DeLELSTM: Decomposition-based Linear Explainable LSTM to Capture Instantaneous and Long-term Effects in Time Series-Supplementary

Chaoqun WANG[1] , **Yijun LI**[1] , **Xiangqian SUN**[1] , **Qi WU**[1*] , **Dongdong WANG**[2] and **Zhixiang HUANG**[2]

[1]School of Data Science, City University of Hong Kong, Hong Kong, China
[2]JD Digits, Beijing, China
{cqwang5-c, yijunli5-c,xqsun4-c}@my.cityu.edu.hk, qiwu55@cityu.edu.hk,
{wangdongdong9, huangzhixiang}@jd.com

1 Dataset

PM2.5 [Liang *et al.*, 2015]:It contains hourly PM2.5 data and the associated meteorological measurements in Beijing from 2010.1.1 to 2014.12.31. PM2.5 is the target series. Aside from PM2.5 values, the meteorological variables include dew point, temperature, pressure, wind direction, wind speed, hours of snow, and hours of rain. In this task, each hour of data from 0:00 to 23:00 of a day composes a multivariate time series sample, and there is a total of 1826 samples. Our task is to predict PM2.5 each hour within a day based on the seven meteorological measurements and the PM2.5 value.

Electricity [Gao *et al.*, 2022]:It records the time series of electricity consumption in the US, from 2017.10.11 to 2020.6.24. The consumption is chosen as the target series sampled hourly. The other 15 time series are exogenous factors, including max temperature, min temperature, dew point, feels like, heat index, wind chill, wind gust, cloud cover, humidity, precipitation, pressure, temperature, visibility, wind direction degree, and wind speed. Similarly to PM2.5, each hour of data from 0:00 to 23:00 of a day composes a series. After deleting samples with missing values of electricity consumption, there are 914 samples. Our task is to predict the electricity consumption each hour within a day based on the 15 exogenous variables and the electricity consumption value.

Exchange [Lai *et al.*, 2018]: It is the collection of the daily exchange rates of eight foreign countries, including Australia, British, Canada, Switzerland, China, Japan, New Zealand, and Singapore, ranging from 1990 to 2016. In this task, a continuous 30-day multivariable time series is a sample. Samples can be formed from any day to the next 30 days, totaling 7559 samples. Our target series is the Singapore exchange, and we aim to predict the exchange rate of Singapore for the following day.

For the three data sets, we divide the total samples into 75% for training, 15% for validation, and 15% for testing. The loss function of all the experiments is MSE. The formula is as follows:

$$loss = \frac{1}{N \times (T-1)} \sum_{i=1}^{N} \sum_{j=2}^{T} (y_i^j - \hat{y}_i^j)^2 \tag{1}$$

*Corresponding author.

Model	Electricity	PM2.5	Exchange
DeLELSTM	1	1	1
IMV-Full	0.74	1.37	1.37
IMV-Tensor	0.93	0.88	0.88
Retain	1.15	1.37	1.37

Table 1: The ratio of parameters

where N is the number of samples, $(T-1)$ is the length of target series to be predicted in a sample, y_i^j is the true value of the sample i at time j, and the $\hat{y}_i^j$ is the predicted value for the sample i at time j.

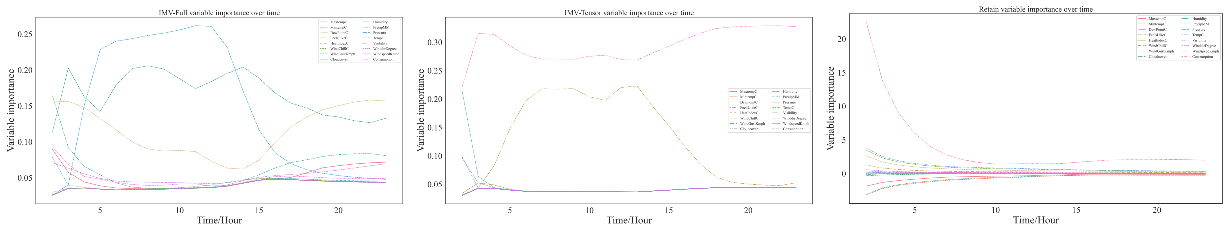
2 Training Details

Parameter Tuning: We conduct the grid search to select optimal parameters. The batch size is selected in $\{32, 64, 128\}$. Learning rate is searched in $\{0.05, 0.01, 0.001\}$. The size of the hidden states is selected in $\{32, 64, 128\}$. The batch size is 32 for PM2.5 prediction, 64 for electricity consumption prediction, and 128 for exchange prediction. The learning rate is set as 0.01 for electricity and exchange prediction and 0.05 for PM2.5 prediction. The hidden size is set as 32 for exchange and PM2.5 prediction and 64 for electricity prediction.

Initialization: The hidden-to-hidden transition, (U_i, U_f, U_o, U_c) in standard LSTM and the $(\mathcal{U}_i, \mathcal{U}_f, \mathcal{U}_o, \mathcal{U}_c)$ in tensorized LSTM, are initialized using Orthogonal initialization [Saxe *et al.*, 2013]. The input-to-hidden transition, (W_i, W_f, W_o, W_c) in standard LSTM and the $(\mathcal{W}_i, \mathcal{W}_f, \mathcal{W}_o, \mathcal{W}_c)$ in tensorized LSTM, are initialized using Xavier initialization [Glorot and Bengio, 2010]. The hidden states in standard LSTM and the tensorized LSTM are initialized as zero.

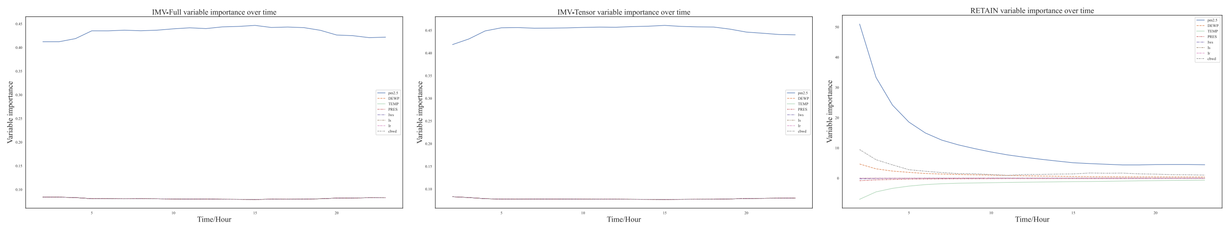
Implementation Details: The maximum number of epochs is 300 in all experiments. We use Adam [Kingma and Ba, 2014] as the optimizer and the adaptive learning rate strategy.

Based on the number of parameters in our model, Table 1 shows the ratio of the number of other models‘ parameters and those in our models on the three prediction tasks.



(a) IMV-Full results. (b) IMV-Tensor results. (c) RETAIN results.

Figure 1: Electricity variable importance over time using baseline models.



(a) IMV-Full results. (b) IMV-Tensor results. (c) RETAIN results.

Figure 2: PM2.5 variable importance over time using baseline models.

3 Interpretation Results of Baseline Models

In this subsection, we provide the interpretation results of three baseline models for the other two tasks, i.e., the PM2.5 prediction and the exchange prediction. Figure 1, Figure 2 and Figure 3 show the variable importance change over time for electricity consumption prediction, PM2.5 prediction, and exchange prediction, respectively. We can notice that the explainable models based on the attention mechanism give the variable importance always the same at most of the time points. In addition, except for the most important variables identified by the models, it could be hard for them to distinguish the importance of the other variables, such as the explanation results for PM2.5 prediction, and the results given by the RETAIN for exchange prediction.

Table 2 shows the prediction results using top 50% features identified by each explanation model. We can see that the features selected by our model can obtain better performance.

4 Computing Infrastructure

All models are run on a NVIDIA V100s GPU with 32GB RAM and AMD EPYC 7742 64-Core processor. All models are implemented in Python 3.9. The versions of main packages of our code are: Pytorch 2.0.0+cu117, Numpy: 1.23.5, Pandas: 1.5.3, Matplotlib: 3.7.1.

5 Codes and Datasets

All experiment codes are in https://github.com/wangcq01/DeLELSTM. All the datasets are public datasets which can be downloaded from:

- PM2.5: https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
- Electricity: https://www.kaggle.com/datasets/unajtheb/homesteadus-electricity-consumption
- Exchange: https://github.com/laiguokun/multivariate-time-series-data

References

[Gao *et al.*, 2022] Penglei Gao, Xi Yang, Kaizhu Huang, Rui Zhang, and John Yannis Goulermas. Explainable tensorized neural ordinary differential equations for arbitrary-step time series prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

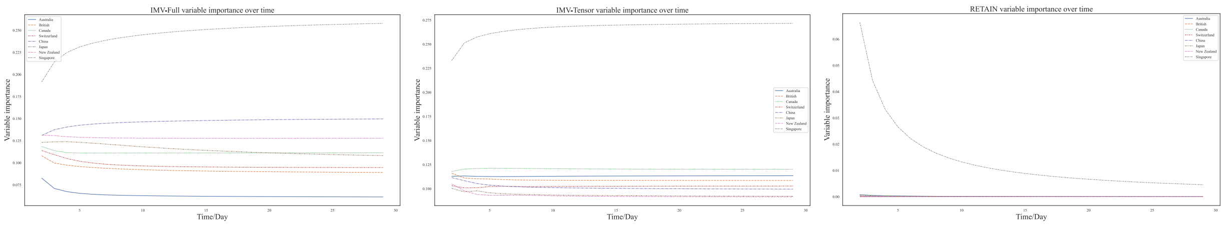
[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lai *et al.*, 2018] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

[Liang *et al.*, 2015] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.

[Saxe *et al.*, 2013] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.



(a) IMV-Full results.

(b) IMV-Tensor results.

(c) RETAIN results.

Figure 3: Exchange variable importance over time using baseline models.

Dataset	Metric	DeLELSTM	IMV-Full	IMV-Tensor	RETAIN
PM2.5	RMSE	**26.0593±0.3953**	26.3685±0.3906	26.3662±0.4189	26.2661±0.5080
	MAE	**13.3563±0.1338**	13.5728±0.1779	13.3801±0.0976	13.5094±0.0968
	MAPE	**21.76% ± 0.73%**	22.99% ± 0.83%	22.40% ± 0.54%	23.04% ± 0.48%
Exchange	RMSE	**0.0026±1.0e-05**	0.0026±1.0e-05	0.0026±7.0e-06	0.0026±6.3e-06
	MAE	0.0015±5.1e-06	0.0015±5.1e-06	**0.0015±3.2e-06**	0.0015±3.7e-06
	MAPE	0.23%±8.1e-06	**0.22%±3.3e-04**	0.23%±4.5e-06	0.23%±5.4e-06

Table 2: Performance based on top 50% important variables