A Deep Learning Framework for Air Quality Prediction

Group Nr. 13
Vicent Descals Carbonell, Priya Prabhakar

11th November 2022

1 Introduction and summary of the selected paper

- The general topic of research is to study spatial-temporal data in order to make predictions. In this case, we are going to study air quality levels in different cities in India.
- The main idea on the original paper is to model spatial dependencies and temporal dynamics as the key of traffic predictions [1]. **Given that air quality also vary depending on the location and the time are analyzed, this makes our problem very similar to the one addressed in the original paper.**
- In the original paper implemented, a Spatial-Temporal Dynamic Network to make accurate traffic predictions for real-world applications. The use two real-world datasets, NYC-Taxi dataset (40 days for training and 20 days for testing) and the NYC-Bike dataset (40 days for training and 20 days for testing). **In our case, we will use a dataset that represents the air quality from different cities in India and split it in data from 60 days (2 months) for training and 30 days (1 month) for testing.**

2 Problem statement

With the development of industry, air pollution has become a serious problem. Monitoring it and understanding its quality is of immense importance to our well-being. Prediction is important for Current India Air Quality Index (AQI) is 189 POOR level with real-time air pollution.
The problem statement is to use deep learning framework as proposed in [1] to estimate air quality in India.
We will use daily air quality data across cities with measurements of different air parameters like Carbon monoxide, Nitric oxide and Sulphur dioxide that will give a score based on all these parameters. With this historical data we want to forecast AQI in a certain city at a certain day. To do this we will give as input to our model all the information about values of the pollutants we want to study for each city and day, and the output will be the prediction of that values for the month after the provided values during training.

3 Research questions

- **Are there big difference between AQI during weekdays and weekend? If so, it happens the same in very city?**
- **Does AQI predictions accurate enough? In which cases the system struggles more to do accurate predictions? Why?**

4 Methodology

We will use Air Quality data for this project, where the goal is to make accurate predictions. The data that will be used can be downloaded from Kaggle (link in Section 6). These data is public domain so everyone is able to use it.
In order to implement accurate Air Quality predictions using STDN system developed in original paper [1], main steps are as follow:

- **Input: The data levels of CO, PM2.5, O3,NO2, SO2, PM10 and NO for various stations across 26 cities in India**
- **Pre-processing: For each day, format of data should be such that the target region i and it surrounding neighbors is considered as a S $\times$ S image with 7 channels.**
- **Local CNN to model the spatial dependency**
- **Long ShortTerm Memory (LSTM) network to capture the temporal sequential dependency**
- **Output: The output will be the log files with the AQI predictions for each city made during test and the corresponding loss values.**

5 Evaluation approach

- **Metrics: Given that we face a similar problem to the one presented in the original paper, we are going to use the same evaluation metrics. These are Mean Average Percentage Error (MAPE) and Rooted Mean Square Error (RMSE).**
- **Baselines:**

  We will evaluate the performance of our system comparing our predictions to the ones presented in this other system [2].

6 Data sources and other resources

- **Air quality data (daily) for 26 cities in India from 2015-2020.**
  https://www.kaggle.com/rohanrao/air-quality-data-in-india
  https://cpcb.nic.in/
- **Code and data of reference paper [1]**
  https://github.com/tangxianfeng/STDN

As explained before we are just going to use the most relevant values from the dataset, and the distribution for train and test will be 60 days (2 months) fro training and 30 days(1 month) for testing.

Ethical statement

The work that will be done is this project has no ethical considerations. It is a study of the air quality in India and it's aim is to be able to predict accurately future values for a specific time and location. There are no privacy concerns as this data has been made publicly available by the Central Pollution Control Board for everyone.
Our work cannot be use in a harmful way, but it can be used by companies and governments to know which are the more affected areas and create solutions to reduce pollution and improve the quality air in those areas.

Division of workload

For this project we will do every task together because it's very difficult to clearly divide the task. At a high level these are the tasks we are going to go through it:

- **Modify the file that load the data and process it to feed it to the model.**
- **Adjust loss functions to work accurate with our input data.**
- **Adapt output log files to include the predictions and respective loss values.**
- **Run the experiments.**
- **Compare results with the other paper mentioned above.**
- **Write the final report.**

Code

- **Link to the Github repository:**
  https://github.com/videscar/Urban-Computing-Project

References

[1] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng and Zhenhui Li. "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 5668–5675. URL: https://arxiv.org/pdf/1803.01254.pdf.

[2] GeorgeCodeHub. *Analysis and prediction of air pollution using BiLSTM Conv1D*. https://github.com/GeorgeCodeHub/Analysis-and-prediction-of-air-pollution-using-BiLSTM-Conv1D. 2021.