Air Quality Predictions in Pakistan using Neural Networks

Mudassir Hanif Sheikh
Department of Computer Science
Habib University
ms03831@st.habib.edu.pk

Rayyan-Ul-Haq
Department of Computer Science
Habib University
ru03588@st.habib.edu.pk

Rizwan Niaz Khajar
Department of Computer Science
Habib University
rk03548@st.habib.edu.pk

Abstract

In modern society Air pollution is a problem whose effects endanger the lives of millions of people all across the world. One measure of Air quality is is $PM_{2.5}$ or Particulate Matter 2.5 which is the term used to describe solid particles that are less then or equal to 2.5 micrometers in width. These Particles are suspended in the air and can cause a host of health problems. This paper aims to create a $PM_{2.5}$ forecasting system using Recurrent Neural Networks, these Neural Networks can use past meteorological data to make predictions of future $PM_{2.5}$ levels up to a week in advance.

1 Introduction

In the past few decades, Asian countries have seen a rapid increase in motorization, urbanization and energy expenditure. The increased number of sources of air pollution, coupled with large population densities and lack of proper air quality management infrastructure has resulted in Air pollution being a serious health hazard for people throughout the continent.

One of the measures of air pollution is Particulate Matter 2.5 or $PM_{2.5}$ this refers to solid particles that have a width of 2.5 micrometers or smaller. The sources of these particles can be the burning of fossil fuels, industrial processes, dust, smoke etc. These suspended particles in the air can severely damage a persons respiratory and cardiovascular symptoms, leading to diseases such as lung cancer and asthma. It is therefore imperative to create proper $PM_{2.5}$ management programs to ensure public health. [1]

Pakistan in particular suffers greatly from the effects of $PM_{2.5}$, with approximately 35% percent of the population living in urban areas, it is the most urbanized country in all of South Asia. In 2006, the Pakistan Strategic Country Environmental Assessment world Bank Report stated that Particulate matter was responsible for 22000 deaths among adults and 700 among children, this also puts a burden on the economy, as $PM_{2.5}$ related health issues account for 1% of the gross domestic product

The main contribution of this paper will be to help in the creation of a proper $PM_{2.5}$ management system as before this paper, there has been no other work done in creating a $PM_{2.5}$ prediction model for the country. [2]

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 The Data

The Data Set used in this Paper contained meteorological data acquired from the Pakistan Air Quality Initiative (PAQI). The Data set consisted of hourly readings of $PM_{2.5}$ levels, USAQI (United States Air Quality Index) levels, CO_2 levels, Temperature readings in Celsius, Relative humidity levels, Wind Direction, Wind Speed, and Visibility levels.

2.1 Preprocessing

In order to better understand how time affected the meteorological variables as well as PM2.5, we started off by specifying the month, the hour, the day of the month and the day of the week the readings were taken on. The Data set included data from several key cities across Pakistan Including, Lahore, Karachi and Islamabad. For this paper, we chose to focus on the city of Karachi. In Karachi the data was taken from several different areas, namely: DHA (Defense Housing Authority), North Karachi, Landhi, Korangi, J-Area Korangi, Saddar, Hyderi Market, Bahadurabad, Karachi Industrial area, Shersha, Liaquatabad, and Civic Center. The data set also required some additional work to be done on it before it could be usable, this was because the data set itself was incomplete, and there were several instances where days and even weeks worth of readings were missing in certain areas. To solve this, the mean values of all the *other* areas was used to fill in the gaps. However, there were still some instances where there were no readings in any area. In this case it was impossible to take the average, so instead the mean of the *previous* 24th time step and the *next* 24th time steps were used to infer the missing data (time steps here refers to an hour i.e the readings were taken every hour, so "previous 24th time step" refers to the previous 24th readings from the missing data, which corresponds to 24 hours ago).

2.2 Normalization

When different attributes have different scales of readings, this usually causes problems as the Neural Network often fails to converge. This is because Neural Networks use gradient descent which may fail to converge if data is not normalized [4] So we converted our data into standard normal by doing the following conversion. [4]

$$X' = \frac{X - mean(X)}{std(X)}$$

2.3 Time Series

We finally did some work to convert the data into a standard time series format. Since we used recurrent neural networks, we sorted our data according to date and time [5]. We arranged the data into τx14 matrix, where 14 is the number of attributes we had whereas τ is the number of time steps considered for predicting the next PM2.5 value.

3 Methodology

In order to create a working forecast system two different types of recurrent neural networks were used. A Gated Recurrent Unit neural network, (GRU) and a Long Short Term Memory (LSTM) neural network. One of the key differences between the two types of neural networks is that GRUs require less computational power than LSTMs with the caveat that they have less accuracy for large sequences of data. Besides this, univariate GRU and LSTM models where only past $PM_{2.5}$ values were used to predict future $PM_{2.5}$ were also tested. The models present in deep learning literature for the same problem are more sophisticated in the sense that they often use a hybrid approach [3, 5 and 6]. However, for this paper we adhered to the basics, and tried basic versions of the models instead of the hybrid approaches.

3.1 Gated Recurrent Unit (GRU)

The GRU model as trained for 10 epochs, it had a batch size of 4, the adam optimizer was used, the learning rate was set to 0.0005, the activation function was $tanh$ and $L2$ loss function was used.

3.2 Long Short Term Memory (LSTM)

The LSTM model was also trained for 10 epochs, it had a batch size of 4, the adam optimizer was used, the learning rate was set to 0.005, the activation function was $tanh$ and once again $L2$ loss function was used.

3.3 Univariate GRU

The Univariate GRU model was trained for 10 epochs, it had a batch size of 4, the adam optimizer was used, the learning rate was set to 0.0002, the activation function used was $tanh$ and the $L1$ loss function was used.

3.4 Univariate LSTM

The Univariate LSTM model, was trained for 10 epochs, it had a batch size of 12, the adam optimizer was used, the learning rate was set to 0.01, the activation function used was $tanh$ and the $L1$ loss function was used.

4 Tests and Results

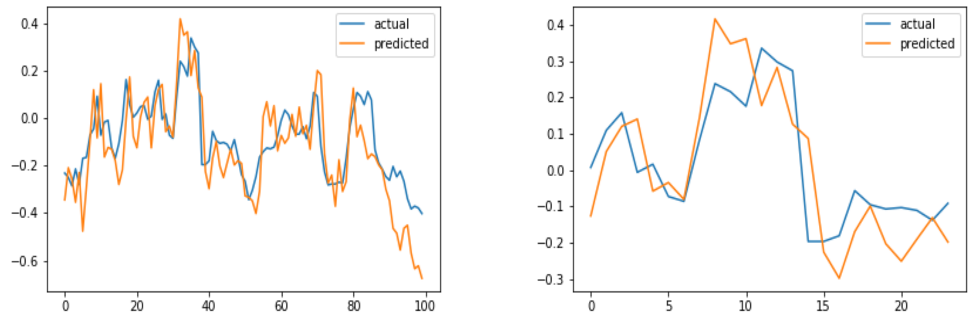
One of the questions that puzzled us and for which we had to do a lot of experimentation was that of a good size for a time step. The literature had different recommendations and different results. For example, [3] used 24-hour time step, that is, they used 24-hour values in the past to predict the current PM$_2.5$ value. Whereas [5] used several different values for the time step, including 2, 4, 8, 24, 48 and 72. The results showed that 72 performed the best. Following their footsteps, we went one step ahead and tried for all values going from 1 time step to 72 time steps, but our model stopped improving as we went past 24 from which we concluded that our models performed the best when taking only the past day's data into consideration.

The non univariate GRU and LSTM models were subject to the same tests. The first test was to see how well the models made predictions when the actual meteorological data was provided for all the time steps. The second test was to see how well the models made predictions when only the meteorological data was available but previous PM2.5 ground truth values were not and instead the predicted results of PM2.5 were used to make even more predictions. These two results were then plotted and compared with the actual data. For the Univariate GRU and LSTM models only the first test was performed where the actual PM$_{2.5}$ value at each time step was used to predict the next PM$_{2.5}$ value. All of the graphs were shown for 100 time steps, corresponding to the X-axis, with 1 time step corresponding to an hour. The value on the y-axis is the normalised value of the PM$_{2.5}$. The blue line represents the actual values and the orange line represents the value predicted by the model.

4.1 GRU

After 10 epochs the final loss was 0.4196364. two tests were performed on the model, one where the actual data was used to predict the next hours PM$_{2.5}$ level and one where only the first 24 hours worth of actual data was used, and for the rest of the time steps, the predicted result was used to predict the next PM$_{2.5}$ value

4.1.1 Test 1: Using measured data for all predictions.



(a) A plot of the GRU models predictions

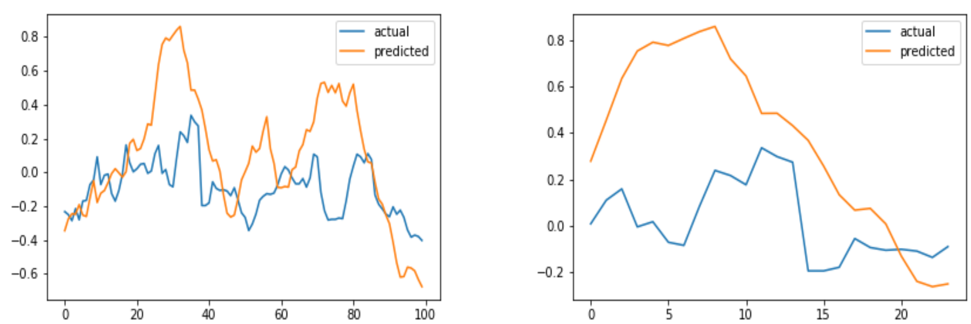
(b) A zoomed in look at the first 24-48 hours of the figure on the left

Figure 1: GRU results using measured data, normalized PM2.5 values vs hours

Figures 1 show the plots of the results of the models predictions when using measured data. The figure on the right provides a closer look into the predicted values.

4.1.2 Test 2: Using predicted values for predicting the next value

Figures 2 show the plots of the results of the models predictions when using predicted values of $PM_2.5$ to make the next prediction. The figure on the right provides a closer look into the predicted values.



(a) A plot of the GRU models predictions

(b) A zoomed in look at the first 24-48 hours of the figure on the left

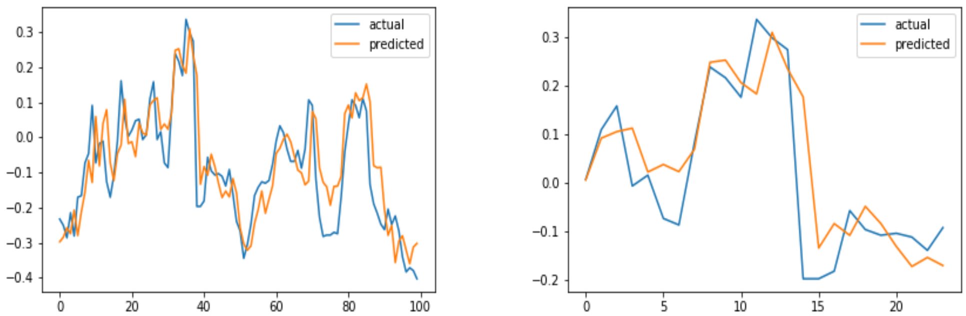
Figure 2: GRU prediction using predicted data, normalized PM2.5 values vs hours

4.2 LSTM

For the LSTM, the final loss after 10 epochs came out to be 0.41297793. The same tests as the GRU were performed on it.

4.2.1 Test 1: Using measured data for all predictions.

Figures 3 show the plots of the results of the models predictions when using measured data. The figure on the right provides a closer look into the predicted values.



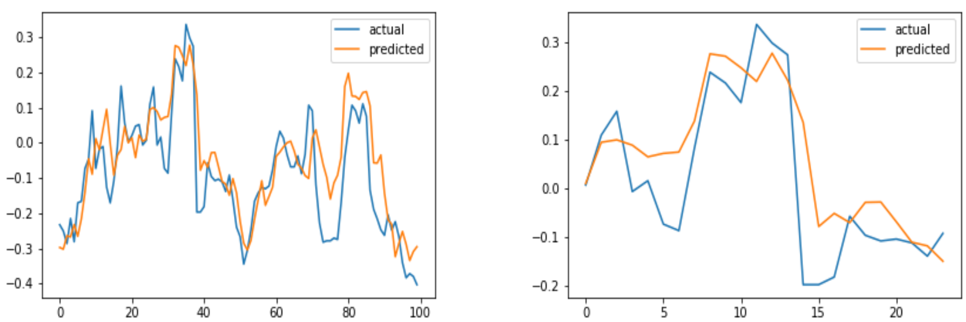
(a) A plot of the LSTM models predictions

(b) A zoomed in look at the first 24-48 hours of the figure on the left

Figure 3: LSTM predictions using measured data, normalized PM2.5 values vs hours

4.2.2 Test 2: Using predicted values for predicting the next value

Figures 4 show the plots of the results of the models predictions when using predicted values of $PM_2.5$ to make the next prediction. The figure on the right provides a closer look into the predicted values.



(a) A plot of the LSTM models predictions

(b) A zoomed in look at the first 24-48 hours of the figure on the left

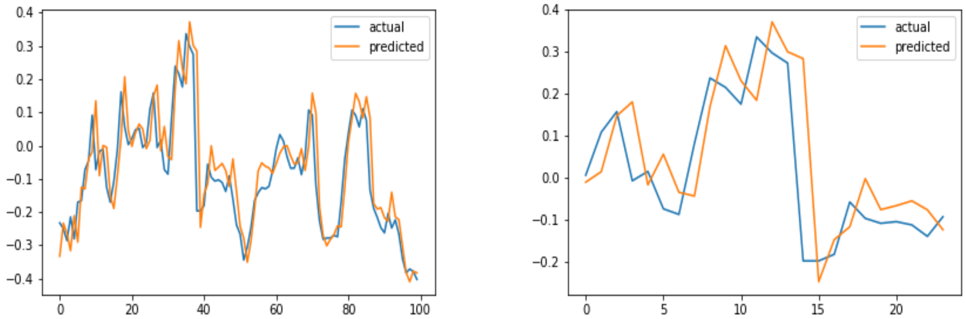
Figure 4: LSTM predictions using predicted data, normalized PM2.5 values vs hours

4.3 Univariate GRU

After being trained for 10 epochs, the Univariate GRU had a loss of 0.42902866. When we performed Test 1 the results as shown in figure 5 seemed promising, but when we attempted Test 2, the results as shown in figure 6 were quite disappointing.

4.4 Univariate LSTM

The univariate LSTM performed even worse than the Univariate GRU, as its final loss was 0.43344387 after 10 epochs. Since, the loss was even higher than the Univariate GRU model, only the first test was performed on it, that is, we only used the measured data to make predictions. Figure 7 shows the results.



(a) A plot of the Univariate GRU models predictions (b) A zoomed in look at the first 24-48 hours of the figure on the left

Figure 5: Univariate GRU predictions using measured data, normalized PM2.5 values vs hours

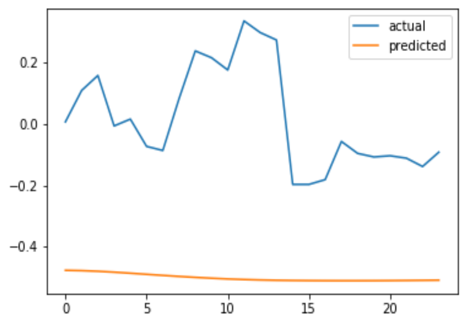
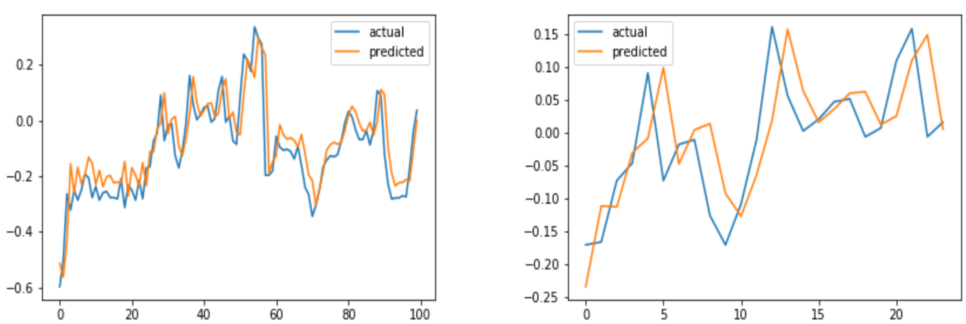


Figure 6: A plot of the Univariate GRU models predictions when using predicted values for all predictions, normalized PM2.5 values vs hours. (only the first 24-48 hours are shown)



(a) A plot of the Univariate LSTM models predictions (b) A zoomed in look at the first 24-48 hours of the figure on the left

Figure 7: Univariate LSTM predictions using measured data, normalized PM2.5 values vs hours

4.5 Observations and Discussion

The results of the 4 models being run for 10 epochs on the same data set yielded the following losses.

Model	L1 loss	L2 loss
GRU	-	0.4196364
LSTM	-	0.4129779
Univariate GRU	0.42902866	-
Univariate LSTM	0.43344387	-

Table 1: Losses for models on measured values on test data

One interesting observation, which may not seem obvious at first, was that the Univariate models did not fare well for predictions that do not have the ground truth values available to them, that is, the Univariate models only worked on measured data and performed very poorly on the predicted data. This is something new that we tried relying on the fact that $PM_2.5$ values follow a pattern and that they depend on their previous values. However, the results were not that successful and the more conventional models present in literature, that is, multivariate LSTM and GRU performed better.

From our losses and the results of our tests, we can see that for the multivariate (non-univariate) case, the LSTM model performs better as expected, while for the univariate models, the GRU performs better. However when comparing all 4 of the models we can see that the multivariate LSTM model performs best for the task at hand, which is making accurate future predictions of the $PM_2.5$ values far in advance when measured data is not available.

Despite the fact that we had very limited data and our dataset was not very mature, we were able to produce results using simple models that are comparable to those present in literature which use more complex models as well as a more mature data.

5 Future Directions

There are a number of future directions this project can be taken in, the first would be to design forecast systems for other major cities in Pakistan, namely Lahore, as such a system would be most beneficial there. Since this is the first time a study such as ours is being carried out in the country, there is a lot of work that could be done. For example the data that we used is in its initial phases, and had a lot of issues that we discussed. As we move forward, the data will be more mature and greater in numbers which will help us train better models and predict better.

Another direction we could take is to use a hybrid approach (One that utilizes a number of different neural network types) to create a more accurate forecast system, as they seem to produce better results[3]. Finally, yet another direction we could take this project is to try and understand the reason for the results we see, i.e understand how the various meteorological factors interact with each other to determine the $PM_{2.5}$ values on any given day using statistical analysis which will help us build a very reliable system for $PM_{2.5}$ forecasting.

References

[1] Monks, PS. & Granier, C. & Fuzzi, S. (2009) Atmospheric composition change – global and regional air quality *Atmospheric Environment, 43(33).*

[2] Colbeck, I. & Zaheer, A.N. & Zulfiqar, A. (2009) The State of Ambient Air Quality in Pakistan - A Review. *Environmental science and pollution research international*.

[3] Huang, C.J. & Kuo, P.H. (2018). A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. *Sensors* .

[4] Jayalakshmi, T. & Santhakumaran, (2011). A. Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng. 3*, 89–93

[5] Y. Tsai, Y. Zeng and Y. Chang. (2018) Air Pollution Forecasting Using RNN with LSTM, *IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing*.

[6] B. Wang, W. Kong, H. Guan and N. N. Xiong, (2019) Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things, *in IEEE Access, vol. 7, pp. 69524-69534*.