CNN-GRU for Air Quality Index Forecasting

Van-Tien Nguyen[1,2,*], Hai-Dang Nguyen[1,2] and Minh-Triet Tran[1,2,3]

[1]University of Science, VNU-HCM, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]John von Neumann Institute, VNU-HCM, Vietnam

Abstract

Recently, air pollutant becomes an urgent problem, especially in urban areas; therefore, a system to predict future air properties is demanded to improve the quality of life. By proposing a neural network using weather and air data to perform the Air Quality Index forecasting task, the model shows a reasonable performance instead of traditional regression methods. Moreover, the dependency of the number of input days, and the hour prediction accuracy are also discussed.

1. Introduction

In developing cities, air quality has become a preoccupation for citizens' health. Governments need to find a way to forecast air conditions on available meteorological data in order to prepare a solution to deal with changes that can appear in the future.

Air Quality Index (AQI) is a scale used by governments for reporting air quality [1]. Air quality sensors are utilized to collect values of air factors (e.g. temperature, humidity, wind average, etc.). There are six main pollutants which are fine inhalable particles PM2.5, PM10, Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), and Ozone (O_3). Based on the AQI level, the community can understand how polluted the air currently is or how polluted it is forecasted to become.

Realizing the necessity of an AQI prediction, Urban Air: Urban Life and Air Pollution [2] introduces the UrbanAir task that provides a streaming dataset from CCTV and air stations network installed in Dalat City, Vietnam. There are two subtasks of AQI prediction. Subtask 1 which is using only the air station data to predict hour-average values and AQI levels of pollutants is the main focus of this work.

In this paper, there are four main contributions:

- Propose a CNN-GRU model to predict the hour-average value of pollutants and AQI from the raw value of weather and pollutants.
- Show the dependency between the number of input days and the prediction performance.
- Show the effective of prediction for each hour in a day.

2. Related Work

In [3], authors use three regression models as feature extractors using preprocessed sensor data, extract two new features Part-Of-Day (cluster 24 hours of a day into 5 groups) and Is-Rush-

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
†These authors contributed equally.
nvtien19@apcs.fitus.edu.vn (V. Nguyen); nhdang@selab.hcmus.edu.vn (H. Nguyen); tmtriet@fit.hcmus.edu.vn (M. Tran)

Hour (identify a Part-Of-Day group is rush hour or not), and employ a stack generalization technique to combine multiple regressions' results into the final output.

In [4], Convolution Neural Network (CNN) [5] is used as a feature extractor. A modified Long Short-term Memory (LSTM) [6] called ILSTM removes the output gate to improve the hidden gate and input gate. A CIM gate is introduced to prevent saturation during training.

In this work, instead of learning from raw and noisy data, Convolution Neural Network (CNN) [5] is proposed to learn rich features. By performing 1D convolutions and Rectified Linear Unit (ReLU) [7], high dimensional features are constructed by linear and non-linear operations on nearby hour values and other pollutants values. Besides, Gated Recurrent Unit (GRU) [8] is used as an auto-regression model to avoid the "gradient explosion" and "gradient vanishing" problems in RNN [9], and reduce the number of parameters in LSTM [6] architecture.

3. Approach

3.1. Data preprocessing

The sensors are facing many errors during running time. The strategy for data preprocessing includes three steps. Firstly, handle missing and wrong behavior data records. After that, resample by hour to calculate the 1-hour-average value, and needed hour-average value for each pollutant (e.g. 24-hour-average value for PM2.5 and PM10). Then AQI, AQI level, and the final AQI with the responsible pollutant. The final total dataset is interpolated by 24 hours and removed invalid records.

3.2. Datasets

The raw data is crawled from a real-time server. After data preprocessing, subsets of data (train, validation, and test) are created strategically to optimize the model's hour-average value of pollutants. Besides, the target set is constructed for Subtask 1 evaluation metrics, and the train value set is used for pre-training the model before the train set.

Firstly, all records (resample by hour) which have not NaN value are selected as valid datetime records.

The target set's construction is based on the Subtask 1 metric mentioned in Section 3.4. A column '0d' represents the last input date (D). There are three datetime columns namely '1d', '5d', '7d' which are short- (D+1), mid- (D+5), and long-term (D+7) periods respectively. These three columns are selected from the valid records with datetime starting from '2022-11-01 00:00:00'. Related sensor id is also noted. The final column contains the number of possible input days which contains 24-hour records.

The test set's time period is the same as the target set; however, all the next 7 days must be valid instead of only the next 1, 5, and 7 days. So that, columns '1d', '5d', and '7d' are not stored.

For the train and validation set, initialization is similar to the test set, but the time period is before November. The proportion of the train and validation set is 9:1 because of the small number of valid 7-day periods.

For pre-training the model, the train value set contains all record which has available hour-average value of pollutants for the next 7 days of the '0d' column (the last input date). For explanation, the main contributor to model performance is the hour-average value of air factors so only these value is required for the pre-training step. The train value set's datetime is not included in the validation set.

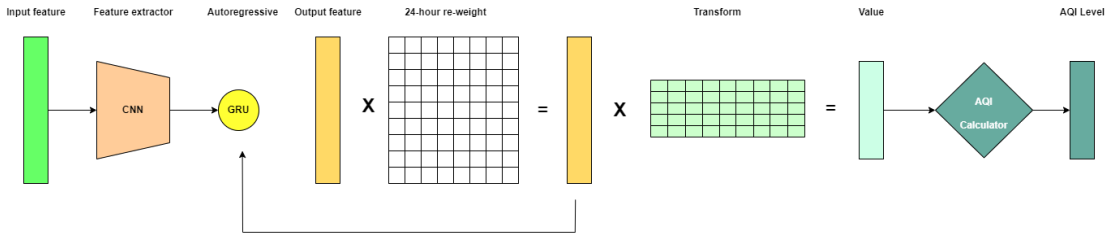


Figure 1: Model architecture. The input features are (24*input_day) hour-average value. The final output features are a concatenation of 7-day output features.

3.3. Methods

3.4. Metrics

There are two main metrics used for the evaluation of Subtask 1 [2] for 6 pollutants (PM2.5, PM10, CO, NO2, SO2, O3): (1) MSE/MAE for the hour-average value of each pollutant. (2) F1-score for AQI for each pollutant.

The prediction is evaluated in short-, mid- and long-term periods which are 1, 5, and 7 days in the future respectively.

For training hour-average values, the only used metric is MSE.

3.5. Model architecture

There are 3 main components in the architecture: a feature extractor, a recurrent neural network, and an AQI calculator (Figure 1).

Feature extractor: A convolution neural network [5] includes 4 layers using 1-D filters to learn the relative hour-average value from raw hour-average values. The BatchNorm and ReLU layers are used after each 1D convolution.

Autoregressive model: Gated recurrent unit [8] is used to handle the time-series features from extractor. The input and output of GRU represent 24-hour features. Each output feature is re-weighted by a 24x24 matrice, then feed into the GRU model until 7-day output features are generated.

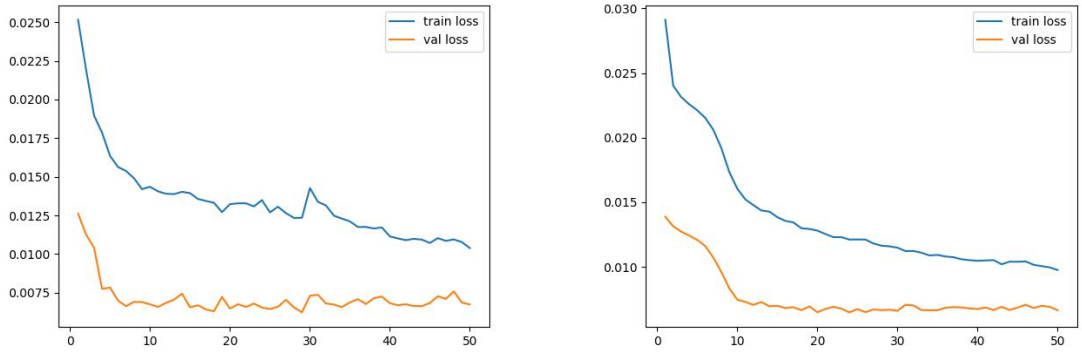
To achieve the pollutants' value, each 24-hour feature which is the output of the GRU is transformed into the value vector representing the hour-value of pollutants.

AQI calculator use the input and output hour-average values to calculate the AQI value by the Equation 1 which is mentioned in Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI) [1].

$$I_P = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + I_{Lo} \tag{1}$$

Where I_P: the index of pollutant p
C_P: the truncated concentration of pollutant p
BP_{Hi}: the concentration breakpoint that is greater than or equal to C_P
BP_{Lo}: the concentration breakpoint that is less than or equal to C_P
I_{Hi}: the AQI value corresponding to BP_{Hi}
I_{Lo}: the AQI value corresponding to BP_{Lo}

[1] https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf



(a) leaning rate = 0.01 (b) leaning rate = 0.001

Figure 2: Pre-training model by train value set in 50 epochs. A high learning rate is chosen for speed.

Set	Normalized		Raw		
	MSE	MAE	MSE	MAE	F1-score
Validation	0.0131	0.009	15981.85	6.32	0.590
Test	0.024	0.017	47645.08	14.70	0.485
Target	0.007	0.004	10486.00	3.30	0.614

Table 1
Final results which are evaluated in three sets: validation, test, and target.

4. Experiments and Results

4.1. Baseline

For the pre-training phase, the configuration includes 1 input day, Adam optimization with a learning rate of 0.01. The training result is plotted in Figure 2. The best validation loss is 0.00624 (normalized) achieved at epoch 13, then the model enters the saturation period. By reducing the learning rate to 0.001, the model converges slower and smoother with a minimal validation loss is 0.00648. The pre-trained model is tested on the test set, and the results are 0.0638 (normalized).

The final model is also evaluated by validation, test, and target sets (Table 1). Although the test set is smaller than the validation set, the performance is reduced.

4.2. Number of input days

The model performance in the train value set is tested from 1 to 10 input days (Figure 3). With 1 input day, the loss is always minimized the best. Furthermore, the model's convergence abilities of 2, 3, and 4 input days are the same, nearly as 1 day. There is a noticeable performance when feeding for 9 days.

4.3. Prediction hour in a day

By plotting the MSE loss of 24 hours per day in 7 prediction days in the train set (Figure 4), there is a common trend in the hour accuracy. In all 7 days, the period between 10 a.m. and 8 p.m. keeps the smallest loss. The MSE loss on days 1, 5, and 7 share the same shape.

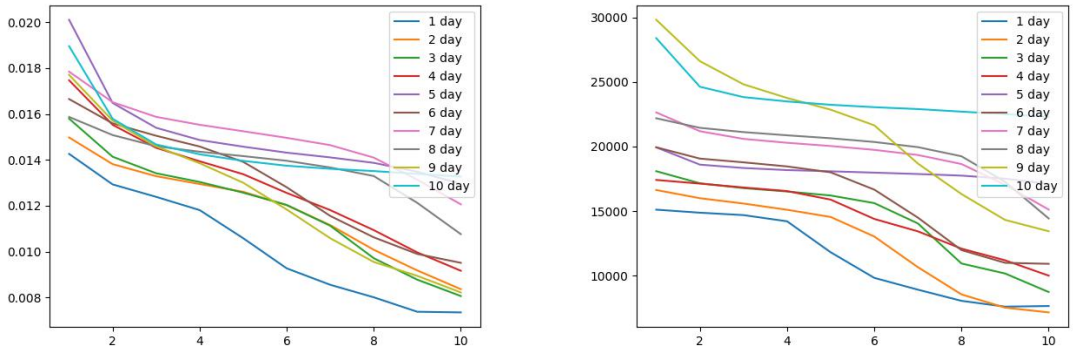


Figure 3: The dependency of the number of input days and model performance. (a) The normalized value for 10 epochs (b) The raw value for 10 epochs.

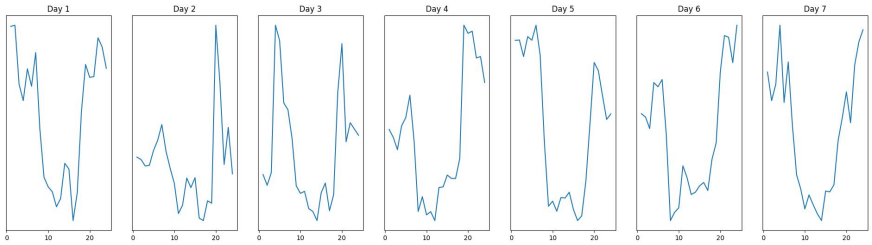


Figure 4: The normalized MSE loss of 24 hours per day in 7 predicted days in the train set.

4.4. Submission result

Table 2 show the submission results from the CNN-GRU model for Subtask 1 of UrbanAir. The model receives only 1 input day.

In the prediction period, the PM factors are not the main reasons for pollution. Besides, because the recorded value of SO_2 is dominated by 0 and 1, the predicted AQI for SO_2 is biased. However, in sensor 8, the model shows that the hour-average values can be more specific.

5. Conclusion

Air Quality Index (AQI) forecasting is crucial for improving the quality of smart cities. In this work, a CNN-GRU model is optimized to predict the hour-average value of six pollutants by feeding raw input data records. The achieved results are noticeable to improve better with data preprocessing methods that reduce noisy data, outlines, and sensor problems, solving the saturation problem.

References

[1] AirNow, Air quality index (aqi) basics, https://www.airnow.gov/aqi/aqi-basics/, 2022. Accessed: 2022-11-28.

sensor id	pm2.5	pm10	co	no2	so2	o3	pm2.5 AQI	pm10 AQI	co AQI	no2 AQI	so2 AQI	o3 AQI
1	5.145	6.456	1.201	0.029	1.000	0.094	0	0	0	0	5	3
	5.574	5.957	1.200	0.028	1.000	0.094	0	0	0	0	5	3
	7.143	7.679	1.201	0.026	1.000	0.094	0	0	0	0	5	3
4	1.288	1.541	1.484	0.282	1.000	0.011	0	0	0	2	5	0
	2.782	2.997	1.634	0.182	1.000	0.020	0	0	0	2	5	0
	3.895	4.350	1.668	0.150	1.000	0.024	0	0	0	2	5	0
5	2.937	2.657	1.200	0.451	1.000	0.020	0	0	0	3	5	0
	3.784	4.147	1.200	0.429	1.000	0.020	0	0	0	3	5	0
	6.184	6.786	1.200	0.414	1.000	0.020	0	0	0	3	5	0
6	2.875	2.939	13.550	1.358	1.000	0.070	0	0	3	5	5	2
	3.108	3.373	14.066	1.294	1.000	0.071	0	0	3	5	5	2
	4.274	4.528	14.120	1.076	1.000	0.071	0	0	3	4	5	2
7	2.859	3.704	1.200	5.975	1.000	0.020	0	0	0	0	5	0
	4.365	4.894	1.200	5.827	1.000	0.020	0	0	0	0	5	0
	5.290	5.992	1.200	5.835	1.000	0.020	0	0	0	0	5	0
8	3.348	4.363	7.930	0.603	0.962	0.000	0	0	1	3	5	0
	3.341	3.514	8.142	0.512	0.955	0.000	0	0	1	3	5	0
	4.733	4.991	8.234	0.479	0.956	0.000	0	0	1	3	5	0
10	0.361	0.520	45.648	0.144	1.000	0.038	0	0	5	2	5	0
	1.680	2.064	45.589	0.144	1.000	0.041	0	0	5	2	5	0
	5.280	6.490	45.618	0.143	1.000	0.042	0	0	5	2	5	0

Table 2
Prediction for submission for 7 sensors on 18, 23, and 25 November 2022 with respect to three rows of each sensor row. The time for predicting dates is set as follows: 8 am-9 am, 11 am-12 pm, and 5 pm-6 pm. The last input date is 17 November 2022.

[2] M.-S. Dao, T.-H. Dang, T.-L. Nguyen-Tai, T.-B. Nguyen, D.-T. Dang-Nguyen, Overview of mediaeval 2022 urban air: Urban life and air pollution, MediaEval'22: Multimedia Evaluation Workshop (2022).

[3] D. Duong, Q. Le, T.-L. Nguyen-Tai, H. Nguyen, M. Dao, B. Nguyen, An Effective AQI Estimation Using Sensor Data and Stacking Mechanism, 2021. doi:10.3233/FAIA210040.

[4] J. Wang, X. Li, L. Jin, J. Li, Q. Sun, H. Wang, An air quality index prediction model based on cnn-ilstm, Scientific Reports 12 (2022). doi:10.1038/s41598-022-12355-6.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation 1 (1989) 541–551. doi:10.1162/neco.1989.1.4.541.

[6] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–80. doi:10.1162/neco.1997.9.8.1735.

[7] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Omnipress, Madison, WI, USA, 2010, p. 807–814.

[8] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, CoRR abs/1409.1259 (2014). URL: http://arxiv.org/abs/1409.1259. arXiv:1409.1259.

[9] D. E. Rumelhart, J. L. McClelland, Learning Internal Representations by Error Propagation, 1987, pp. 318–362.