International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Air Quality Index Prediction

Dr Kanmani P[1], Sarthak Gupta[2], Gurleen Kaur[3]

[1]Department of Data Science and Business Systems, SRM Institute of Science and Technology, Chennai, India
[2] Department of Data Science and Business Systems, SRM Institute of Science and Technology, Chennai, India
[3] Department of Data Science and Business Systems, SRM Institute of Science and Technology, Chennai, India

---***---

Abstract - We address air index forecast by inferring, using deep and machine learning algorithms, the concentration per hour of air pollutants (such ozone, Particulate Matter (2.5), and Sulphur). DL is one of the most popular methods, may efficiently apply optimization algorithms to train a model on large amounts of information and data. The air quality prediction model is trained using time series data. In our approach, the sequence module makes use of deep CNN, and time series data is progressively fed into the CNN model for training. In CNN, there are many different functioning layers, including RNN, and LSTM. The CNN layer can be utilized to successfully extract the features that are sequential of time series. Advanced features outperform normal features when analyzing data of time series. CNN's pooling layer oversees down sampling. CNN is good at forecasting air quality, according to test data.

***Key Words*:** AQP, CNN, PM, RNN, DL

1.INTRODUCTION

Predicting air quality is crucial because as the economy grows, air pollution gets worse and poses a risk to people's health. For example, diesel oil and industrial emissions are the main outlets and factors that cause cancer risks (70%), in central Tehran [1]. Air around us contains a variety of exhaust gases from factories, autos, and power plants that burn coal, such as those that use suspended particles like Particulate Matter (10), Particulate Matter (2.5), nitrogen dioxide (NO2), ozone, and Sulphur dioxide. PM2.5, which has a diameter of less than 2.5 Micrograms, which has attracted particular interest from the medical community because of the tiny particles that it contains that penetrate the passage through the nose. Including rhinitis, tiny particles can affect the hearts' health and increase the risk of cancer concerning lungs by entering the lungs through the respiratory system. The main sources of particulate matter are industrial exhaust and diesel exhaust from automobiles, Polycyclic aromatic hydrocarbons are present in both. After entering the body, Particulate Matter (2.5) can stimulate blood vessel inner walls, leading to clot coagulation, and ultimately affect the heart and degrade it's condition . As a result, living in an environment with high PM2.5 concentrations for a prolonged period elevates the danger of premature death due to chronic heart and lung diseases. [2]. Predicting the amount of pollution present in the air is a common application of time series prediction. Air pollution data is often recorded over time and can exhibit trends, seasonal patterns, and other time-dependent variations. Time series analysis techniques, such as autoregressive integrated moving average models i.e. ARIMA models or seasonal decomposition, can be used to analyze historical air pollution data and forecast future levels. Accurately predicting the amount of pollution present in the air is important for informing policy decisions, public health measures, and individual behavior choices. AQP has traditionally employed ways such as linear and nonlinear regression. With small data sets, simple causality analysis is highly effective, but regression analysis is speedy for modeling, it might be difficult to describe nonlinear data and analyses excessively complicated data sets. Neural networks are more effective at resolving the problems and are also more adapted to the challenge to forecast air for huge, complicated data sets. SVR [3] is an extension of support vector classification (SVC) [4], which uses kernel functions to translate data to a high-dimensional space. Time series prediction, including financial prediction, time series-related disciplines include translation by machine, verbal recognition, text identification, etc., is now becoming more and more dependent on CNN, long-term short memories (LSTM), and RNN. In image processing, object recognition, automated driving, and facial recognition, convolutional neural networks are widely used. Convolutional neural networks' main benefit is the variety of its units, which include convolution, pooling, rectification, hyperbolic functions, normalizing, and regularization. - Different processing units for time series signals can give a variety of local features and produce a variety of processing effects. Deep learning models, particularly recurrent neural networks (RNNs), can stack many processing units, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs), on top of one another to represent sequential elements of a time series with vectors varying dimensionally. These processing units can capture long-term dependencies in the time series data by learning to selectively forget or remember information from previous time steps. The use of deep learning models has been shown to be effective in predicting time series data in a variety of domains, including finance, healthcare, and environmental monitoring.

2. PREDICTION MODEL

We have introduced an AQP model that relies heavily on the Convolutional Neural Network architecture. The model consists of convolutional and pooling layers, along with a connected layer, as depicted in Figure 1.

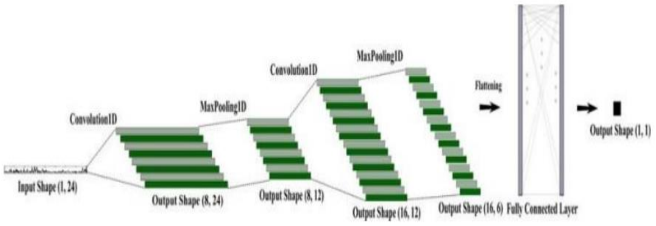

Figure 1. Architecture Diagram of CNN.

2.1 Convolution Neural Network Layer

The kernel vector, commonly referred to as the sliding window, is an important part of a convolution layer. Convolution being carried out moves successively over the input vector. The kernel vector is multiplied element wise with the input vector, and the sum added by a bias is put through the activation function to form the feature map for the next layer [5].

2.2*Pooling*

The pooling layer's basic idea is to offer translational invariance, with the goal of preserving identified features in a smaller form by rejecting less relevant data at the expense of resolution. [6]. The max algorithm slides the input vector by the window in stages, selecting the highest value of the window region and considering remains as garbage value. Instead, the values in the window area are averaged to conduct average pooling.

2.3Connected Layer (Fully)

The Fully Connected layer is typically a Multilayer Perceptron, a type of neural network. An MLP is made up of three node levels: the input, the output, and the hidden layer. Except for the input node layer, every node's output is linked to the function called activation function. During training, the backpropagation technique is frequently used to optimize the weights and biases that are linked to an MLP.

2.4Convolutional Neural Networks Architecture

The recommended convolutional neural networks design for Air Quality Prediction is shown in Figure no. 2. As they depend on time, the AQP data are time-series data. As a result, most of the elements in Figure no. 2 are 1-D vectors. All the kernels visible and feature mind maps are vectors as an example. Convolutional neural networks are fed a 24-vector containing 24 concentration values of the pollutants under consideration. The input vectors listed below are given to Convolutional Neural Network. The input vector consists of twenty-four values collected over twenty-four subsequent time steps. The ideal output for training data is the outcome from the following step. Consider 100 values that were attained in a row over time. The input vector consists of the first through twenty-fourth values, and the twenty-fifth value is used as the intended output; similarly, the input vector consists of the second through twenty-fifth values, and the twenty-sixth value is used as the expected output; etc. The forecast for the value at the kth time is made using the twenty-four values from the previous time steps as the input vector. With each layer in the Convolutional Neural Network, four segments are created. Much of portion one is devoted to convolution.

2.5DL Frameworks

The development of machine learning is fairly rapid. The key to making applications smarter is to construct the machine learning model such that it can be used successfully by embedded devices. It has been demonstrated that deep learning has positive outcomes in business operations. Artificial intelligence research enables complicated issues to be solved. Deep learning is a difficult undertaking. Engineers and scientists, constructing useful mathematical models and applying them in embedded systems is a demanding task. Today, a number of well-known e-commerce companies have made Deep Learning frameworks accessible to academics, including Tensor Flow from Google, MXNet from Amazon, PyTorchfrom Facebook, PaddlePaddle from Baidu, and x-deep learning from Alibaba, among others. These frameworks simplify the building of complex mathematical models while facilitating difficult programming tasks. This study utilizes PyTorch for constructing the convolutional neural networks model. PyTorch is a scientific computing framework that provides comprehensive assistance for deep learning techniques. It is widely adopted by major companies like Facebook, Twitter, and Google, and aims to enhance the efficiency of developing models and overall flexibility through the utilization of Compute Unified Device Architecture and C/C++ libraries for processing. Asa Python port of Torch, PyTorch can be used by anyone with a basic understanding of Python to build their own deep-learning models [7].

3. EXPERIMENTATION

Under this we showcase some experimental results that exhibit the efficacy of the model we proposed. Particulate Matter (2.5) values are used in our testing. Methods that are comparable may be utilized for the prediction of the congregation of additional impurities.

3.1Different sets of data used

Every hour, we gathered air quality and meteorological data from 77 sites from the Environmental Protection Administration's Executive Yuan database. [8]. The dataset is accompanied by 15 attributes. The properties' values were noted and observed. The 15 characteristics include THC, NO, NOX, CO, PM10, PM2.5, SO2, CO, andO3.

3.2*Ground truth and Loss Function*

After comparing the predicted Particulate Matter levels in accordance with the actual data at each location, the RMSE is used to assess the performance. RMSE stands for Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}}$$

Where yi(^) and yi are the projected and actual values for the i-th hour, and n is the size of the batch number of measurements. We determine the average RMSE for each epoch. After computing the RMSE error for each batch (batch size = 32). A forecast with less error is more likely to be accurate.

4. RESULT AND DISCUSSIONS

Table 1 displays the results and performance of the different Convolutional Neural Network models presented using the Air Quality Monitoring dataset. Root Mean Square Error method is used to assess the performance of the models. The table shows the results of the two recommended Convolutional Neural Network models, maximum pooling, and pooling average. The current kernel sizes are represented by the numerals one, three, and five. In the table below, the best of all has been bolded.

Table 1. RMSE RESULTS

	CNN-Max-1	CNN-Max-3	CNN-Max-5	CNN-Avg-1	CNN-Avg-3	CNN-Avg-5
Train	9.17	**6.07**	6.22	8.85	6.43	**6.11**
Test	9.72	8.20	**7.25**	9.49	8.31	**7.37**

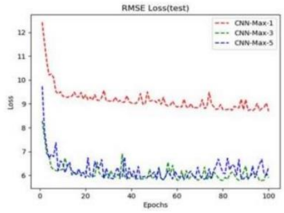

Figure 2. Max pooling performance Performance

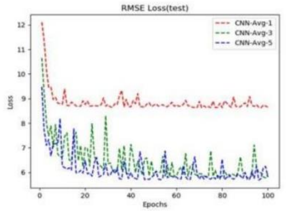

Figure 3. Convolutional Neural Network

5. CONCLUSION

We discussed how to use Convolutional Neural Network is a deep learning system utilized to predict the concentration per hour and buildup of air impurities. Time series data has always been used to train the air quality prediction algorithm. There are several important layers in Convolutional Neural Network, including convolution, pooling, and recurrent neural network. The convolution layer can successfully extract time series data's sequential qualities. According to test findings, Convolutional Neural Network is effective at forecasting air quality.

REFERENCES

[1] Sina Taghvaee, Mohammad H.Sowlat, Mohammad Sadegh Hassanvand, Masud Yunesian, Kazem Naddafi, Constantinos Sioutas, Source-specific lung cancer risk assessment of ambient PM2.5-bound polycyclic aromatic hydrocarbons (PAHs) in centralTehran, 2018, abstract.
[2] Wen-Chi Pan, Chih-Da Wu, Mu-Jean Chen, Yen-Tsung Huang, Chien-Jen Chen, Huey-Jen Su, Hwai-I Yang, “Fine Particle Pollution, Alanine Transaminase, and Liver Cancer: A Taiwanese Prospective Cohort Study (REVEAL-HBV),“ JNCI: Journal of the National CancerInstitute, Volume 108, Issue 3, 1 March 2016, djv341.
[3] Vladimir Vapnik, The Nature of Statistical Learning Theory, Springer New York, NY, 1995.
[4] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik, “A training algorithm for optimal margin classifiers,“ In Proceedings of the Fifth Annual Workshop on Computational. LearningTheory, pp. 144-152, ACM Press, 1992.
[5] Zhifei Zhang, University of Tennessee, Knoxville, TN, America, “Derivation of Backpropagation in Convolutional Neural Network,“ 2016, unpublished.
[6] Environmental Protection Administration Executive Yuan R.O.C.(Taiwan) environmental resources database.
[7] Jeff Hale, Deep Learning Framework Power Scores 2018, unpublished. Environmental Protection Administration Executive Yuan R.O.C. (Taiwan) air quality Monitoring website