Research and Evaluation of Deep Learning Models for Air Quality Prediction

Authors: Li Jiaming

Date: 2023-09-22T00:00:00+00:00

Abstract

Objective: Timely and accurate air quality prediction data is crucial for environmental management, especially during heavy air pollution episodes, providing data support for decision-making by government ecological environment management departments to respond to pollution conditions and precisely allocate social resources. Methods: The deep learning-based air quality prediction model AirNet6, developed by the author, balances accuracy and real-time performance, achieving predictions of ozone, sulfur dioxide, carbon monoxide, and other pollutants for 7 days or longer. Results: Unlike traditional chemical model calculations, this model employs Spatio-Temporal Graph Convolutional Network (STGCN) to capture patterns from historical monitoring data, weather forecast data, social activity data, etc., completing predictions for over 100 sites for the next 168 hours within 2 minutes. Conclusion: Experimental results demonstrate that the AirNet6 model achieves significant improvements in speed, energy efficiency, and accuracy compared to traditional chemical models and time series AI models. Keywords: air quality prediction, artificial intelligence, deep learning model, spatio-temporal graph convolutional network

Full Text

Preamble

Research and Practice of Deep Learning Models for Air Quality Prediction

Li Jiaming[1]
[1](Guangdong Eco-environment Monitoring Center, Guangzhou 510000, China)

[**Objective**] Timely and accurate air quality prediction data is crucial for environmental management, particularly during heavy pollution episodes when such forecasts provide essential data support for government ecological environment management departments to respond to pollution conditions and allocate

social resources precisely. [**Methods**] The author developed AirNet6, a deep learning-based air quality prediction model that balances accuracy and real-time performance to achieve 7-day or longer forecasts for ozone, sulfur dioxide, carbon monoxide, and other pollutants. [**Results**] Unlike traditional chemical model calculations, this model employs Spatio-Temporal Graph Convolutional Networks (STGCN) to capture patterns in historical monitoring data, weather forecasts, and social activity data, completing predictions for over one hundred monitoring stations across 168 hours within two minutes. [**Conclusion**] Experiments demonstrate that the AirNet6 model achieves significant improvements in speed, energy efficiency, and accuracy compared to traditional chemical models and time-series AI models.

Keywords: air quality prediction; artificial intelligence; deep learning model; STGCN
Classification Numbers: X823; TP183

Traditional chemical model-based air quality prediction methods typically rely on complex simulations using pollution source emission data and weather forecasts. These computations require massive server clusters or even supercomputers, taking several hours to generate results and thus usually permitting only one daily forecast. However, air quality prediction data is vital for environmental management, serving as the fundamental basis for government decision-making, resource allocation, and response measures. During heavy pollution events, management agencies and experts need to make real-time decisions based on weather and pollution conditions, often holding several consultations daily. This demands more timely prediction data reflecting current conditions—a requirement that chemical models struggle to meet.

With the development of artificial intelligence, common time-series prediction models such as RNN, LSTM, and Prophet can forecast subsequent data changes based on historical monitoring data (e.g., from the previous hour) and deliver results quickly, satisfying real-time prediction needs. However, these time-series models only capture temporal patterns at individual monitoring points without considering inter-site correlations or weather factors, resulting in insufficient accuracy.

To simultaneously improve both accuracy and computational speed, the author developed a novel deep learning model based on Spatio-Temporal Graph Convolutional Networks (STGCN). This model leverages not only historical air quality data from monitoring sites and their spatial relationships but also incorporates corresponding weather forecast data and social activity factors to rapidly predict air quality data within 168 hours. STGCN was originally developed for traffic flow prediction in transportation networks ("Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," 2018). It comprises several spatio-temporal convolutional blocks and a final output layer, with each block containing temporal and spatial convolutional layers to capture temporal features and spatial dependencies of node data, respectively.

1. Model Design

[Figure 1: see original paper] Model Data Flow Diagram

1.1 Model Input

As shown in Figure 1, the model input is a 3-dimensional tensor. The first dimension represents time with an hourly granularity—for instance, using one year of data yields a time dimension length of 365$×$24=8760. The second dimension represents monitoring stations; for example, Guangdong has 125 national-controlled air monitoring stations, making the station dimension length 125. The third dimension represents channels corresponding to different factors: the first channel contains monitoring data (e.g., ozone concentrations), the subsequent seven channels contain weather forecast data (air pressure, solar radiation, precipitation, temperature, humidity, u-wind speed, and v-wind speed), and the final channel encodes social activity data.

Notably, weather forecast and social activity data are predictable, meaning they extend further into the future than historical monitoring data. For example, when predicting air quality seven days ahead, the model can utilize already-available 7-day weather forecasts and social activity projections rather than relying solely on current historical data, thereby improving prediction accuracy.

Since air quality data is point-based from monitoring stations while weather forecast data is gridded—such as NOAA's GFS (Global Forecast System) data at 0.25-degree latitude/longitude resolution—their spatial domains must be unified for integration. The author uses monitoring station locations as the unified spatial domain, significantly reducing data volume. Weather forecast values for each monitoring station are obtained through interpolation based on station coordinates.

Furthermore, research and testing revealed that micro-level social activity data such as pollution source emissions, economic activities, and traffic flow are difficult to collect accurately and comprehensively, and their temporal and spatial domains differ substantially from monitoring data, making integration challenging like meteorological data. Therefore, the author employs a macro-level statistical value as a predictable indicator of social activity intensity. Since social activity frequency primarily correlates with the alternation between "working days (w)" and "rest days (r)," enumerating combinations of "previous day," "current day," and "next day" yields eight patterns: www, wwr, wrw, wrr, rww, rwr, rrw, and rrr. By calculating historical monitoring data statistics for these eight scenarios, 24-hour average values for each station under each condition can be derived as empirical values representing social activity intensity, as shown in Figure 2:

[Figure 2: see original paper] Mean monitoring data under 8 working day patterns

1.2 Network Structure

The model architecture, depicted in Figure 3, consists of three spatio-temporal convolutional blocks and one output layer. The output layer contains four connection blocks to simultaneously predict data for four time points. In the spatio-temporal convolutional layers, the third block interfaces with the output layer through four convolutional blocks connected to the four output blocks, while the first two blocks each have a single unified convolutional block.

[Figure 3: see original paper] Network structure diagram of the model

1.3 Model Output

The initial model output was a 1-dimensional tensor with the station dimension representing air quality at n hours ahead for each station. This required training a separate model for each time point −168 models for 7-day predictions, or 1008 models for 6 pollutants (168$×6). $Through iterative research, the author implemented a time-shifting approach that requires only 8 mode$ $n = [3, 6, 12, 24, 48, 72, 120, 168]. This reduces the total to 48 models (8×$6)$. For time points between these values, predictions use the nearest subsequent time point' s model with appropriately shifted input data. For example, lacking a model for time point 10, the model for time point 12 is used, but input data is shifted forward by 2 hours.

As illustrated in Figure 4, assuming the model for time point 12 requires 24 hours of input data, with current time t and all data up to t available, predictions for t+7 through t+12 all use the time point 12 model. Specifically, predicting t+12 uses input data from t-23 to t, while predicting t+10 uses input data from t-25 to t-2.

[Figure 4: see original paper] Model data input-output mapping diagram

Similarly, the prediction time coverage for each time point n model is shown in Table 1:

Time point coverage table

Further iterative research refined the model to multi-output architecture (Figure 3), requiring only three models per pollutant: one simultaneously predicting n=[3,6,12,24] (four time points), one predicting n=[48,72] (two time points), and another predicting n=[120,168] (two time points), totaling just 18 models (3$×$6). Testing revealed that the 4-2-2 time point segmentation reduces model count and training time with negligible accuracy loss. However, 4-4 segmentation decreases accuracy for the latter four time points (48,72,120,168), and using a single model to output all eight time points significantly reduces accuracy. Thus, the 4-2-2 segmentation represents the optimal choice.

The final iterated model outputs a 2-dimensional tensor where the first dimension represents time points (length 2 or 4) and the second dimension represents stations.

2. Experiments

2.1 Dataset

The author utilized hourly air quality data from 125 national-controlled stations in Guangdong Province as training and validation data, covering six pollutants: ozone, nitrogen dioxide, carbon monoxide, sulfur dioxide, PM10, and PM2.5. Station latitude and longitude coordinates enable interpolation with weather forecast data.

Weather forecast data was sourced from NOAA' s GFS0.25 dataset, covering a rectangular latitude-longitude region over Guangdong ("north": 26, "west": 109, "south" : 20, "east" : 118). Real-time GFS forecast data for routine operation is obtained from NOMADS (NOAA Operational Model Archive and Distribution System), while historical GFS data (retained for only 8 days on NOAA's website) was acquired from UCAR (University Corporation for Atmospheric Research). Both channels use the OPeNDAP interface.

As previously described, social activity data uses macro-level "working day-rest day" empirical values. The author employed Python' s holidays module and chinese_{calendar} module to obtain holiday data, combined with historical air quality statistics.

The dataset spans from April 1, 2021, to October 1, 2022, with data from April 1, 2021, to July 1, 2022, used for training and data from July 1, 2022, to October 1, 2022, used for testing—representing five quarters for training and one quarter for testing.

2.2 Data Preprocessing

Since national-controlled air quality monitoring relies on online instruments, missing data occurs regularly. The model cannot process empty values, so missing data must be imputed. The imputation first employs spatial interpolation using Kriging with data from neighboring stations at the same time. If missing data persists (typically due to network or central server failures affecting most stations simultaneously), temporal imputation is applied using the station' s data from the previous day.

Weather forecast data rarely has missing values, and redundancy is substantial (multiple predictions exist for the same time point from preceding days). Missing values are simply filled with data from the previous or following day at the corresponding time.

The macro-level social activity indicator has no missing values and requires no imputation.

All input data must be normalized before entering the model. Normalization compresses all values proportionally into the [0,1] range to facilitate neural network computation. While normalization bounds are typically set to the maximum and minimum of training samples, the author expanded these bounds

twofold to enable the trained model to handle out-of-sample data. Letting Smax, Smin, and Savg represent the maximum, minimum, and average values of all sample data, the normalization bounds are calculated as:

Normalizedmax = 2×Smax - Savg
Normalizedmin = 2×Smin - Savg

2.3 Experimental Setup

The author conducted training and validation using three PCs, each equipped with an NVIDIA 3080 or 3090 GPU providing computational power. Power consumption during training was approximately 300W, with each model requiring about 2 hours to train. The model programs ran in an Ubuntu 20.04 environment with Python 3.9, using PyTorch 1.12.0 and CUDA 11.7. To accelerate computation and reduce GPU memory usage, all operations used float32 data types.

2.4 Experimental Results

Following the "6.2 Statistical Evaluation of Single Pollutant Concentration Forecast" section in the "Technical Specification for Numerical Air Quality Forecasting (HJ 1130-2020)," the author calculated bias, error, and correlation between predicted and observed values for four models from July 1, 2022, to October 1, 2022. Results are presented below:

(1) Normalized Mean Bias (NMB)

Comparison of normalized mean bias results

Due to large biases in the chemical model, its data underwent preliminary bias correction (dividing by (1+bias)) to produce "Chemical Model (Bias-Corrected)" results.

(2) Root Mean Square Error (RMSE)

Comparison of root mean square error results

(3) Correlation Coefficient (r)

Comparison of correlation coefficient results

Experimental results demonstrate that AirNet6 achieves significantly higher accuracy than other models for four pollutants: ozone, nitrogen dioxide, PM10, and PM2.5. For carbon monoxide and sulfur dioxide, its accuracy is very close to Prophet and significantly superior to other models.

3. Routine Operation

After training, models typically remain effective for one quarter to one year. Since the model only captures patterns in training data (April 1, 2021, to July 1, 2022, in this study), prediction accuracy declines when internal or external

environmental factors change, necessitating retraining with new data for better alignment. During routine operation, no training is required—only inference using trained models—consuming minimal resources. On a standard PC (AMD Ryzen 7 1700 CPU, NVIDIA 970 GPU), predictions for six pollutants across 168 hours complete within two minutes, enabling hourly forecast density in production and facilitating responses to heavy pollution events.

4. Resource Consumption Comparison

Comparison of resource consumption

5. Conclusion and Future Work

This paper presents research and practice on a novel deep learning model that captures regional air quality patterns over time and in relation to weather and social activities, enabling rapid and accurate future air quality predictions. Experiments on real-world national-controlled monitoring datasets from Guangdong Province demonstrate the new model' s superiority over alternative methods, proving highly valuable for implementing more refined environmental management.

Future work will focus on continued model optimization to improve accuracy and expand applicability to other domains such as river water quality monitoring and noise prediction.

References

(1) Bing Yu, Haoteng Yin, Zhanxing Zhu. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, arXiv:1709.048

(2) Tomasz Stańczyk and Siamak Mehrkanoon. Deep Graph Convolutional Networks for Wind Speed Prediction, arXiv:2101.10041 [cs.LG]

(3) Yi-Fan Zhang[1](B), Peter J. Thorburn[1], and Peter Fitch[2]. Multi-task Temporal Convolutional Network for Water Quality Sensor Prediction https://link.springer.com/chapter/10.1007/978-3-030-36808-1_{14}

(4) Taylor Letham. Forecasting at Scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2

(5) Fjellstrom, C. (2022). Long Short-Term Memory Neural Network for Financial Time Series. 2022 IEEE International Conference on Big Data (Big Data). https://arxiv.org/abs/2201.08218

Author Contribution Statement: Li Jiaming: proposed research ideas, designed research methodology, implemented experiments, and wrote the paper.