Air Quality Prediction Based on a Deep Learning Hybrid Model

Keke Han
School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou, Henan, China

Abstract: Accurate prediction of air quality is crucial for the management of public health and environmental. This paper uses Zhengzhou air quality data from 2018 to 2020 to predict AQI. Among them, six major pollutant concentrations, including sulfur dioxide, nitrogen dioxide, carbon monoxide, inhalable particles, ozone fine particles, are selected as key influencing points to predict the air quality index. Addressing the limitation of traditional models in fully capturing the complex spatiotemporal dependencies within air quality data, this study proposes a deep learning hybrid model integrating Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and Attention Mechanisms. This model first utilizes CNN to effectively extract spatial correlations among pollutants. In order to ensure the long-term dependence of air quality, the air quality characteristics extracted from CNN will be input into LSTM. Finally, add the attention mechanism to the LSTM layer. In the attention mechanism, a higher weight can be automatically assigned to key information. In this process, the model's interpretation of important features can be enhanced. Experimental results show that the CNN-LSTM-Attention model achieves lower mean absolute error and root mean square error compared to other models, while exhibiting a higher R^2 value. This demonstrates that the prediction accuracy of the hybrid model is higher and more suitable for air quality index prediction.

Keywords: Air Quality; Convolutional Neural Network; Long Short-Term Memory Neural Network; Attention Mechanism

1. Introduction

In recent years, the social economy has progressed rapidly, and the air quality problem has become more serious, which has attracted wide attention from the public. Therefore, constructing air quality models to forecast air quality is of great importance for people's work and daily lives.

Research models for air quality forecasting mainly include traditional statistical models, machine learning statistical models, and hybrid models. Traditional statistical models are based on statistics. They built a mathematical model for analysis based on the characteristics of previous air quality data and other related factors. For example, Wang Siyuan et al. [1] used PM2.5 concentration data from all four seasons in Yan'an City to develop a multiple linear regression model, and the prediction errors were minimal. Nastaran Talepour et al. [2] used multiple linear regression analysis based on meteorological factors to forecast seasonal and annual PM concentrations. They found that the multiple linear regression model had higher accuracy in autumn.

As artificial intelligence advances, more scholars have found that machine learning models have higher predictive accuracy in many complex tasks. Zhong Lin et al. [3] used random forests and three-fold cross-validation to select optimal features and then input them into a DE-ELM model to predict the AQI. The experimental results showed that the models with variable selection achieved higher precision and enhanced credibility. Pooja Chaturvedi [4] considered the concentration levels of seven pollutants and used various machine learning models to forecast air quality in Varanasi. The results indicated that the models based on decision trees and random forests achieved an accuracy of approximately 100%. Shelly Sachdeva et al. [5] used meteorological factors and pollutant concentration to study and analyze air quality. They used machine learning models such as LSTM, ANN, and SVR for air quality prediction and found that different methods were suitable for different pollutants, with ANN being able to predict nearly all pollutants.

As research phenomena become increasingly complex, scholars have discovered that combining different models into hybrid models can integrate their respective strengths, enabling the handling of more complex data types and relationships. Wen Changjun et al. [6] optimized a BP neural network using the ISMA algorithm to construct an ISMA-BP model. Experiments showed that this model achieved higher prediction accuracy than others and improved the forecasting performance. Chen Changfeng et al. [7] proposed a Bi-LSTM-MA air quality prediction model for Shanghai and found that it effectively reduced prediction errors. Anh Tuan Nguyen et al. [8] combined XGBoost and ACNN-ARIMA-QPSO-LSTM, which addressed the nonlinearity and randomness of air particulate matter. They observed reductions in various errors and improvements in the coefficient of determination, demonstrating the enhanced effectiveness of the model in predicting air quality indices. Guo et al. [9] constructed a hybrid model of VMD-CSA-CNN-LSTM to forecast air quality indices for nine representative Chinese cities. This study not only enhanced the accuracy of air quality index predictions but also provided new tools for air quality management. Andri Pranolo et al. [10] combined attention mechanisms with long-short-term memory models to develop the BA-LSTM model. Using Beijing's air pollutant data to forecast PM2.5 concentrations, they found that the BA-LSTM model achieved the optimal mean RMSE and MAPE and demonstrated versatility in multivariate time-series analysis. Aysenur Gilik et al. [11] proposed a method integrating CNN with LSTM to predict air pollutant concentrations across multiple urban locations by leveraging spatiotemporal relationships.

The air quality index is an indicator used to quantitatively describe the air quality situation. It comprehensively considered the concentration of the six major pollutants. It converts complex air pollution conditions into numerical values, enabling the public to quickly understand the current level of air pollution. The Air Quality Index is influenced by multiple factors, and traditional prediction models face limitations in forecasting its values. To enhance prediction accuracy, this study utilizes air quality data from Zhengzhou City between 2018 and 2020. By integrating Convolutional-Neural-Networks (CNN), Long-Short-Term Memory (LSTM), and attention mechanisms, it leverages the strengths of each model to forecast the Air Quality Index, thereby supporting air quality monitoring.

2. Algorithm Principles

2.1 Affiliations

A convolutional neural network is a positive propagation neural network that can process spatial data. Designed to address the complexity of image classification, they are now widely applied across various visual tasks. Unlike traditional image processing techniques, CNNs do not require manual feature selection; they can directly identify relevant features from the data. The CNN architecture includes multiple levels. The convolutional layer is convolutionized by the output characteristic vector and convolutional kernel of the previous layer. Apply the nonlinear activation function RERU to construct output characteristic vectors so that the model can learn complex data relationships. The pooled layer generally uses the maximum pooled layer. This reduces the number of trainable parameters, improves efficiency, and preserves essential features. Finally, the weighted sum of these features is computed through the output layer to yield the final result. Figure 1 shows the framework of the model.

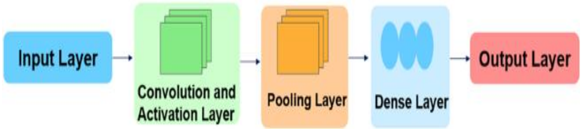

Figure 1. CNN Model

2.2 LSTM

LSTM is a kind of deformation of RNN. It addresses the issues of gradient vanishing or gradient explosion encountered by RNNs when processing long sequences.

LSTM employs a “gating” mechanism to control information flow, primarily comprising four gates: the input gate, the forget gate, the output gate, and the memory gate. Specifically, the input gate determines which new information should be stored. It outputs a value between 0 and 1 via the activation function σ, which is used for inputting information. Forgetting decides some information that should be forgotten. The output gate determines which information should be output. Based on the current input and the hidden state from the previous time step, it outputs a value between 0 and 1 to generate the hidden state. The memory cell is an internal state introduced in the LSTM network, specifically designed for linear recurrent information transmission. The hidden state represents the network's internal representation when processing sequential data. It can incorporate information from any node preceding the current sequence node, enabling LSTMs to leverage past information for predicting future outputs. This grants them enhanced expressive power when handling sequential data.

$$I_t=\sigma(X_t W_{xi}+H_{t-1} W_{hi}+b_i \quad (1)$$

$$F_t=\sigma(X_t W_{xf}+H_{t-1} W_{hf}+b_f) \quad (2)$$

$$O_t=\sigma(X_t W_{xo}+H_{t-1} W_{ho}+b_o) \quad (3)$$

2.3 Attention Mechanism

The attention mechanism is a mechanism that mimics human visual attention. It can make the model pay more attention to important data while ignoring less important information during information processing. Its structure is shown in Figure 2. This approach effectively reduces the prediction model's focus on non-critical regions at any given time, thereby increasing its attention to critical regions. This method of weighting different areas improves the processing efficiency.

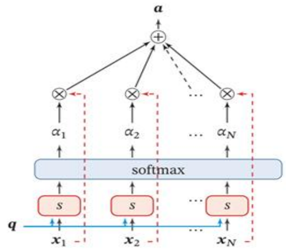

Figure 2. Attention Mechanism Structure

2.4 CNN-LSTM-Attention Model

To enhance the accuracy of Air Quality Index (AQI) predictions, this paper proposes a hybrid model based on Convolutional Neural Network-Long Short-Term Memory-Attention Mechanism (CNN-LSTM-Attention) for forecasting. Figure 3 illustrates the architecture of this hybrid model. The CNN component extracts local features from data through convolutional computations. These features serve as input to the LSTM network. Subsequently, use the attention mechanism in the output layer of LSTM. This dynamically assigns weights based on feature importance, making the model to pay more attention to information critical for air quality prediction. This approach enhances both prediction accuracy and generalization capability.

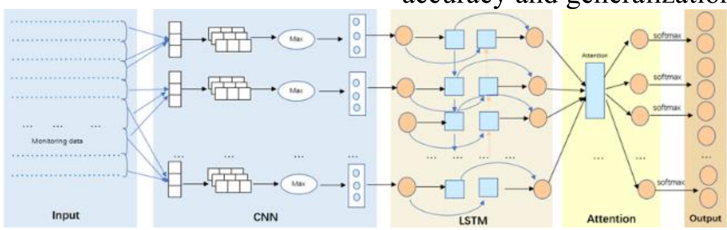

Figure 3. CNN-LSTM-Attention Structure

3. Experimental Analysis

3.1 Data Sources

This study utilizes daily air quality data from Zhengzhou City from 2018 to 2020, comprising 1,096 data points. The dataset originates from the Zhengzhou Atmospheric Environment Monitoring Station and the Zhengzhou Meteorological Bureau, primarily including seven indicators: AQI, O_3, PM10, NO_2, CO, SO_2 and PM2.5. Figure 4 illustrates the temporal variation of AQI.

3.2 Data Processing

Input data undergoes normalization using the minimum-maximum normalization method, scaling feature values proportionally to fall within the range [0,1]. This normalization facilitates model learning and mitigates disproportionate impacts caused by inconsistent feature value ranges. The formula is as follows. Data is partitioned chronologically at an 80:20 ratio: the first 80% constitutes the training set, while the remaining 20% forms the test set.

$$x\_scaled=(x-x\_min)/(x\_max-x\_min) \quad (4)$$

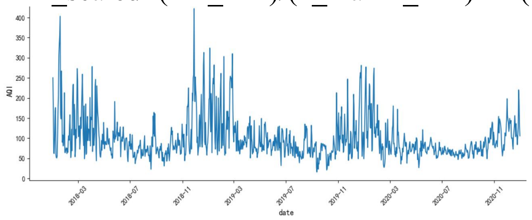

Figure 4. AQI Variation

3.3 Evaluation Criteria

To compare the predictive capabilities of different models for the Air Quality Index (AQI), this paper employs the following four evaluation metrics: R^2, MAE, RMSE, and MAPE. The lower the values of MAE, RMSE and MAPE, the better the prediction effect of the model. R^2 takes the value between 0 and 1. The closer the value is to 1, the higher the degree of fit between the model and the data.

3.4 Parameter Settings

This paper employs a hybrid CNN-LSTM-Attention model for air quality prediction. Parameter configuration must balance fitting capability and generalization performance. Input data comprises six features—concentrations of six major pollutants—with the input dimension set to 6 and output dimension to 1. The time step is set to 7, utilizing the previous seven days' data to forecast the AQI at future time points. The CNN module extracts features from input data using one convolutional layer with ReLU activation to mitigate vanishing gradients, employing max pooling. The LSTM layer captures long-term dependencies of local features extracted by CNN across time, feeding all hidden states from this layer's time steps into the subsequent attention mechanism. The attention layer assigns different weights to LSTM at different time steps. The attention weights are computed using the dot product function with the tanh activation function. Finally, LSTM output and attention weight weighting summation. After passing through the attention layer, the data enters a fully connected layer. The output of this layer is subjected to nonlinearity enhancement via a hyperbolic tangent activation function, and the final prediction is completed through another fully connected layer. Table 1 presents the parameter settings.

Table 1. Parameter Settings

Parameter	Quantity
Convolution layer	64
Pooling layer	Max
Activation function	ReLU
padding	Same
units	64
Optimizer	Adam
Batch size	32
epoch	100
Learning rate	0.001

3.5 Prediction Results

This paper employs BP, LSTM, LSTM-Attention, CNN-LSTM, and CNN-LSTM-Attention models to predict AQI based on AQI and pollutant data. Evaluation metrics reveal significant differences in predictive performance across models.

As shown in Table 2, the BP model exhibits higher prediction errors across all metrics compared to other models and has the lowest R^2 value, indicating lower prediction accuracy. Compared to LSTM, LSTM-Attention reduces MAE, RMSE, and MAPE while increasing R^2 by 4.2%, validating that the attention mechanism enhances the base model's performance by dynamically weighting key temporal features. Compared to CNN-LSTM, CNN-LSTM-Attention reduced MAE by 14.2% and RMSE by 16.6%, demonstrating that the attention mechanism optimizes the integration efficiency of CNN-extracted spatial features in temporal modeling. While CNN-LSTM showed some performance improvement over LSTM, its MAE was higher than LSTM-Attention when used alone, indicating that modeling complex temporal dynamics requires incorporating the attention mechanism. In terms of overall model performance, CNN-LSTM-Attention achieved an MAE of 7.1460, RMSE of 8.3597, and MAPE of 10.0166. Both MAE and RMSE were lower than other models, while R^2 reached 0.8680, surpassing other models. This indicates superior error reduction and data fitting capabilities compared to alternative models. Figure 5 demonstrates that the predicted values from the CNN-LSTM-Attention model closely align with actual values, exhibiting a similar trend to the true values and indicating strong performance.

Table 2. Table Example

Model	MAE	RMSE	MAPE	R2
BP	12.3033	17.2699	14.6501	0.7058
LSTM	8.1384	10.8998	9.4620	0.7938
LSTM-Attention	7.8193	9.6178	8.8791	0.8273
CNN-LSTM	8.3295	10.0281	8.1252	0.8317
CNN-LSTM-Attention	7.1460	8.3597	10.0166	0.8680

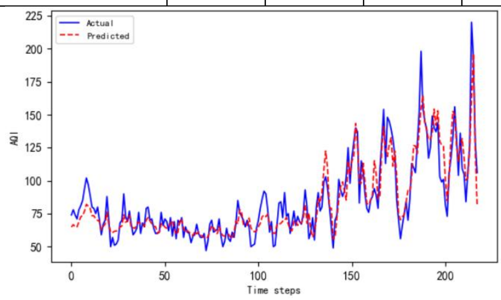

Figure 5. CNN-LSTM-Attention True Values vs. Predicted Values Comparison Chart

5. Conclusions

To address the issue of poor air quality forecasting caused by the complexity and volatility of pollutant concentrations, this paper proposes a hybrid CNN-LSTM-Attention model for air quality prediction. The CNN component extracts correlations among pollutant concentrations in air quality data. These features are then fed into an LSTM network to learn temporal patterns in air quality data, capturing how pollutant concentrations influence air quality. Finally, apply the attention mechanism on LSTM, dynamically assigning weights based on feature importance to focus on information critical for air quality prediction. To validate model performance, this study compared five models: BP, LSTM, CNN-LSTM, LSTM-Attention, and CNN-LSTM-Attention. Results demonstrate that the performance of CNN-LSTM-Attention is better than that of other models, exhibiting lower prediction errors. Future research may consider incorporating seasonal and meteorological factors to further enhance model performance.

References

[1] Wang Siyuan, Ren Ying, Xia Bisheng, et al. PM2.5 Concentration Prediction in Yan'an City Based on Multiple Regression. Information Technology, 2024, (05): 15-21.

[2] Talepour, N., Birgani, Y.T., Kelly, F.J. et al. Analyzing meteorological factors for forecasting PM10 and PM2.5 levels: a comparison between MLR and MLP models. Earth Sci Inform,2024,17: 5603–5623.

[3] Zhong Lin, Yan Qisheng. Interpretable Machine Learning Modeling Method for Air Quality Prediction. Computer Simulation, 2023, 40(10): 511-518.

[4] Chaturvedi, P. Air Quality Prediction System Using Machine Learning Models. Water Air Soil Pollut, 2024, 235(578).

[5] Shanshan S, Qilong H. Multi-view multi-task spatiotemporal graph convolutional network for air quality prediction. The Science of the total environment,2023, 893164699-164699.

[6] Wen Changjun, Chen Yangyang, He Yonghao, et al. Air Quality Index Prediction Based on Multi-Strategy SMA-BP Neural Network. Electronic Measurement Technology, 2023, 46(22): 78-86.

[8] Nguyen, A.T., Pham, D.H., Oo, B. et al. Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. Big Data, 2024, 11: 71.

[9] Guo, Z., Jing, X., Ling, Y. et al. Optimized air quality management based on air quality index prediction and air pollutants identification in representative cities in China. Sci Rep, 2024, 14: 17923.

[10]Pranolo A, Zhou X, Mao Y. A novel bifold-attention-LSTM for analyzing PM2.5 concentration-based multi-station data time series. International Journal of Data Science and Analytics,2024, (prepublish): 1-18.

[11] Aysenur G, Selcuk A O, Atilla O. Air quality prediction using CNN+LSTM-based hybrid deep learning architecture. Environmental science and pollution research international, 2021, 29(8): 11920-11938.