Learning Decomposed Spatial Relations for Multi-Variate Time-Series Modeling (Supplemental Material)

Yuchen Fang,[1*] Kan Ren,[2] Caihua Shan,[2] Yifei Shen,[2]
You Li,[4] Weinan Zhang,[1] Yong Yu,[1] Dongsheng Li [2]

[1]Shanghai Jiao Tong University, [2]Microsoft Research Asia, [3]Central South University
{arthur_fyc, wnzhang}@sjtu.edu.cn, kan.ren@microsoft.com

Appendix

We present detailed experiment settings and additional experiment results in this section. Moreover, the implementation codes of the proposed method (SRD-TCN and SRD-GRU) are offered in the supplementary materials.

Experiment settings

This section will give more detailed experiment settings, including dataset description and hyper-parameter settings.

Dataset description A detailed description of the datasets is presented below. We split the forecasting datasets into train, valid, and test sets in chronological order and conducted cross-validation on the neonatal dataset. The statistics and split details are presented in Table 2.

- **Solar** (Lai et al. 2018) includes the solar power production records in the year of 2006, which are sampled every 10 minutes from 137 PV plants in Alabama State.
- **Electricity** (Lai et al. 2018) records the hourly electricity consumption in kWh ranging from 2012 to 2014, for $n = 321$ clients.
- **Pems-bay** (Li et al. 2017) contains average traffic speed measured by 325 sensors in the Bay Area ranging from Jan 2017 to May 2017, offered by California Transportation Agencies (CalTrans).
- **Metr-la** (Li et al. 2017) contains average traffic speed measured by 207 loop detectors on the highways of Los Angeles County provided by the Los Angeles Metropolitan Transportation Authority ranging from Mar 2012 to Jun 2012.
- **Neonatal seizure detection dataset** (Stevenson et al. 2019) contains multi-channel EEG recordings from human neonates at the Helsinki University Hospital, and the visual interpretation of the EEG by the human experts.

[*]The work was conducted during the internship of Yuchen Fang and You Li at Microsoft Research. Correspondence to Kan Ren and Weinan Zhang.

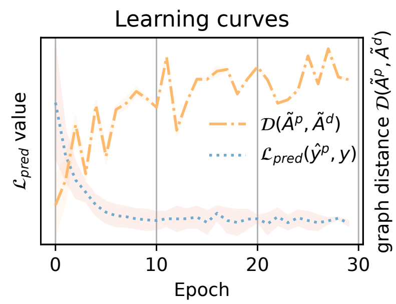

Figure 1: The learning curve of w/o PL

Hyper-parameter settings The hyper-parameters of SRD-GRU and SRD-TCN are presented in Table 1 for reproducibility. The search spaces of all common hyper-parameters, e.g., hidden dimension or learning rate, are shared among all compared methods for fair comparison. More implementation details can be referred to in our code.

Additional ablation study

Recall that in Eq. (19), the prediction loss $\mathcal{L}_{\text{prediction loss}}$ is the combination of $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}}, \boldsymbol{y})$, $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^p}, \boldsymbol{y})$ and $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^d}, \boldsymbol{y})$ to ensure both prior and dynamic graphs model meaningful spatial relation information for final predictions. To justify the design of our prediction loss, we conduct ablation study on SRD-TCN with $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^p}, \boldsymbol{y})$ and $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^d}, \boldsymbol{y})$ removed, denoted as **w/o PL**. As $\hat{\boldsymbol{y}^p}$ and $\hat{\boldsymbol{y}^d}$ is not optimized, the final prediction of w/o PL is instead generated as $\hat{\boldsymbol{y}} = \text{PM}(\text{Concat}[\boldsymbol{H}_Q^d, \boldsymbol{H}_Q^p])$.

The R^2 results on Pems-bay are shown in Table 3. w/o PL performs worse then SRD-TCN and achieves similar results with w/o MM (SRD-TCN without min-max learning paradigm), which illustrates the importance of $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^p}, \boldsymbol{y})$ and $\mathcal{L}_{\text{pred}}(\hat{\boldsymbol{y}^d}, \boldsymbol{y})$ in improving the effectiveness of graph learning. The reason for the performance drop is that the dynamic graphs are only optimized to maximize the graph distance without capturing meaningful information for the final prediction, as it is relatively simple for the dynamic graphs of all samples to maximize their distances between a global prior graph and there is no direct limitation brought by the

hyper-parameter	Electricity		Solar-energy		Pems-bay		Metr-la		Neonatal	
	SRD-GRU	SRD-TCN	SRD-GRU	SRD-TCN	SRD-GRU	SRD-TCN	SRD-GRU	SRD-TCN	SRD-GRU	SRD-TCN
learning rate	1e-3	1e-4	1e-3	1e-4	1e-3	1e-4	1e-3	1e-4	1e-3	1e-3
embedding dimension e	16									
k in topk operation	20								2	
density controller α	3									
teleport probability β	0.05									
network depth Q	3	3	1	3	2	3	3	3	3	10
GM depth S	3	2	3	3	3	2	3	2	2	2
hidden dimension D	8	8	8	8	8	8	8	8	16	8
min-max loss coefficient α_1, α_2	0.0, 10.0	1.0, 1.0	1.0, 10.0	1.0, 1.0	1.0, 1.0	1.0, 1.0	10.0, 1.0	10.0, 1.0	10.0, 1.0	1.0, 1.0

Table 1: The hyperparameters of SRD-GRU and SRD-TCN

Dataset	# Samples	# Variables	Length (T)	Train-Valid-Test ratio
Solar-energy	52,560	137	168	(0.6, 0.2, 0.2)
Electricity	26,304	321	168	(0.6, 0.2, 0.2)
Metr-la	34,272	207	12	(0.7, 0.2, 0.1)
Pems-bay	52,116	325	12	(0.7, 0.2, 0.1)
Neonatal	13,389	3	7680	4-fold cross validation

Table 2: The statistics of datasets.

Ablations \ Horizon L	6	24	48	96
w/o MM	0.872	0.750	0.708	0.695
w/o PL	0.852	0.752	0.718	0.692
SRD-TCN	**0.876**	**0.772**	**0.723**	**0.720**

Table 3: The R^2 results of ablated experiments on Pems-bay. The higher, the better.

prediction losses. Figure 1 illustrates the learning curves of w/o PL, where the graph distance $\mathcal{D}(\tilde{\boldsymbol{A}}^p, \tilde{\boldsymbol{A}}^d)$ keeps increasing rather than first drop to a reasonable value then rise slowly as shown in Figure 4(b) in the main paper, indicating that the dynamic graphs do not first learn reasonable meaningful spatial relations and the minimization phase fails. The R^2 results together with the learning curves show that the prediction losses are required for effective graph learning.

References

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Stevenson, N. J.; Tapani, K.; Lauronen, L.; and Vanhatalo, S. 2019. A dataset of neonatal EEG recordings with seizure annotations. *Scientific data*, 6(1): 1–8.