Using Recurrent Neural Networks to Predict Air Quality

Devin Kulatunga, Daniel Amran, Zachery Abdallah, Nouhad Rizk

Department of Computer Science, University of Houston

Abstract

Air quality becomes a greater factor towards a person's health closer to dense population centers. An increase in human activity generates a greater amount of waste which some individuals need to take steps to avoid. Concentrated areas are not direct indicators of the Air Quality Index, however they suggest a higher likelihood of man-made particulates being produced. This article features machine learning algorithms used to predict the Air Quality Index (AQI) over time in localized areas. This data not only keeps the population one step ahead on weather conditions, but also helps data analysts and scientists glean into what its correlations and potential effects are. The purpose of studying the Air Quality Index (AQI) is to give people control of this detrimental impact to their well-being. To predict a value in the future with respect to the past, a time-series analysis model like the Recurrent Neural Network (RNN) is the most appropriate to use for this experiment.

Background

One of the growing concerns of pollution is its influencing factor towards climate change. Many population centers experience some degree of air pollution which carries health risks to people as well as affecting the environment around them. Much like how pesticides can affect the ecology of an area and contaminate the foods people eat, air pollution can saturate the body with toxins which can lead to heart disease, cancer, and lung disease. Any sort of exposure could have an influence on an individual's day-to-day life.

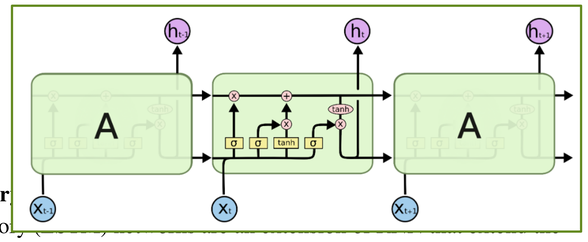
Air Quality Index (AQI) Values	Levels of Health Concern
When the AQI is in this range:	*...air quality conditions are:*
0 to 50	Good
51 to 100	Moderate
101 to 150	Unhealthy for Sensitive Groups
151 to 200	Unhealthy
201 to 300	Very Unhealthy
301 to 500	Hazardous

The primary motivation for these experiments is to observe the effects industrialization has had on the environ--ment and its inhabitants. Studying emissions is essential to understanding the impact of geological and scientific processes. Air pollution data helps data analysts and scientists glean into what the environmental and potential health effects are and also keeps individuals informed through weather applications. Air pollution affects everyone on an individual level and should be monitored and predicted. The main goal of these experiments is to predict future air pollution trends a week in advance using previously recorded data.

Methods

RNN Recurrent Neural Network for Forecasting

- Since the data selected was recorded in a time series order we concluded that using a Recurrent Neural Network (RNN) most appropriate for accomplishing our task.
- RNNs work by retaining a memory of what it has already processed and then learn from previous iterations or sequences during its training. At each element of the sequence, the model considers not just the current input, but what it remembers about the preceding elements which is important for predicting future sequences.

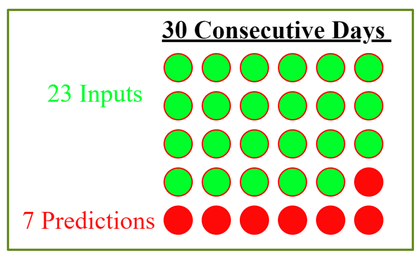


Long short-term Memor

- Long short-term memo memory.
- LSTMs also reduce the vanishing gradient problem with LSTM cells, which are designed to learn important information and forget less important.

Time-Based Windows

- Time-series data calls for spliced windows to feed into the RNN.
- In order to create a training dataset a loop was used to find a consecutive 30 days of recorded data from between 2000-2020 for each city.
- A testing set was then created for dates after 2020
- The sum total of the train/test and predictions is 30 which is supposed to resemble the days of one month.



Implementation

Implementing LSTM

- When implementing the model it was determined that in order to optimize runtime and accuracy using a 2 layer LSTM model was the most optimal.
- For the activation function relu was the best at avoiding negative results.
- 2 dropout layers were added with a value of 20% to avoid overfitting within the model.

Results

Results

- What is surprising is the accuracy did not lower significantly despite the predictions for days were going further from the training portion.

RMSE day 1: 4.490280417065027
RMSE day 2: 4.567630252350579
RMSE day 3: 4.435634553963542
RMSE day 4: 4.637608755374881
RMSE day 5: 4.588242820812866
RMSE day 6: 4.454803899876893
RMSE day 7: 4.52112995892734

Model Variation

- To try increase the model's accuracy we created a second variation of the model which only predicted one day in advance and then adds that predicted data back into the dataset and repeat it for 7 days.
- However this variant received a larger RMSE which indicated a worse model.

RMSE day 1: 5.19365279743138
RMSE day 2: 5.182089040854947
RMSE day 3: 5.352778805679363
RMSE day 4: 5.167112635293334
RMSE day 5: 5.229091275858283
RMSE day 6: 5.288437938393027
RMSE day 7: 5.363647589066005

Conclusion

- To reiterate, air quality is an important feature of everyday life. There are many factors that contribute to the quality of the air you breathe. Although impossible to change on your own, the ability to avoid unhealthy air would go a long way in helping an individual to improve their health.
- The results indicate that Air Quality Index (AQI) can be predicted for the future with sufficient accuracy by identifying past AQI trends of a particular region.

Future Direction

- Optimizing Pre-Processing of Data
- Retraining with differing Time-Based windows as needed for different applications
- Comparing and contrasting performance with a Multiple-output Support Vector Regression (MSVR) model

Acknowledgments

Wang, J., Li, X., Jin, L. *et al*. An air quality index prediction model based on CNN-ILSTM. *Sci Rep* 12, 8373 (2022). https://doi.org/10.1038/s41598-022-12355-6

Reigada, Caleb (2022, June). US Air Quality 1980-Present, Version 1. Retrieved November 17, 2022 from https://www.kaggle.com/datasets/calebreigada/us-air-quality-1980present