Real-Time Air Quality Forecasting Using a Long Short-Term Recurrent Neural Network

Alan Uthuppan

Temple University

CIS 4523: Knowledge Discovery and Data Mining

Professor Zoran Obradovic

May 2nd, 2024

Abstract

LSTM networks are used to predict particulate matter concentrations in real-time in urban environments. This study focused on Philadelphia due to the critical impact of urban air quality on public health. The LSTM network was used to predict air quality concentrations over 24-hour and 2-hour time periods. Data sources for the study included OpenAQ data for air quality, OpenWeatherMap data for weather conditions, with both datasets spanning from January 1st, 2023 to January 1st, 2024. Scraping issues such as API rate limitations and computational constraints significantly impacted the scope of the dataset. The performance of the models was measured by metrics such as MAE, MSE, RMSE and R-Squared. While the 24 hour forecast model had issues with precision, the 2 hour model showed strong predictive capabilities. The difference in accuracy between the two models highlights the need to incorporate more data features and more complex modeling techniques for future work.

Introduction

Air quality is a major environmental concern that has a significant impact on public health, ecosystems, and the economy. Especially in urban areas, where industrial activity and traffic emissions are high, air pollution is a major risk factor for a litany of negative health issues. Therefore, air quality monitoring and prediction is a key component of public health policies and individual healthcare.

Air pollution is made up of particulate matter and gasses that are harmful to human health and to the environment. Particles are defined by their diameter for air quality regulatory purposes. Particulate matter is made up of tiny particles that are small enough to reach deep into the lungs or even enter the bloodstream and cause various health problems. Specifically, those with a diameter of 10 microns or less are inhalable into the lungs and can induce adverse health effects. Particulate matter PM2.5 and PM10 are two of the most significant pollutants (see Figure I). Vehicle emissions, industrial emissions, and natural events such as wildfires and volcanic activity are some of the sources of PM2.5 or PM10 (California Air Resources Board).

In Philadelphia, the geographical focus for this study, air quality varies greatly due to its dense population and industrial activity. Monitoring these changes is crucial for foreseeing public health risks and making informed policy decisions.

Traditional ways of monitoring air quality consist of collecting data from different monitoring stations and reporting this data at regular intervals. This approach does not provide real-time intervention or warnings to the public that air quality conditions are deteriorating. However, with the help of machine learning, air quality can now be forecasted in real-time. This can improve the timing of public health responses and improve environmental planning, offering a proactive approach rather than a reactive one to managing air quality.

The motivation for the goals of this project stems from issues in the researcher's personal healthcare, specifically with them having struggled with severe asthma as a child. Sensitivity to air pollutants makes it more complicated for those with preexisting respiratory difficulties to visit and live in large cities where air quality conditions are often poor. This is the base issue this project aims to solve, as creating predictive models that can deliver real-time and accurate air quality forecasts could potentially open up several new avenues for those living with cardiovascular issues.

This project explores the application of Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, to forecast concentrations of PM2.5 and PM10 in Philadelphia for the subsequent 24-hour and 2-hour time periods.

Related Work

Current Work in LSTM for Air Quality Prediction

Recent advances in air quality prediction have largely focused on the use of LSTM networks. These networks are particularly well-suited in predicting time-series data, which is crucial in environmental studies. For example, Chang et al, (2021) and Zhang et al., (2021) showed that sophisticated LSTM models can improve air quality forecasting accuracy by capturing temporal dependencies in atmospheric datasets, along with additional weather data as well.

Zhu et al., (2021) also looked at modeling techniques with LSTM networks and discovered that they provided a more detailed analysis of pollutant concentration patterns, increasing predictability compared to conventional models.

Challenges and Improvements in Machine Learning Approaches

However, while LSTMs and other ML models have demonstrated great potential in environmental applications, according to Jiang et al, (2021), they often struggle with overfitting and as a result end up with models that are not generalizable. To address this issue, researchers such as Fu et al, (2021), Zhang et al, (2021), and Xiao et al, (2021), discussed improvements such as the incorporation of ensemble methods and regularization techniques to increase model robustness and minimize error rates. These techniques not only reduced the likelihood of overfitting, but also enhanced the models' ability to generalize across different locations. Computational requirements were an additional limiting factor due to the need to preprocess large amounts of data.

Gaps in Current Research and Contributions of This Study

Despite significant progress, there still exist gaps in the current research, especially when it comes to using predictive models in real-time for diverse urban environments (such as Philadelphia). Karimian et al, (2021) and Chang et al., (2021) point out that there is a lack of large-scale studies that incorporate different environmental factors such as weather variability and urban topography that affect pollutant behavior. In this study, we attempt to use LSTM models that combine real-time air quality and weather data from two sources to more accurately predict PM2.0 and PM10 levels.

Methodology

Data Acquisition

Datasets on air quality and weather from OpenAQ and OpenWeatherMap respectively were manually scaped and transformed using Python and its Pandas, requests, and NumPy libraries. The air quality data, initially intended to be sourced from real-time scraping via

OpenAQ's API, encountered limitations due to API license constraints, leading to decrease in the volume of data samples. Consequently, the air quality data was only present from the time period of 01/01/2023 to 01/01/2024, significantly limiting the capabilities of the LSTM model. Weather data collection faced similar API rate limits, prompting the researcher to instead purchase a historical dataset from OpenWeatherMap which included various meteorological variables and matched the time frame of the air quality data.

Data Preprocessing

The air quality data was first standardized by converting the “Timestamp” column into pandas DateTime format. Rolling 24-hour averages for PM2.5 and PM10 were calculated. Additionally, rolling 24-hour averages and lagged features for both PM2.5 and PM10 were calculated so the model could use past data points in forecasting future pollution levels. Finally, change features were added to record differences between consecutive timestamps.

For the weather data, preprocessing started with removing duplicate entries to ensure data alignment with the air quality data. Missing “visibility” values were linearly imputed, and zeroes were filled in for absent “wind_gust” data values to maintain consistency across records. Temperature readings originally in Fahrenheit were converted into Celsius to standardize measurements. The weather dataset was further processed by aligning and combining with the air quality data based on matching timestamps, removing the timestamps from the weather data that did not appear in the scraped air quality data. This ensured that each weather observation correlated directly with its specific air quality measurements.

The combined dataset was finally processed by converting all columns to float type to standardize the format of the data. Normalization was applied across the dataset to scale all values appropriately, necessary for the improvement of LSTM models. To accommodate the

LSTM networks' requirements, missing timestamp rows were reintroduced and filled with zeros, allowing the model's masking layer to appropriately ignore these values during training. This produced a final dataset that aligned with the necessary requirements of the models' learning processes. See Figure II for the full, resultant dataset and all of its features.

Model Development

Two models were implemented using the TensorFlow and Keras libraries, each aiming to predict PM2.5 and PM10 levels for the upcoming 24-hour and 2-hour intervals for each respective pollutant. The previous two weeks (336 hours) were used to forecast the 24-hour pollutant values, and the previous 24 hours were used to forecast the 2-hour pollutant values. As it was the researcher's first time implementing an LSTM, this phase encountered unexpected delays due to unanticipated model training times, delaying overall model development. These time constraints resulted in the absence of adjusting hyperparameters and tuning loss functions within the project's timeline, which could have potentially enhanced model performance. See Figure III for the full LSTM implementation.

Model Evaluation

The evaluation of the models was conducted through metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), alongside visual assessments. Predictive accuracy was represented through graphs plotting predicted versus actual pollutant levels for both PM2.5 and PM10, with both forecasting time periods accounted for. This offered a direct visual comparison of the models' performances. Error distribution plots were also generated to assess the variance in the predictions.

Results

Model Performance\

	MAE	MSE	RMSE	R^2
24 Hour Forecast				
PM2.5	0.01376	0.00030	0.01733	-1.02327
PM10	0.02960	0.00149	0.03858	-0.09490
2 Hour Forecast				
PM2.5	0.04189	0.00530	0.07283	0.89342
PM10	0.04883	0.00462	0.06796	0.88467

Visualization of Results

The results of both models were further visualized to provide a clearer understanding of their performance across the study period. Figures IV and V display the prediction accuracy and error distribution graphs for PM2.5, whilst Figures VI and VII concern PM10 values.

Similarly, the prediction accuracy and error distribution for PM2.5 values for the 2-hour forecast model is visualized by Figures VIII and IX, whilst Figures X and XI examine the forecasted results for PM10 values.

Discussion

Model Performance

The 24-hour forecast model showed less-than-ideal performance, particularly for PM2.5, with a negative R^2 values of -1.02327 and -0.09490 for PM2.5 and PM10 respectively. This indicates that the model could not accurately forecast air pollution concentration values, possibly

performing worse than even a simple baseline model that would predict the mean value at all times.

Since air quality can be influenced by a range of factors including sudden changes in weather, traffic patterns, and unforeseen events like wildfires, the 24-hour model might not capture these events adequately, especially with the limited time frame of the training data. Given the high frequency of data needed to predict air quality over a full day, the model may have overfitted to noise rather than capturing the underlying patterns effectively. Additionally, while the model used lagged and rolling average features, it may have benefited from more complex feature engineering, such as capturing cyclical daily patterns or integrating other types of features, such as traffic and industrial emissions data.

Contrastingly, the 2-hour forecast model showed much stronger performance, with R^2 values of 0.89342 for PM2.5 and 0.88467 for PM10. These values suggest that the model was successful in capturing a substantial portion of the patterns that significantly affect air quality over shorte

This model's better performance could be attributed to a number of reasons. In general, shorter-term predictions typically deal with less variability and are less likely to encounter changes in input variables. This allows the 2-hour model to make more accurate forecasts based on past values that are more indicative of immediate future conditions. The 2-hour model may also be better suited to respond to rapid temporal changes in pollutant levels, given its training data is more relevant and dense, providing it with a narrower scope to accurately forecast near-term changes.

The stark difference in performance between the two models highlights the importance of choosing appropriate model configurations based on the time interval and feature characteristics.

The poor performance of the 24-hour model could potentially be fixed by incorporating more complex data features or employing advanced regularization techniques. Furthermore, augmenting the models' inputs with a larger dataset (in terms of both number of samples and features) that capture sudden environmental changes could possibly increase its prediction accuracy.

Future Research

Regarding future research directions, implementing hybrid models that combine LSTM with other predictive techniques, such as convolutional neural networks that can process spatial features, may provide more accurate predictions. Moreover, implementing ensemble methods that aggregate predictions from multiple models could potentially combat individual model failures and result in a more generalized model with better performance.

Acknowledgements

I would like to express my sincere gratitude to Professor Zoran Obradovic, whose expertise in data modeling and coursework provided a solid foundation for my research. I would also like to thank Hussain Otundi for his patience and readiness to assist with any questions regarding the course material, which enhanced my project's execution.

Furthermore, I extend my appreciation to Temple University for access to their research database and computational resources.

Lastly, I would like to thank my peers and colleagues who offered their feedback and moral support, which helped refine the project's work and maintain a steady level of motivation.

References

California Air Resources Board. (n.d.). California Air Resources Board. Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board. https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=Particles%20are%20defined%20by%20their,diameter%20(PM2.5)

Chang, H., Karimian, H., & Zhu, X. (2021). Advanced LSTM Models for Air Quality Forecasting in Urban Environments. Journal of Environmental Sciences, 42(1), 104-112.

Fu, H., Jiang, M., & Sulian, T. (2021). Enhancing Air Quality Predictions Using Machine Learning: A Review of Models and Key Findings. Environmental Modelling & Software, 141, 104948.

Jiang, H., Fu, R., & Karimian, H. (2021). Machine Learning in Air Quality Forecasting: Challenges and Opportunities. Atmospheric Environment, 244, 117834.

Karimian, H., & Chang, H. (2021). A Comparative Study of Machine Learning Models for Air Quality Prediction. Science of the Total Environment, 765, 142810.

Zhu, X., Chang, H., & Fu, H. (2021). Refined Modeling of Air Pollutants Using LSTM Networks. Journal of Cleaner Production, 291, 125233.

Supplemental Materials:

https://drive.google.com/drive/folders/151-WWf5C82rLwERKpVuNiTN5zYRlnObF?usp=drive_link