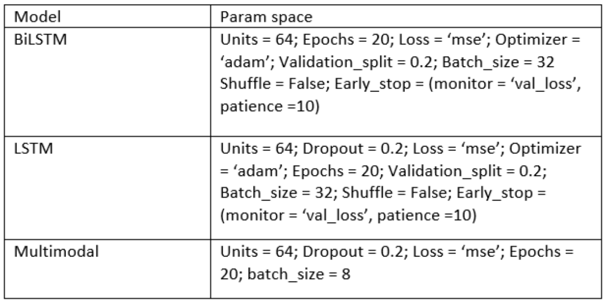
3 EXPERIMENTS

3.1 Data sets

The organizer has provided us with a data set. Weather and air quality data are included in the dataset, with the training set spanning 2010-2017 and the test set spanning 2018-2019.

3.2 Model settings

To improve the performance of the presented approaches, we use a random search method to select ideal hyperparameters based on performance indicators. For each model, this method entails scanning a predefined parameter space and picking the best performing hyperparameters which is shown in Table 1.



Model	Param space
BiLSTM	Units = 64; Epochs = 20; Loss = 'mse'; Optimizer = 'adam'; Validation_split = 0.2; Batch_size = 32 Shuffle = False; Early_stop = (monitor = 'val_loss', patience =10)
LSTM	Units = 64; Dropout = 0.2; Loss = 'mse'; Optimizer = 'adam'; Epochs = 20; Validation_split = 0.2; Batch_size = 32; Shuffle = False; Early_stop = (monitor = 'val_loss', patience =10)
Multimodal	Units = 64; Dropout = 0.2; Loss = 'mse'; Epochs = 20; batch_size = 8

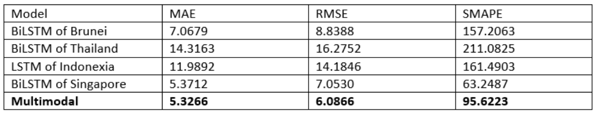
Table 1: parameters of the models

3.3 Results

In this study, we compare the proposed methodologies' performance to the performance measures listed in Section 2.5.

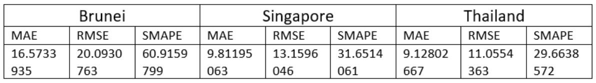
The MAE and RMSE results of multimodal are better than the rest of the models, as shown in Table 2. However, because multimodal employs the average PM10 value of countries, the forecast results will be skewed for countries with excessively low or excessively high PM10.

The MAE and RMSE values in Table 2 are better than those in Table 3, but the SMAPE results are worse.



Model	MAE	RMSE	SMAPE
BiLSTM of Brunei	7.0679	8.8388	157.2063
BiLSTM of Thailand	14.3163	16.2752	211.0825
LSTM of Indonexia	11.9892	14.1846	161.4903
BiLSTM of Singapore	5.3712	7.0530	63.2487
Multimodal	**5.3266**	**6.0866**	**95.6223**

Table 2: The results of MAE, RMSE SMAPE of the models



Brunei			Singapore			Thailand		
MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
16.5733935	20.0930763	60.9159799	9.81195063	13.1596046	31.6514061	9.12802667	11.0554363	29.6638572

Table 3: The results of MAE, RMSE SMAPE of organization

4 CONCLUSIONS AND FUTURE WORKS

After three months for analyzing, we illustrated the benefits of combining numerous models with generalization and deep learning methods to address the PM10 index estimation problem utilizing various types of features such as timestamps, location, and public weather data. By incorporating a variety of characteristics, the test results reveal that PM10 level prediction is fairly accurate when compared to ground-truth. Transnational air pollution can be predicted using the strategy of merging multiple models.

We're excited to continue our research by examining additional forms of data, such as image data, video, and new deep learning models. One of the new approachs is multivariate transformer learning because this learning can learn across long timespans.

REFERENCES

[1] AR Varkonyi-Koczy. A Mosavi, S Ardabili. 2019. List of Deep Learning Models. *International Conference on Global Research and Education* (2019).

[2] Jason Brownlee. 2018. *Deep Learning for Time Series Forecasting*.

[3] Quang M.; Nguyen-Tai Tan-Loc; Bo Dong; Nguyen Dat; Dao Minh-Son; Nguyen Binh T. Duong, Dat Q.; Le. 2020. Multi-source Machine Learning for AQI Estimation. 5 (2020).

[4] Wida Susanty Haji Suhailia-Peijiang Zhaob. Effa Nabilla Aziza, Asem Kasema. 2021. Convolution Recurrent Neural Network for Daily Forecast of PM10 Concentrations in Brunei Darussalam. *AIDIC* 13 (2021).

[5] D Roggen. FJ Ordóñez. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *mdpi* 5 (2016).

[6] Y. T. GU. G. R. LIU. 2001. A POINT INTERPOLATION METHOD FOR TWO-DIMESIONAL SOLIDS. *INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING* (2001).

[7] Asem Kasem, Minh-Son Dao, Effa Nabilla Aziz, Duc-Tien Dang-Nguyen, Cathal Gurrin , Minh-Triet Tran, Thanh-Binh Nguyen, and Wida Suhaili. Overview of MediaEval 2021: Insights for Wellbeing Task Cross-Data Analytics for Transboundary Haze Prediction.

[8] Pingqing Fu Xiangdong Li Ling Jin, Xiaosan Luo. 2016. Airborne particulate matter pollution in urban China: a chemical mixture perspective from sources to impacts. *National Science Review* 4 (2016), 593–610.

[9] S Suryati Widdha Mellyssa. M Basyir, M Nasir. 2017. Determination of Nearest Emergency Service Office using Haversine Formula Based on Android Platform. *EMITTER* 5 (2017).

[10] H Smith NR Draper. 1998. *Applied regression analysis*.

[11] G Hinton. Y LeCun, Y Bengio. 2015. Deep learning. *Nature* (2015).

[12] Min-Hua Shi-Yi-Xin Lian. Yu-Fei Xing, Yue-Hua Xu. 2016. The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease* (Jan 2016).