Deep Multi-task Learning for Air Quality Prediction

Bin Wang, Zheng Yan, Jie Lu,
Guangquan Zhang, and Tianrui Li

School of Information Science and Technology
Southwest Jiaotong University
Chengdu, China
Faculty of Engineering and Information Technology
University of Technology, Sydney
Sydney, Australia
wangbin@my.swjtu.edu.cn
{yan.zheng,jie.lu,guangquan.zhang}@uts.edu.au
trli@swjtu.edu.cn

Abstract. Predicting the concentration of air pollution particles has been an important task of urban computing. Accurately measuring and estimating makes the citizen and governments can behave with suitable decisions. In order to predict the concentration of several air pollutants at multiple monitoring stations throughout the city region, we proposed a novel deep multi-task learning framework based on residual Gated Recurrent Unit (GRU). The experimental results on the real world data from London region substantiate that the proposed deep model has manifest superiority than shallow models and outperforms 9 baselines.

Keywords: Deep learning, Recurrent neural networks, Neural networks, Air quality prediction, Urban computing

1 Introduction

In recent years, along with economic development, air pollution in developing countries has become a serious issue [1]. Air pollutants comprise molecule (e.g., $PM_{2.5}$ and PM_{10}) and harmful gas (e.g. NO_2) are threatening the public health [4]. For monitoring real-time air pollution, Chinese governments have built the amount of air quality monitoring stations and collect air quality data every hour in recent years [18]. Besides monitoring, there is a rising demand for forecasting future air quality index (AQI). Accurately measuring and estimating the concentration of air pollution particles makes the citizen and governments can behave with suitable decisions, such as reducing outdoor activities, to remarkably reduce the adverse results of air pollution.

Air quality prediction methods mainly fall into two categories: classical dispersion models and data-driven models [17,**?**]. Classical dispersion models identify the root cause of air pollution from chemical, emission, climatological and

combinations of these factors. These models are most a numerical function of emissions from industry and vehicular, meteorology, and other factors. However, it is very difficult to get all these factors completely and accurately. Thus, the prediction accuracy is hard to be guaranteed. Also, the computation complexity is very high.

Data-driven approaches, e.g. artificial neural networks, forecast air pollutions based on massive observed data. Deep learning, as a cutting-edge technique of machine learning, has made great success in computer vision tasks and is very suitable and robust when modeling complicated spatio-temporal data e.g., videos. Inspired by this, more and more researchers begin to improve and apply it to solve urban computing problems and achieve considerable results. In this paper, we propose to use a novel deep learning model to predict the concentration of air pollutants. Following are the major contributions of this paper.

1. *W*e introduce a novel Residual-GRU based on vanilla GRU for short-term air pollutants. Experimental results demonstrate this model can speed up convergence and perform better during the test.
2. We formalize the air pollutants prediction into a multi-task end-to-end framework. Previous related studies are single-task which predict only for one station or only one pollutant but our method can forecast for all stations and all pollutants at one time.

The rest of the paper is organized as follows: In Section 2, we first explain some basic variants of recurrent neural networks (RNN) and then introduce the proposed model in more details. Then, we experimentally evaluate our proposed prediction models and compare them with other baselines in Section 3. In Section 4, we present related works. Finally, conclusions and future works are given in Section 5 and Section 6, respectively.

2 Recurrent Neural Networks

In this section, we introduce the different versions of RNN and our proposal, the improved model Residual-GRU. For simplicity, all bias terms are omitted.

Vanilla RNN This model is specially designed to incorporate sequential information and has achieved great success particularly in NLP tasks. The formulas of vanilla RNN are shown as below:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$y_t = f(W_{ht}h_t)$$

where x_t is the input at time t, W is the transformation weights, f is the element-wise activation function such as $tanh$, and h_t and y_t is the hidden state and output at time t respectively. A serious drawback of vanilla RNN is it can hardly capture long-term sequential dependency due to vanishing gradients.

LSTM To overcome the drawback of vanilla RNN, LSTM is designed as below:

$$i_t = \sigma(W^i_{hh}h_{t-1} + W^i_{xh}x_t)$$

$$f_t = \sigma(W_{hh}^f h_{t-1} + W_{xh}^f x_t)$$

$$o_t = \sigma(W_{hh}^o h_{t-1} + W_{xh}^o x_t)$$

$$\tilde{C_t} = tanh(W_{hh}^g h_{t-1} + W_{xh}^g x_t)$$

$$C_t = \sigma(f_t \odot C_{t-1} + i_t \odot \tilde{C_t})$$

$$h_t = tanh(C_t) \odot o_t$$

where i, f, o are input, forget and output gates, respectively. $\odot$ is the Hadamard product. Such a three-gates mechanism can effectively alleviate vanishing gradients problems.

GRU GRU could be regarded as the light LSTM. It only utilizes two gates to control information flow, which is shown as below:

$$z_t = \sigma(W_{hh}^z h_{t-1} + W_{xh}^z x_t)$$

$$r_t = \sigma(W_{hh}^r h_{t-1} + W_{xh}^r x_t)$$

$$\tilde{h_t} = tanh(W_{hh}^h (r_t \odot h_{t-1}) + W_{xh}^h x_t)$$

$$h_t = z_t \odot \tilde{h_t} + (1 - z_t) \odot h_{t-1}$$

where z_t is called update gate and r_t is called reset gate.

Proposed Residual-GRU Based on GRU, inspired from residual convolutional networks [3], the proposed Residual-GRU combines residual learning with GRU. The unique difference between Residual-GRU and GRU resides in $h_t = z_t \odot \tilde{h_t} + (1 - z_t) \odot h_{t-1} + W_{xh}^{res} x_t$. By providing a shortcut path (i.e., $W_{xh}^{res} x_t$) between adjacent layer outputs, it could release gradient flow more smoothly and accelerate the training process. Please note that W_{xh}^{res} can be omitted if the dimension of x_t is equal to h_{t-1} and $\tilde{h_t}$, that is, $h_t = z_t \odot \tilde{h_t} + (1 - z_t) \odot h_{t-1} + x_t$. In our experiments, we set the their dimension equal, and hence omit the W_{xh}^{res}. The graphic framework is shown in Fig. 1. Particularly, the involved hyperparameters include

1. *Input length* is set as 5, i.e., we use the previous 5-hours vector sequences as inputs to predict next time vector. At each timestamp, the dimension of the input vector is $19 * 3 = 57$, which means 19 monitoring stations times by 3 pollutants concentration ($PM_{2.5}$, PM_{10}, NO_2). In this way, we can implement a multi-task learning in a unified end-to-end framework.
2. *Number of layers* and *hidden states* are set as 2 and 53 respectively. The last timestamp is furthermore followed by fully connection with 53 nodes for deep representation learning.
3. *Epochs* and *batch size* are set as 100 and 32 respectively.
4. *Activation functions* are all set as *tanh* except the output layer with *sigmoid*.

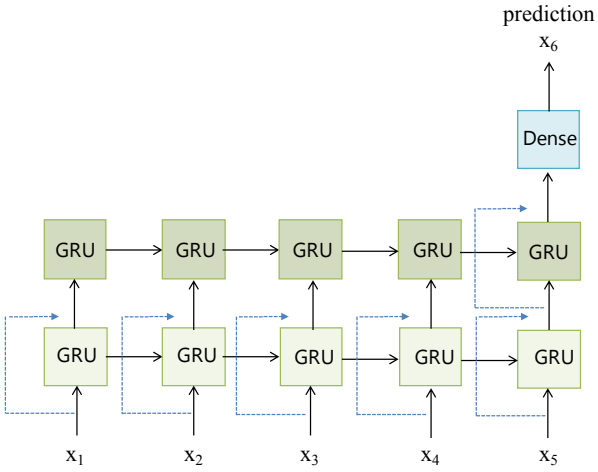


Fig. 1: The framework of Residual-GRU, where the dash line means the residual connection and x_i is the input vector. *Dense* means the fully connected layer.

3 Experiments

3.1 Dataset

The hourly concentration data of three air pollutants, i.e., $PM_{2.5}$, PM_{10}, and NO_2 at London region from 1st/Mar/2017 to 27th/Mar/2018 were downloaded from KDD CUP of Fresh Air. This dataset includes 9385 timestamps for 19 stations. we first impute the missing value with historical mean and then use max-min to normalize features into [0,1]. In our experiments, we split data as 0.8/0.2 for training and test.

3.2 Baselines

Our baselines are generally separated into 2 categories, i.e., classic machine learning methods and deep learning models. The classic machine learning methods mainly include:

- **SVR** SVR is the support vector machine designed for regression. The kernel function usually includes 'rbf', 'linear', 'poly' and so on.
- **LASSO** Lasso is a regression analysis method that performs both variable selection and regularization.
- **RandomForest**. Random forest is a popular ensemble method for classification and regression. The main drawback of it is time-consuming for prediction.

The deep learning baselines include:

- **Residual-LSTM** This baseline just replaces the GRU unit with LSTM to demonstrate the superiority of GRU.
- **LSTMs-100-100** This baseline removes residual connections and consists of 2-layers-depth LSTM and 100 hidden states for each layer by which we can observe the effectiveness of residual learning.
- **LSTM-100** This 1-layer-depth LSTM is devised to check the effect of depth.
- **Dense-100_LSTM-100** Different from LSTM-100, we add a time distributed fully connected layer for each time step. Hence this model introduces more powerful learning ability and might be with weaker generalization.

3.3 Evaluation Metric

We measure our method by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

$$RMSE = \sqrt{\frac{1}{z}\sum_{i}(x_i - \hat{x_i})^2} \tag{1}$$

$$MAE = \frac{1}{z}\sum_{i}|x_i - \hat{x_i}| \tag{2}$$

where $\hat{x}$ and x are the predicted value and ground truth, respectively; z is the number of all predicted values.

3.4 Results

We train all deep models on a server with Quadro P5000 GPU and the programming environment is Keras with TensorFlow backend. Table. 1 shows the experimental results which illustrate our proposed model is state-of-the-art. Fig. 2 reveals the loss variation during training and test. Please note that we do not fine-tune the hyper-parameters exhaustly. There are some important conclusions we can summarise:

1. Above all, we can see that classic machine learning methods are not very suitable than deep learning models. This demonstrates the effectiveness and importance of deep learning for complicate spatio-temporal feature extract.
2. Residual-GRU performs better than Residual-LSTM. This may be GRU module has fewer parameters and hence make it easier for learning in this spatio-temporal task.
3. By comparing LSTMs-100-100 with LSTM-100, we conclude more layers do not always mean better performance, the learning process can be difficult due to the deeper layers.
4. By analyzing LSTM-100-100 and Dense100-LSTM100, we find that even though time distributed fully connected layers induce more powerful learning ability, it degrades with poor generalization.

5. Fig. 2a reflects the residual connect can boost convergence, especially in the first few epochs. From Fig. 2b, we can see that Residual-GRU has better learning ability i.e., with less loss at the end of training. Fig. 2c demonstrates the proposed has better generalization capability during the test.
6. By observing yellow, blue, and red loss curves from Fig. 2b and Fig. 2c, we know that less training loss does not mean better generalization. However, the proposed model has both the lowest loss in the trade-off. This indicates the effectiveness of Residual-GRU for air pollutants prediction.

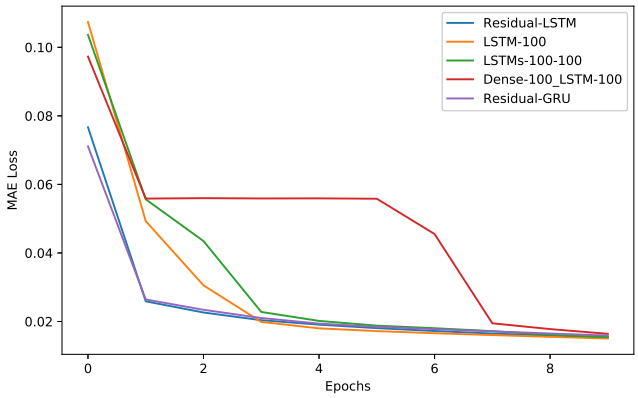
Table 1: RMSE and MAE among different models

Model	RMSE		MAE	
	Train	Test	Train	Test
Residual-GRU	**1.59**	**1.85**	**0.96**	**1.15**
Residual-LSTM	1.62	1.87	0.98	1.16
LSTM-100	1.65	1.90	1.00	1.18
LSTMs-100-100	1.66	1.97	1.02	1.23
Dense100-LSTM100	1.59	1.91	0.97	1.18
SVR-rbf	10.69	10.43	8.91	8.55
SVR-poly	10.61	10.35	8.83	8.46
SVR-linear	10.36	10.09	8.67	8.30
LASSO	3.88	4.64	2.33	2.76
RandomForest	2.11	2.69	1.29	1.64

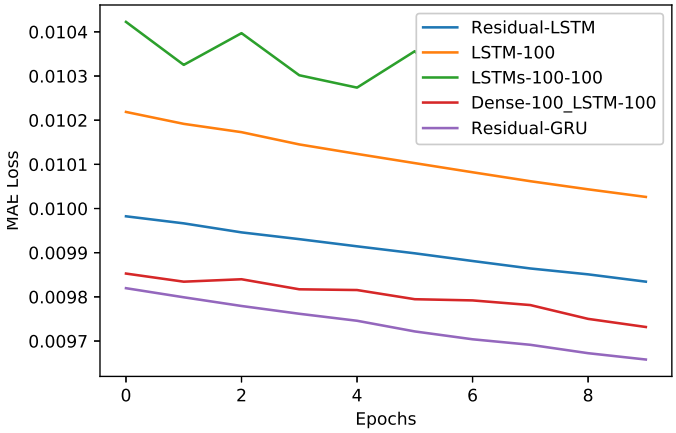
4 Related Works

4.1 Deep Spatio-Temporal Learning

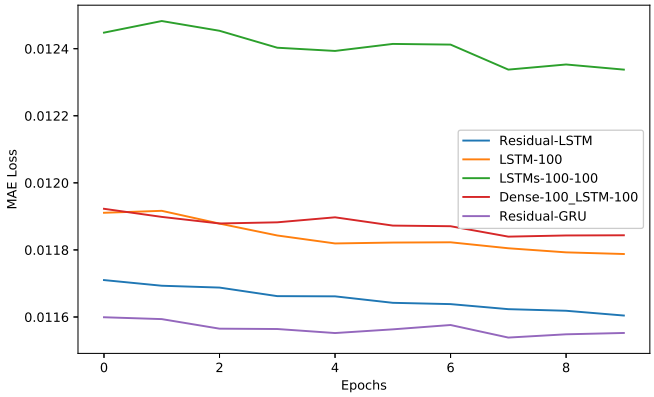
Inspired by deep learning on computer vision, [16] firstly applied deep learning to crowds flow prediction and proposed DeepST. This model adopted 3 deep CNNs to capture closeness, period and seasonal trend respectively. Besides spatio-temporal data, it also fused different source data such as weather for a better generalization. Furthermore, [15] proposed ST-ResNet which is an improved version of DeepST. To add deeper layers and get better precision, it introduced residual convolutional neural networks (CNN). With the similar philosophy as predicting citywide crowds flows, [9] adopted deep CNN on grid-based spatio-temporal data to make transportation network speed prediction. [10] applied ST-ResNet to real-time crime forecasting. [13] proposed a deep multi-view network (DMVST-Net) to predict taxi demand based on the hybrid of CNN, LSTM, and fully connected networks. By fusing different models, [5] proposed fusion convolutional LSTM (FCL-Net) to forecast passenger demand under on-demand ride services. They also suggested a random forest employed for feature



(a) MAE loss of the first 10 epochs during training

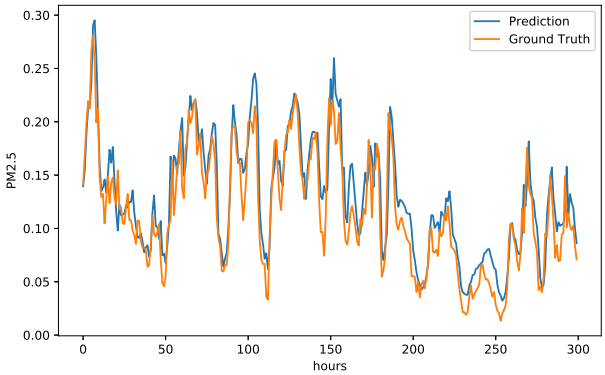


(b) MAE loss of the last 10 epochs during training

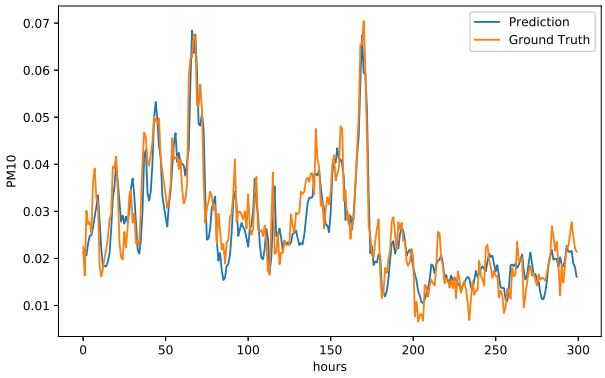


(c) MAE loss of the last 10 epochs during test

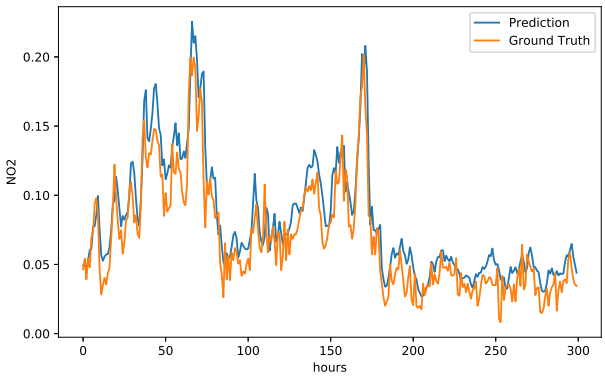
Fig. 2: Training MAE loss



(a) Concentration prediction of $PM_{2.5}$



(b) Concentration prediction of PM_{10}



(c) Concentration prediction of NO_2

Fig. 3: Concentration prediction at one certain station

selection can save training time without losing much accuracy. [12] creatively integrated convolution and LSTM together and introduced ConvLSTM for precipitation nowcasting. [2] proposed LSTM-based spatio-temporal learning for wind speed forecasting. [11] proposed short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework (CLTFP).

4.2 Air Quality Forecasting

[19] proposed the concept of spatial partition and aggregation and a hybrid shallow model based on multi-view learning to implement real-time air quality prediction over future 48 hours. [7] developed a stacked auto-encoder network to extract deep representations of the concentration of PM2.5 and then use a logistic regression to predict. Such a two-stage training and fine-tuning framework is tedious and cannot learn from end-to-end. [6] used two vanilla LSTMs to predict the concentration of O_3 and NO_2 respectively. They furthermore transformed the predictive results into category label according to the AQI thresholds and measure results from accuracy. [14] introduced a new distributed fusion framework to fuse heterogeneous multi-source data, which can simultaneously capture the individual and overall effects from all influential factors for AQI prediction. [8] proposed GeoMAN using a multi-level attention-based RNN that considers multiple sensors' and information fusion. The main contribution is the multi-level attention mechanism, i.e. local attention and global attention. A clear difference between previous works and ours is that the previous study is single-task but Residual-GRU can implement multi-task at one time. To the best of our knowledge, we are the first to utilize residual GRU to implement multi-task learning for air quality prediction.

5 Conclusion

In this paper, we propose a deep multi-task learning model to predict air quality. This model integrates residual connections into GRU and can predict multiple concentrations of pollutants at all sites simultaneously. We evaluate our method based on a real-world dataset and extensive experiments show the superiority of our method against 9 baselines.

6 Future Works

In the future, we will extend our method to solve the problem of long-term prediction and design more metrics for different considerations. Moreover, we will incorporate more external factors (e.g., weather) and explore new processing techniques such as fuzzy granulation.

Acknowledgment

This work was supported by the Natural Science Foundation of China (No. 61773324), the Fundamental Research Funds for the Central Universities (No. 2682015QM02) and the Australian Research Council (No. DP150101645).

References

1. Akimoto, H.: Global air quality and pollution. Science 302(5651), 1716–1719 (2003)
2. Ghaderi, A., Sanandaji, B.M., Ghaderi, F.: Deep forecast: Deep learning-based spatio-temporal forecasting. arXiv preprint arXiv:1707.08110 (2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645 (2016)
4. Kampa, M., Castanas, E.: Human health effects of air pollution. Environmental pollution 151(2), 362–367 (2008)
5. Ke, J., Zheng, H., Yang, H., Chen, X.M.: Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. Transportation Research Part C: Emerging Technologies 85, 591–608 (2017)
6. Kök, ., Şimşek, M.U., Özdemir, S.: A deep learning model for air quality prediction in smart cities. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 1983–1990 (Dec 2017)
7. Li, X., Peng, L., Hu, Y., Shao, J., Chi, T.: Deep learning architecture for air quality predictions. Environmental Science and Pollution Research 23(22), 22408–22417 (2016)
8. Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y.: Geoman. In: International Joint Conference on Artificial Intelligence (IJCAI-18) (2018)
9. Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., Wang, Y.: Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. Sensors 17(4), 818 (2017)
10. Wang, B., Yin, P., Bertozzi, A.L., Brantingham, P.J., Osher, S.J., Xin, J.: Deep learning for real-time crime forecasting and its ternarization. arXiv preprint arXiv:1711.08833 (2017)
11. Wu, Y., Tan, H.: Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. arXiv preprint arXiv:1612.01022 (2016)
12. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems. pp. 802–810 (2015)
13. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J.: Deep multi-view spatial-temporal network for taxi demand prediction. arXiv preprint arXiv:1802.08714 (2018)
14. Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y.: Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2018)
15. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1655–1661 (2017)
16. Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X.: Dnn-based prediction model for spatio-temporal data. In: Proceedings of the International Conference on Advances in Geographic Information Systems. p. 92 (2016)

17. Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A.: Real-time air quality forecasting, part i: History, techniques, and current status. Atmospheric Environment 60, 632–655 (2012)
18. Zheng, Y., Liu, F., Hsieh, H.P.: U-air: When urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1436–1444. KDD '13, ACM (2013)
19. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., Li, T.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2267–2276. KDD '15, ACM (2015)