Supplementary Materials for *Memory In Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity from Spatiotemporal Dynamics*

Yunbo Wang*, Jianjin Zhang*, Hongyu Zhu, Mingsheng Long (✉), Jianmin Wang, and Philip S. Yu
KLiss, MOE; BNRist; School of Software, Tsinghua University, China
Research Center for Big Data, Tsinghua University, China
Beijing Key Laboratory for Industrial Big Data System and Application
{wangyb15,zhang-jj16}@mails.tsinghua.edu.cn, {mingsheng,jimwang}@tsinghua.edu.cn

1. Key Equations of Spatiotemporal LSTM

Spatiotemporal LSTM (ST-LSTM) [1] has four inputs in Equation (1): $\mathcal{X}_t$ is either the input frame for $l = 1$ or the output hidden states by the previous layer $\mathcal{H}_t^{l-1}$ for $l > 1$; $\mathcal{H}_{t-1}^l$ and $\mathcal{C}_{t-1}^l$ are the hidden states and memory cells from the previous timestamp; and $\mathcal{M}_t^{l-1}$ is the spatiotemporal memory cells either from the top layer at the previous timestamp or the last layer at the current timestamp. All states are represented by $\mathbb{R}^{C\times W\times H}$ tensors, where the first dimension is the number of their channels, and the following two dimensions denote the width and height of feature maps. The output of a certain unit at timestamp t and layer l is determined by the spatiotemporal memory $\mathcal{M}_t^{l-1}$ from the previous layer, as well as the temporal memory $\mathcal{C}_{t-1}^l$ from the previous timestamp:

$$
\begin{aligned}
&g_t = \tanh(W_{xg} * \mathcal{X}_t + W_{hg} * \mathcal{H}_{t-1}^l + b_g)\\
&i_t = \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1}^l + b_i)\\
&f_t = \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1}^l + b_f)\\
&\mathcal{C}_t^l = f_t \odot \mathcal{C}_{t-1}^l + i_t \odot g_t\\
&g_t' = \tanh(W_{xg}' * \mathcal{X}_t + W_{mg} * \mathcal{M}_t^{l-1} + b_g')\\
&i_t' = \sigma(W_{xi}' * \mathcal{X}_t + W_{mi} * \mathcal{M}_t^{l-1} + b_i')\\
&f_t' = \sigma(W_{xf}' * \mathcal{X}_t + W_{mf} * \mathcal{M}_t^{l-1} + b_f')\\
&\mathcal{M}_t^l = f_t' \odot \mathcal{M}_t^{l-1} + i_t' \odot g_t'\\
&o_t = \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1}^l + W_{co} * \mathcal{C}_t^l + W_{mo} * \mathcal{M}_t^l + b_o)\\
&\mathcal{H}_t^l = o_t \odot \tanh(W_{1\times 1} * [\mathcal{C}_t^l, \mathcal{M}_t^l]),
\end{aligned}
\tag{1}
$$

where σ is the sigmoid function, $*$ is the convolution, and $\odot$ is the Hadamard product. The input gate i_t, input modulation gate g_t, forget gate f_t and output gate o_t control the spatiotemporal information flow. The biggest highlight of ST-LSTM is its zigzag memory flow $\mathcal{M}$. It provides a great modeling capability of the short-term trends in longer pathways through the vertical layers. However, it also suffers from the problem of blurry predictions as it still uses the simple forget gate inherited from previous methods. The extremely complex non-stationarity cannot be fully captured by such simple temporal transitions.

*Equal contribution, in alphabetical order

References

[1] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NeurIPS*, 2017.