IMAGE-BASED AIR QUALITY FORECASTING THROUGH MULTI-LEVEL ATTENTION

Tony Zhang and Robert P. Dick

Electrical and Computer Engineering Department
University of Michigan, Ann Arbor, MI, USA

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897241

ABSTRACT

The problem of air quality forecasting is important but also challenging because air quality is affected by a diverse set of complex factors. This paper describes the first image-based air quality forecasting model. It fuses a history of $PM_{2.5}$ measurements with colocated images. We construct a multi-level attention-based recurrent network that uses images and $PM_{2.5}$ data to represent variation over space and time. Experiments on Shanghai data show that our model improves $PM_{2.5}$ RMSE prediction accuracy by 15.8% and MAE by 10.9% compared to previous forecasting methods. In addition, we evaluate the impact of each model component via ablation studies.

***Index Terms*—** $PM_{2.5}$ forecasting, image analysis, data fusion, attention

1. INTRODUCTION

$PM_{2.5}$ is a pollutant consisting of particles smaller than 2.5 micrometers. $PM_{2.5}$ is especially dangerous to human health; the particles are small enough to bypass the immune system and travel in the respiratory and cardiovascular systems [1]. Because $PM_{2.5}$ is harmful and also difficult to forecast, many studies focus on $PM_{2.5}$ forecasting [2]. Time series methods such as ARIMA have been used, as has deep learning [3, 4].

Past work developed data-driven models for time-series forecasting of air quality [3]. For example, researchers designed a dual-stage attention model for time series prediction [5]. Also, deep neural networks have been used to combine multiple sources of data such as weather and geo-context data for $PM_{2.5}$ forecasting [3, 4]. However, $PM_{2.5}$ forecasting remains a challenging problem despite existing work because $PM_{2.5}$ levels are affected by many complex factors.

$PM_{2.5}$ levels at one location are affected by surrounding $PM_{2.5}$ levels, and $PM_{2.5}$ varies at small spatial scales [6]. Utilizing digital cameras and webcams is beneficial in estimating $PM_{2.5}$ concentrations at different areas of an image [7, 8]. Also, images can capture external factors correlated with $PM_{2.5}$ levels; for example, images can track meteorological conditions such as humidity and cloud cover on its own.

Images can be valuable for air pollution forecasting since cameras capture a large amount of data over large spatial regions, whereas air pollution data are commonly collected by sparsely distributed single-point monitoring stations. Different regions in a city exhibit spatial correlation of $PM_{2.5}$ levels over time, and images can help track that spatial correlation for many spatial regions. By augmenting $PM_{2.5}$ particle sensing data with images, we can better estimate the air quality at a particular location and improve forecast accuracy.

Researchers have yet to consider image-based air quality forecasting. Cameras (e.g., webcams) are less expensive and easier to maintain than most commonly used air quality sensors. Using images for $PM_{2.5}$ estimation typically increases field estimation MAE accuracy by 14.3% compared to only using particle sensors [8]. Moreover, existing research in visibility physics demonstrates significant correlation between visibility and $PM_{2.5}$ levels [9, 10]. Finally, Zhang et al. improves the accuracy of image-based air quality prediction by 22% compared to existing image-based techniques [7]. On the other hand, applying images in forecasting future $PM_{2.5}$ levels has not yet been attempted.

Past research includes numerous image-based haze detection techniques [11, 12], but they do not capture the complex spatio-temporal correlations of haze in images over time. Our objective is to forecast $PM_{2.5}$ concentrations by fusing $PM_{2.5}$ concentrations with colocated images, which requires spatio-temporal analysis of air quality in the images. In this paper, we jointly use a convolutional neural network (CNN) and a long short-term memory (LSTM) to model the level of haze in the images over time.

Accurate image-based forecasting requires a quantitative knowledge of intricate relationships between the haze in each image region and the $PM_{2.5}$ data. Inspired by the success of attention networks in low-level computer vision [13, 14], our method incorporates spatial attention, which learns the image regions to focus on, and feature attention, which learns the importance of each feature extracted from the image. Spatial attention selects the regions based on their similarity with $PM_{2.5}$ latent features. We hypothesize that spatial attention can improve predictions by identifying image regions with $PM_{2.5}$ concentrations that are better correlated with the ground-truth sensor location.

The main contributions of this paper are (1) addressing the image-based air pollution forecasting problem for the first time, (2) developing a forecasting model capturing the level of haze in images over time with a combined CNN and RNN,

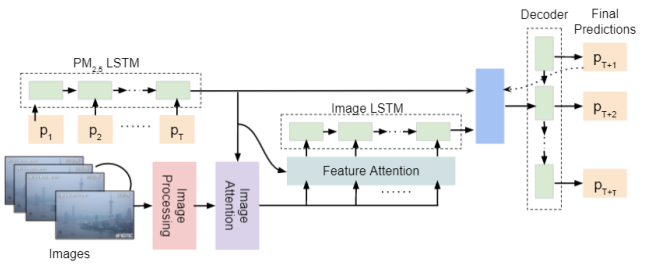

Fig. 1. Image-based air quality forecasting model overview.

which is novel in this context, and (3) incorporating multi-level attention to learn intricate relationships between images and the $PM_{2.5}$ data. We evaluate our model on Shanghai data containing hourly $PM_{2.5}$ measurements. Using images for PM2.5 concentration forecasting improves accuracy by 15.8% compared to previous forecasting methods.

2. PROBLEM FORMULATION

This section describes the mathematical notation used throughout the paper and the problem formulation for air pollution forecasting. We forecast $PM_{2.5}$ concentrations measured by a monitoring station every hour. Images are captured hourly near the monitoring station. Assuming a time window of length T, we are given as input $PM_{2.5}$ data $\mathbf{P_i} = \{p_t\}_{t=1}^{T} \in R^T$. The corresponding images are specified as $\mathbf{I} = \{i_t\}_{t=1}^{T} \in R^{T \times C \times H \times W}$, where $i_t \in R^{C \times H \times W}$, C is the number of channels, H is the height, and W is the width of the image. In this paper, $C = 3$, $H = 32$, and $W = 32$. We aim to predict the $PM_{2.5}$ concentrations over the next τ hours where the ground-truth is represented by $\mathbf{P_f} = \{p_{T+t}\}_{t=1}^{\tau} \in R^{\tau}$.

We formulate the problem as $\hat{P}_{\mathrm{f}} = M(\mathbf{P_i}, \mathbf{I})$, where M is the forecasting model, and the predictions are $\hat{P}_{\mathrm{f}} = \{\hat{p}_{T+t}\}_{t=1}^{\tau} \in R^{\tau}$.

3. IMAGE-BASED FORECASTING MODEL

We describe a novel multi-level attention LSTM network designed for air pollution forecasting. Unlike previous haze detectors [11, 15], our model can represent changes in haze over both space and time. Our proposed model integrates a CNN and an LSTM. The CNN extracts the haze from each image and the LSTM predicts $PM_{2.5}$ concentrations over time. We also incorporate multi-attention to learn intricate relationships between images and $PM_{2.5}$ data.

Figure 1 shows the architecture of the proposed model, resembling the encoder-decoder framework for time-series forecasting [16]. We develop three LSTM sequences, one encoding the previous $PM_{2.5}$ time-series data, another encoding the sequence of images, and another forecasting future $PM_{2.5}$ concentrations. We feed the past $PM_{2.5}$ data into an LSTM encoder to obtain its latent representation, and the image processing module learns to identify hazy regions from the images. Next, the image attention module weights each image region using the $PM_{2.5}$ hidden representation. The feature attention module then embeds each image and weights each image feature, and the image features are fed into another LSTM encoder. Finally, the LSTM decoder forecasts future $PM_{2.5}$ concentrations from the outputs of the two encoders.

Layer	# of Filters	Filter size	Activation
Conv	16	3 x 3	-
RDB 1-3	16	3 x 3	-
Conv	32	3 x 3	ReLU
Pool	32	H/2 x W/2	-
Conv	64	3 x 3	ReLU
Pool	64	H/4 x W/4	-
Conv	128	3 x 3	ReLU
Pool	128	H/8 x W/8	-

Table 1. The architecture of the image processing module. Padding keeps image sizes consistent. For the pooling layers, the filter output size is given.

Layer	Input size	Output size	Activation
FC	$128 \times \frac{H}{8} \times \frac{W}{8}$	128	ReLU
FC	128	128	ReLU
FC	128	output size	ReLU

Table 2. The fully connected (FC) layers of the image embedding layers.

3.1. Data Representation

3.1.1. $PM_{2.5}$ Data Representation

The encoder of the $PM_{2.5}$ data is comprised of a sequence of LSTMs of length T. The $PM_{2.5}$ concentration p_t at time t is fed as an input to the encoder as $h_t = f_e(h_{t-1}, p_t)$, where f_e represents an LSTM unit and h_t represents the t-th hidden state. We obtain hidden states for each time step $H = \{h_1, ..., h_T\}$, where $h_t \in R^n$ is the t-th hidden state and n is the size of each hidden state. The output of the encoder is the hidden representation h_T of the entire $PM_{2.5}$ sequence.

3.1.2. Image Representation

Since images can capture the level of air quality, we learn to identify hazy regions from the images. For this purpose, we adapt the Residual Dense Block (RDB) [17], which has been used for single-image dehazing [15]. We develop the image sequence processing module outlined in Table 1. From the input i_t, the module begins with a Conv. layer and proceeds

with three RDB blocks[1], and finally three Conv-Pool layers. The output i'_t consists of feature maps representing the level of haze for each region. Its output dimensions are $128 \times \frac{H}{8} \times \frac{W}{8}$, or $128 \times 4 \times 4$.

3.2. Image Attention Module

While the RDB has the ability to capture haze in an image effectively [17], it treats every pixel equally although images may contain uneven haze. Different regions in a city exhibit spatial correlation of $PM_{2.5}$ levels, and images can help track that spatial correlation via image attention for each pixel. It is important to weight image regions according to their relationship with the $PM_{2.5}$ data because image attention computes the correlation of $PM_{2.5}$ of each pixel relative to the $PM_{2.5}$ latent features.

For each time t, we calculate the attention weight for each 4×4 region using the the latent representation of the $PM_{2.5}$ data $h_T \in R^n$. We compute the dot product between $W_i i'_t(x,y)$ and $W_h h_T$, where (x,y) is the location of the region, and $i'_t(x,y) \in R^{128}$ is the 128-dimensional representation of the region at (x,y). The learned parameters are $W_i \in R^{128\times128}$ and $W_h \in R^{128\times n}$. The attention weight $s_t(x,y)$ denotes the importance of the (x,y) region at time t and represents the similarity between the region and h_T.

$$s_t(x,y) = [W_i i'_t(x,y)]^T W_h h_T. \quad (1)$$

The attention weights are then normalized over all regions using softmax . Finally, we multiply the attention weight matrix by i'_t to obtain the output i''_t.

$$\alpha_t(x,y) = \frac{\exp[s_t(x,y)]}{\sum_{x=1}^{4}\sum_{y=1}^{4}\exp[s_t(x,y)]} \text{ and} \quad (2)$$

$$i''_t = \alpha_t i'_t. \quad (3)$$

3.3. Feature Attention Module

The image feature attention module represents the relationship between each image feature and the latent features h_T of $PM_{2.5}$. It adaptively selects the image features most relevant to h_T when predicting the future time series. We flatten the output i''_t to one dimension where $i'' \in R^{m\times T}$ and feed it to the image embedding layers for each time t as described in Table 2. The size of the output layer m is a hyper-parameter selected during training.

For time t, we calculate the attention weight of each image feature j via h_T. We compute the dot product between $W'_i i''(j)$ and $W'_h h_T$, where $i''(j) \in R^T$, and $h_T \in R^n$. The parameters to learn are $W'_i \in R^{n\times T}$ and $W'_h \in R^{n\times n}$. The attention weight $s(j)$ represents the importance of j-th feature.

$$s(j) = [W'_i i''(j)]^T W'_h h_T. \quad (4)$$

The weights are normalized by the softmax over all m features.

$$\alpha(j) = \frac{\exp[s(j)]}{\sum_{k=1}^{m}\exp[s(k)]}. \quad (5)$$

The attention weights denote the importance of the individual features. Then, the input vector for time t follows:

$$\tilde{x}_t^{img} = [a(1)i''_t(1), a(2)i''_t(2), ..., a(m)i''_t(m)]^T. \quad (6)$$

We hypothesize that feature attention can improve predictions by identifying image features (e.g., resembling weather conditions) that are better correlated with the ground-truth sensor location. $PM_{2.5}$ is correlated with external factors such as meteorology, time of day, and land use. Feature attention calculates the correlation of each image feature with the $PM_{2.5}$ latent features via a dot product.

3.4. Model Architecture

The image features $\tilde{x}^{img}$ are fed into an LSTM encoder for images, comprised of a sequence of LSTMs of length T. The image features $\tilde{x}_t^{img}$ at time t are fed as an input to the encoder as $h_t^{img} = f_e^{img}(h_{t-1}^{img}, \tilde{x}_t^{img})$, where f_e^{img} represents an LSTM unit for the image and $h_t^{img} \in R^n$ represents the t-th hidden state of size n for the image. The output is the hidden representation h_T^{img} of the entire image sequence.

In the decoder with length τ, we concatenate the hidden representation of the image sequence h_T^{img} and the $PM_{2.5}$ data h_T. $h_0^d = [h_T^{img}; h_T] \in R^{2n}$ is then initialized as the first hidden state of the decoder. The previous output of the LSTM becomes the input of the next LSTM p'_t to update the decoder hidden state.

$$h_t^d = f_d(h_{t-1}^d, p'_t), \quad (7)$$

where f_d is an decoder LSTM unit. Afterward, we can estimate y_t:

$$y_t = W_y^T h_t^d + b_y. \quad (8)$$

The learned parameters are $W_y \in R^{2n}$, $b_y \in R$, which determine the prediction y_t.

4. EXPERIMENTAL RESULTS

We evaluate our proposed model on air quality data and images from Shanghai. We first introduce the dataset and the experimental protocol. Next, we evaluate our proposed forecasting method and compare it with other methods. Afterward, we investigate the effect of each individual component of our proposed forecasting model.

[1]For RDB, the depth rate (number of input features) is 16, the number of dense layers is 4, and the growth rate is 16. More details about RDB are in Zhang et al. 2018.

Method	RMSE	MAE
HA	54.84	44.43
SVR	43.97	29.52
GBR	38.75	24.57
RNN	28.12	18.43
LSTM	27.81	18.69
GRU	28.25	18.27
Seq2seq	27.99	17.78
Only image processing module	25.59	16.90
Proposed approach with only image attention	24.78	16.32
Proposed approach	23.57	15.84

Table 3. Comparisons with previous forecasting methods in Shanghai (in $\mu g/m^3$) for six-hour forecasts.

4.1. Dataset and Implementation Details

We use air quality data from July 1st, 2014 to December 31, 2014 from the U.S. Consulates in Shanghai. The data contain hourly $PM_{2.5}$ measurements in $\mu g/m^3$. We also use webcam images taken by the Shanghai Environmental Monitoring Center near the air quality measurement station [18, 19]. The images were taken at the Oriental Pearl Tower. Our dataset includes images in the same date range approximately every hour from 8:00 am to 10:00 pm. We resized the images to $3 \times 32 \times 32$ (C $\times$ H $\times$ W). There are 2,296 chronologically ordered images.

The sequence length of the encoder T is 6 (the window size) and the decoder time-step τ is 6. During the training phase, we conduct grid search to determine hyperparameter values. We set the learning rate to 0.005 and the batch size to 4, and apply early stopping for model training. The hidden size of each LSTM unit is 32, and the output size of the FC unit for the image processing module is 16 units.

We divide the dataset using an 8:1:1 ratio for training, validation, and testing data, which do not overlap. We use Adam to optimize parameters during training and use mean squared error (MSE) as the loss function. We evaluate our model's root mean squared error (RMSE) and mean absolute error (MAE). We also use gradient clipping with a parameter of 0.1. All experiments are run on a machine with an NVIDIA GeForce 940MX GPU.

4.2. Model Comparison

We compare our method with existing pollutant forecasting methods. We present the best performance of each method under different parameter settings. The methods we include are Historical Average (HA), where we predict $PM_{2.5}$ concentrations using the mean of previous $PM_{2.5}$ concentrations, Support Vector Regression (SVR), Gradient Boosting Regression (GBR), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Seq2seq. These time-series methods use only the $PM_{2.5}$ measurements as input.

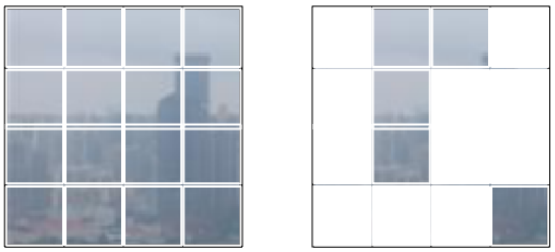

Fig. 2. The left figure is the original image with a 4×4 grid. The image attention module emphasizes certain regions of the image shown in the right figure.

Table 3 compares several air quality forecasting methods. We average the results of three runs. Experiments on Shanghai data show that our forecasting model improves accuracy by 15.8% in RMSE and 10.9% in MAE.

We also evaluate the impact of each model component via ablation studies (see Table 3). These methods use both the $PM_{2.5}$ measurements and images as input. Notably, using images improves accuracy by 8.6% relative to Seq2seq when we add an encoder that extracts image features through the image processing module. Furthermore, adding the image attention module improves accuracy because it selects image regions by computing a dot product with the $PM_{2.5}$ latent features. This module further improves accuracy by 11.5% relative to Seq2seq. Finally, adding the feature attention module improves accuracy by 15.8% since the module weights the extracted image features through a dot product with the $PM_{2.5}$ latent features.

We hypothesize the image attention module improves accuracy because it can identify image regions with $PM_{2.5}$ levels that are better correlated with the ground-truth sensor location. As shown in Figure 2, the image attention module emphasizes certain regions of the image. Since those regions tend to be clustered around the same area, we believe that the attention module can evaluate the correlations in $PM_{2.5}$ concentrations of different regions with sensor location.

5. CONCLUSION

This paper has described an image- and attention-based LSTM architecture to forecast $PM_{2.5}$ concentration. It uses multi-level attention to represent the spatio-temporal relationship of visual haze with measured $PM_{2.5}$ concentration over time. Experiments performed in Shanghai show that the proposed model achieves improved performance relative to previous air pollution forecasting methods. Fusing $PM_{2.5}$ data with images enables more accurate $PM_{2.5}$ forecasts.

6. REFERENCES

[1] Shaolong Feng, Dan Gao, Fen Liao, Furong Zhou, and Xinming Wang, "The health effects of ambient PM2. 5 and potential mechanisms," *Ecotoxicology and Environmental Safety*, vol. 128, pp. 67–74, 2016.

[2] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li, "Forecasting fine-grained air quality based on big data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 2267–2276.

[3] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng, "Deep distributed fusion network for air quality prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, p. 965–973.

[4] Zhipeng Luo, Jianqiang Huang, Ke Hu, Xue Li, and Peng Zhang, "Accuair: Winning solution to air quality prediction for kdd cup 2018," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, p. 1842–1850.

[5] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, April 2017.

[6] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng, "Geoman: Multi-level attention networks for geo-sensory time series prediction," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 7 2018, pp. 3428–3434.

[7] Tony Zhang and Robert P. Dick, "Estimation of multiple atmospheric pollutants through image analysis," in *Proceedings of the International Conference on Image Processing*, 2019, pp. 2060–2064.

[8] Zuohui Chen, Tony Zhang, Zhuangzhi Chen, Yun Xiang, Qi Xuan, and Robert P. Dick, "Hvaq: A high-resolution vision-based air quality dataset," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[9] Xiaoxin Fu, Xinming Wang, Qihou Hu, Guanghui Li, Xiang Ding, Yanli Zhang, Quanfu He, Tengyu Liu, Zhou Zhang, Qingqing Yu, Ruqin Shen, and Xinhui Bi, "Changes in visibility with pm2.5 composition and relative humidity at a background site in the pearl river delta region.," *Journal of environmental sciences*, vol. 40, pp. 10–9, 2016.

[10] Xiaoxiao Zhang, Xiang Ding, Dilinuer Talifu, Xinming Wang, Abulikemu Abulizi, Mailikezhati Maihemuti, and Suwubinuer Rekefu, "Humidity and pm2.5 composition determine atmospheric light extinction in the arid region of northwest china.," *Journal of environmental sciences*, vol. 100, pp. 279–286, 2021.

[11] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, December 2011.

[12] Qingsong Zhu, Jiaming Mai, and Ling Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, November 2015.

[13] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Raymond Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018.

[14] Saeed Anwar and Nick Barnes, "Real image denoising with feature attention," *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[15] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[17] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Nathan Jacobs, Nathaniel Roman, and Robert Pless, "Consistent Temporal Variations in Many Outdoor Scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, March 2007.

[19] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless, "The global network of outdoor webcams: Properties and applications," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November 2009, pp. 111–120.