Sensor-Based Air Pollution Prediction Using Deep CNN-LSTM

Kabir Nagrecha*, Pratyush Muthukumar *, Emmanuel Cocom*, Jeanne Holm**, Dawn Comer**
Irene Burga**, and Mohammad Pourhomayoun *
* Department of Computer Science, California State University Los Angeles, Los Angeles CA,
[knagrec2, pmuthuk2, mpourho]@calstatela.edu,
** City of Los Angeles, Los Angeles CA, [jeanne.holm, dawn.comer, irene.burga]@lacity.org,

CSCI-ISAI Short Paper

***Abstract*—The devastating impacts of air pollution have become more and more evident in recent years. As our measurement technologies improve, we gain better insight into the true impact of this deadly, yet often ignored, threat. The first step in reducing the damages caused by this problem is being able to analyze and predict its patterns. The problem of predicting air quality and the presence of particulate matter lies in the nature of the data needed to create an accurate system. The sheer number of factors affecting air quality mean that previously proposed approaches often utilize a great many sources of data, aiming to incorporate images, wind graphs, traffic information, and more. Yet in truth, most areas outside large metropolises lack ready access to high-quality data, preventing them from ever implementing an effective system. We propose a system utilizing a 1-D deep convolutional neural network to analyze past sensor readings and predict air pollutant concentrations up to a day in the future at a 3-hour resolution. We specifically developed this model for predicting PM2.5 values. The system receives PM2.5 sensor values and discovers temporal pattern in the data, which will be later used for prediction. By removing the dependency on complex data inputs, the system becomes accesible and easily implementable for any region. Despite this simplified approach, the results are comparable to — and often better than — any current state-of-the-art predictive systems in this domain.**

***Index Terms*—air pollution prediction, deep CNN-LSTM, convolutional neural network, low-cost prediction, atmospheric air pollution**

I. Introduction

It is impossible to overstate the need for an accurate air pollution prediction system. Millions die annually to this pervasive danger, with young children being particularly badly affected [1] [2]. Air pollution has also taken their toll on local economies, with California alone suffering a loss of more than $15 billion a year [3]. In order to mitigate the impacts, many researchers have attempted to build systems to forecast air pollution concentrations. However, most of these systems have largely focused on data-rich environments, where a great many sources of information can be leveraged [4] [5] [6].

We propose a universal solution; one that can be implemented anywhere with access to ground-based pollutant sensors. By recasting the gathered sensor data into a modified pseudo-image, we can utilize a Deep 1-D Convolutional Neural Network paired with a Long-Short-Term-Memory unit (LSTM) to forecast pollutant concentrations.

Deep Convolutional Neural Networks belong to a class of neural networks which utilize weighted filters to transform a given image into a new representation of the relevant information. The 1-D variant of this takes a graph of datapoints and interprets it as an image across which the filter is slid horizontally. At each index of the filter, the existing data is transformed linearly. By repeating this operation several times, we can extract a great deal of information from relatively simple inputs [7].

The output is then fed sequentially to an LSTM. The LSTM is a special type of neural architecture which is capable of learning time-based relationships by utilizing a self-feeding loop in its inner layers — thus retaining information from past inputs to incorporate into its analysis of the upcoming inputs. The LSTM is what allows us to accurate forecast upcoming air pollution levels [8].

II. Methods

A. Dataset

We utilize historical sensor data gathered across the Port of LA from 4 major sites [9]. The data includes hourly samples of 6 sensor readings, recording O3, CO, SO2, NO2, PM2.5, and PM10. Together, these 6 compose the majority of air pollution in the atmosphere.

We utilize approximately 4 years of data, which we split into a roughly 75-25 proportion between training and testing. Thus we get a thorough analysis of the performance of the system across a year in varying conditions.

B. Model Architecture and Implementation

The predictive model consists of two connected components — a Convolutional Neural Network, and an LSTM.

Convolutional Neural Networks won their fame in the domain of image processing, wherein each convolutional filter is ”slid” accross the image both vertically and horizontally, thus extracting relevant information from the pixels.

Translating this to the 1-D sensor domain requires us to create a 1-D analogue of the image.

By stacking multiple layers of convolution in series, we can generate a transformed sequence of values extracted from the original sensor inputs. This data will hold far more value for the LSTM's predictive stage.

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00127

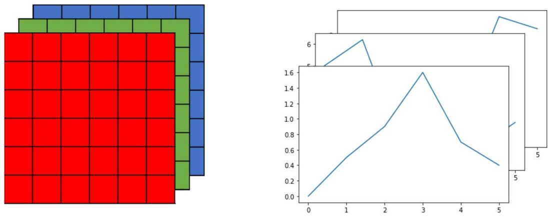

Fig. 1. An analogue between the RGB-filter 2D image used for traditional CNNs and the pseudo-image of multi-input sensor data fed to a 1D-CNN LSTM.

The LSTM is derived from the standard artificial neural network, wherein a series of neurons apply multiplicative and additive operations upon input data to generate output. The LSTM creates a new connection between each neuron's output and it's own input, moderated by a "forgetting" mechanism to prevent information overload.

These two systems are linked together so that they can be trained in series to produce predictions at a 3-hour scale for the next 24 hours — 8 outputs in all. We choose to train the network to predict PM2.5, but this approach is easily extensible to any pollutant.

III. Results and Methodology

Our CNN model was a simple 12-filter single-layer system with a size 5X1 kernel, whose output was flattened and sent to an LSTM for timeseries-based prediction. The last stage of processing involved the use of several linear layers for post-processing and output formatting. The initial layer allows the model to extract relations across a 5-timestep-span. Since our data is hourly, the initial phase of processing can only derive localized patterns at a 5-hour scale. Intra-sensor relations are transformed from our initial "6-filter" input composed of 6 sensors to a 12-filter output. Thus the CNN learns localized patterns in the data and extracts relevant information at a small scale.

This local information is then fed to the LSTM for a more "global" level of spatiotemporal analysis. The LSTM's memory cells allow it to put the local patterns in the context of historical air quality across months, or even years. It's evident from our results that this stage of processing is critical — it is the model's global pattern recognition that allows it to hold up in different seasons and weather conditions across the entire year.

Our model predicted a sliding 24-hour window of pollutant values at a resolution of 3-hours each. Each frame of output predictions refers to a 3-hour distance from the previous frame, with 8 frames in all. For the sake of clarity, we will chart a comparison between the 8th prediction (24 hour mark) against the real values at that time. The system is remarkably accurate, usually with an error in the range of 25% relative to the mean for that site. The full results are displayed in Table 1.

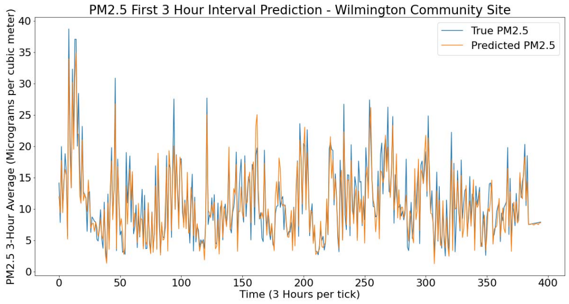

Fig. 2. 24-hour distanced prediction of PM2.5 concentrations at the Wilmington sensor site.

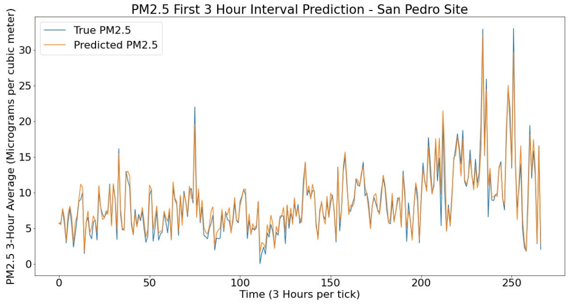

Fig. 3. 24-hour distanced prediction of PM2.5 concentrations at the San Pedro sensor site.

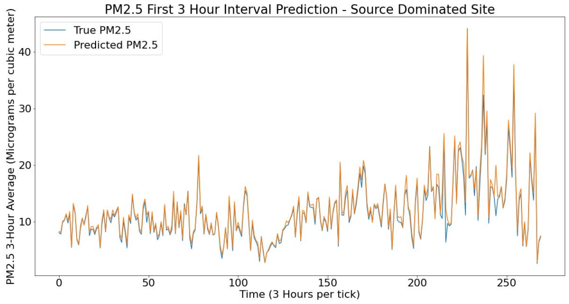

Fig. 4. 24-hour distanced prediction of PM2.5 concentrations at the Source-Dominated sensor site.

The general patterns are similar across sites, which is to be expected given their relative geographical closeness, all being within L.A. More striking is the apparent accuracy of the system, which is capable of forecasting PM2.5 values to a great degree of accuracy.

Existing systems produce similar results, but require high-dimensional and complex data input to produce them [10] [11] [12]. In comparison, these results are produced using low-cost, readily accessible data.

IV. Error Analysis

To evaluate our model's outputs relative to the labels, we utilize the Root-Mean-Squared-Error (RMSE) as a measure of prediction error. The RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2}$$

where N denotes the number of samples and x denotes the per-sample error. We analyze this error as a fraction of the average value of PM2.5 at that site during the testing period. This normalization is necessary to allow for an accurate interpretation of the results given their scale.

TABLE I
RMSE BY 3-HOUR INDEX AS A OF THE MEAN CONCENTRATION FOR RELEVANT SITE.

RMSE Values Across All Sites			
	San Pedro Site	Wilmington Site	Source-Dominated Site
Frame 1	19.46%	14.81%	20.01%
Frame 2	19.48%	15.50%	19.96%
Frame 3	21.84%	17.88%	20.51%
Frame 4	22.51%	19.84%	20.87%
Frame 5	23.65%	20.03%	21.08%
Frame 6	24.43%	21.12%	22.01%
Frame 7	24.61%	22.45%	22.54%
Frame 8	25.21%	20.98%	23.50&

V. Conclusion

It is evident from our results that this model is capable of producing highly accurate results, whilst only utilizing low-dimensionality data. This removes the dependency on possibly unavailable inputs such as satellite imagery.

Satellite imagery is often unusable due to cloud cover or other unavoidable problems. By comparison, we utilize a relatively straightforward and safe source of data with ground-based sensors. By reverting to the use of a reliable and readily accessible data source, our system can be used universally in all locales.

The accuracy of the system demonstrates that the use of this simpler data form is no compromise. This provokes a deeper question as to what data is really needed for the purposes of air pollution prediction. Reliance upon the natural encoding of environmental information into pollution concentrations does not seem to hinder our deep learning system, implying that the necessary information is still present.

This work can be used to allow cities and local governments to track and analyze the cycles of pollution in their region, giving them an advance warning on any troubling trends.

VI. Future Work

It is evident from this study that the 1D-CNN-LSTM approach is is effective and efficient in discovering and predicting temporal patterns in the data such as predicting the air pollution.

The most natural extension of this is to apply a similar approach to the 2D domain, taking satellite imagery, analyzing it, and flattening the convolved results to feed to an LSTM.

Additionally, the system as presented can be used in any locale with ground-based sensors, not just L.A. The flexibility of the general architecture also means it is possible to predict any desired pollutant even beyond PM2.5.

The most important implications of this work lie in its demonstration that deep learning can produce highly accurate results even given low-dimensionality data in this domain.

References

[1] United Nations. With a premature death every five seconds, air pollution is violation of human rights, 2019.

[2] Earl Swigert. Unicef: An urban world. Website.

[3] B. Holmes-gen and W. Barrett. Clean air future, health and climate benefits of zero emission vehicles, 2016.

[4] Pratyush Muthukumar, Emmanuel Cocom, Jeanne Holm, Dawn Comer, Anthony Lyons, Irene Burga, Christa Hasenkopf, and Mohammad Pourhomayoun. Real-time spatiotemporal air pollution prediction with deep convolutional lstm through satellite image analysis. In *Proceedings of the 16th International Conference on Data Science (ICDATA)*, 2020.

[5] Sanam Narejo and Eros Pasero. Meteonowcasting using deep learning architecture. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 8(8), 2017.

[6] Aditya Grover, Ashish Kapoor, and Eric Horvitz. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–386, 2015.

[7] Y. Bengio Y. Lecun, L. Bottou and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE Conference*, pages 2278–2324, 1998.

[8] Chiou-Jye Huang and Ping-Huan Kuo. A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. *Sensors*, 18(7):2220, 2018.

[9] Satish Vutukuru and Donald Dabdub. Modeling the effects of ship emissions on coastal air quality: A case study of southern california. *Atmospheric Environment*, 42(16):3751–3764, 2008.

[10] Tongshu Zheng, Michael H Bergin, Shijia Hu, Joshua Miller, and David E Carlson. Estimating ground-level pm2. 5 using micro-satellite images by a convolutional neural network and random forest approach. *Atmospheric Environment*, page 117451, 2020.

[11] Lianfa Li, Mariam Girguis, Frederick Lurmann, Nathan Pavlovic, Crystal McClure, Meredith Franklin, Jun Wu, Luke D Oman, Carrie Breton, Frank Gilliland, et al. Ensemble-based deep learning for estimating pm2. 5 over california with multisource big data including wildfire smoke. *Environment International*, 145:106143, 2020.

[12] Lianfa Li, Mariam Girguis, Frederick Lurmann, Jun Wu, Robert Urman, Edward Rappaport, Beate Ritz, Meredith Franklin, Carrie Breton, Frank Gilliland, et al. Cluster-based bagging of constrained mixed-effects models for high spatiotemporal resolution nitrogen oxides prediction over large regions. *Environment international*, 128:310–323, 2019.