ETH zürich

Exploring interpretable LSTM neural networks over multi-variable data

Sebastian U. Stich (MLO, EPFL)

on behalf of the authors

Tian Guo, COSS, ETH Zurich

Tao Lin, MLO, EPFL

Nino Antulov-Fantulin, COSS, ETH Zurich

Problem formulation

- Multi-variable time series

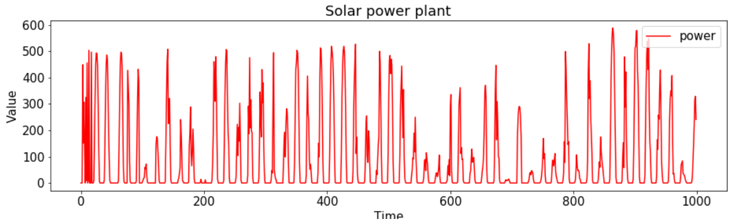
  - Target and exogenous variables

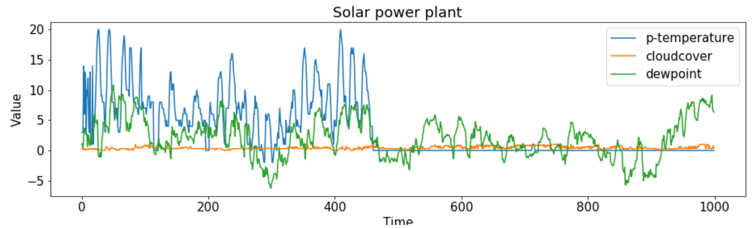
$$\mathbf{X}_T = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$$

$$\mathbf{x}_t = [\mathbf{x}_t^1, \cdots, \mathbf{x}_t^{N-1}, y_t] \quad \mathbf{x}_t \in \mathbb{R}^N$$

  - Predictive model

$$\hat{y}_{T+1} = \mathcal{F}(\mathbf{X}_T)$$





Problem formulation

- Weak interpretability of RNNs on multi-variable data

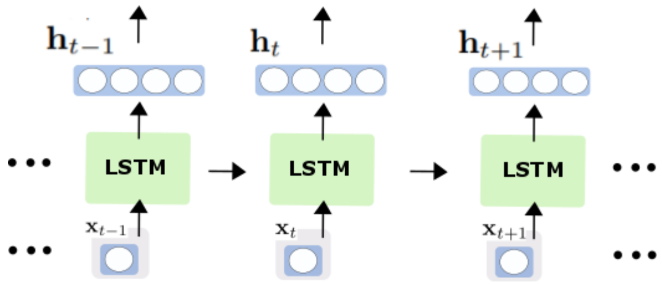
$$\mathbf{X}_T = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$$

  - Multi-variable input to hidden states
    i.e. vectors

$$\mathbf{x}_t = \begin{bmatrix} x_t^1 \\ \vdots \\ x_t^N \end{bmatrix} \xrightarrow{\phi(\mathbf{x}_t)} \mathbf{h}_t = \begin{bmatrix} h_t^1 \\ \vdots \\ h_t^D \end{bmatrix} \longrightarrow \hat{y}_{t+1} = \sigma(\mathbf{h}_t)$$

⚠ No correspondence between hidden states and input variables

⚠ Different dynamics of variables are mingled in hidden states



$\phi(\mathbf{x}_t)$

$$\mathbf{j}_t = \tanh\left(\mathbf{W}_j \left[\mathbf{x}_t \oplus \mathbf{h}_{t-1}\right] + \mathbf{b}_j\right)$$
$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \left[\mathbf{x}_t \oplus \mathbf{h}_{t-1}\right] + \mathbf{b}_f\right)$$
$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \left[\mathbf{x}_t \oplus \mathbf{h}_{t-1}\right] + \mathbf{b}_i\right)$$
$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \left[\mathbf{x}_t \oplus \mathbf{h}_{t-1}\right] + \mathbf{b}_o\right)$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{j}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

Problem formulation

- Interpretable prediction model on multi-variable time series

$$\hat{y}_{T+1} = \mathcal{F}(\mathbf{X}_T)$$

Accurate

- Capture different dynamics of input variables

Interpretable

- Variable importance w.r.t. predictive power
  i.e. which variable is more important for RNNs to perform prediction
- Temporal importance of each variable
  i.e. short or long-term correlation to the target

Interpretable multi-variable LSTM

- IMV-LSTM
  - Key ideas:

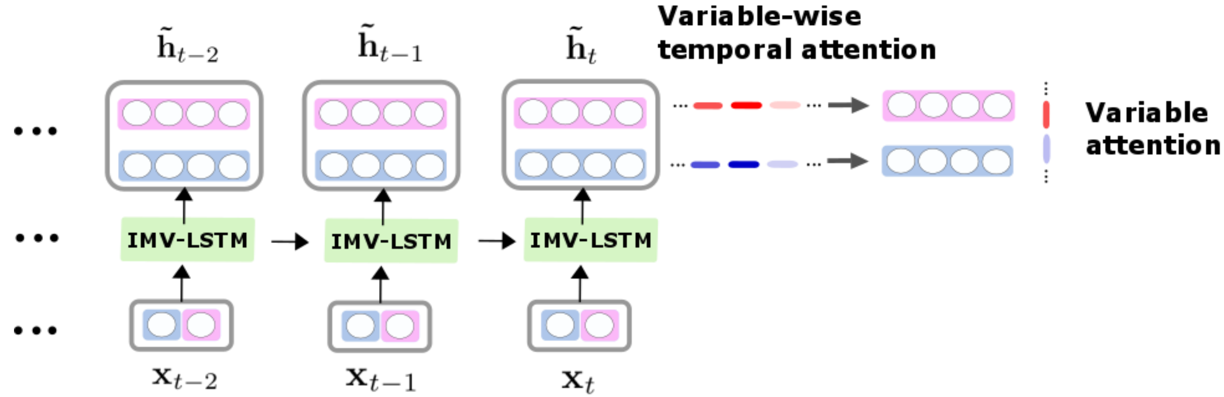
Variable-wise hidden states	Opaque hidden states
↓	
Mixture attention	Temporal and variable level attention
↓	
Learning to interpret and predict	Network parameter and importance learning

IMV-LSTM

- IMV-LSTM with variable-wise hidden states

$$\tilde{\mathbf{h}}_t = \begin{bmatrix} \mathbf{h}_t^1 \\ \vdots \\ \mathbf{h}_t^N \end{bmatrix}^\top$$

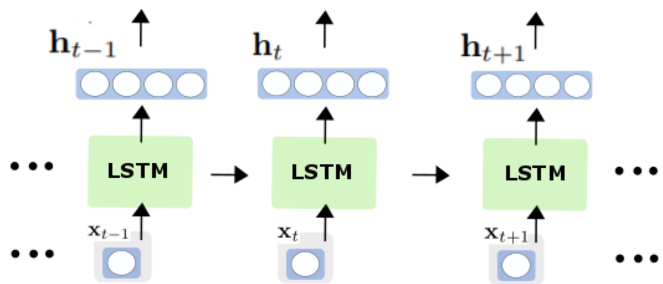
$$\mathbf{h}_t^n \in \mathbb{R}^d$$



- Conventional LSTM with hidden vectors

$$\mathbf{h}_t = \begin{bmatrix} h_t^1 \\ \vdots \\ h_t^D \end{bmatrix}$$

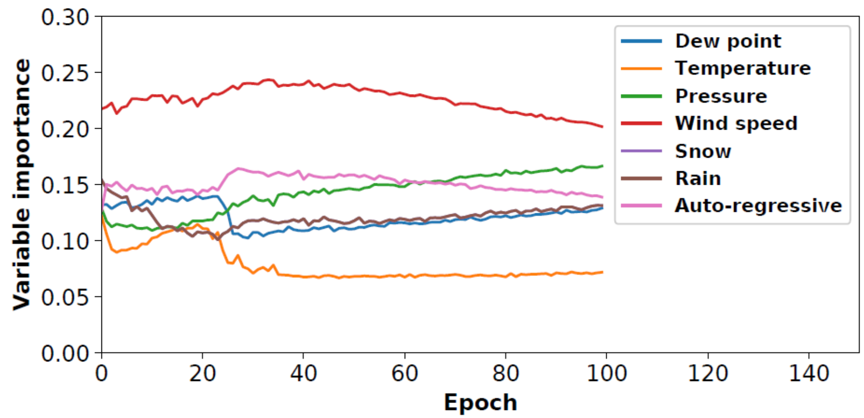
$$h_t^d \in \mathbb{R}$$



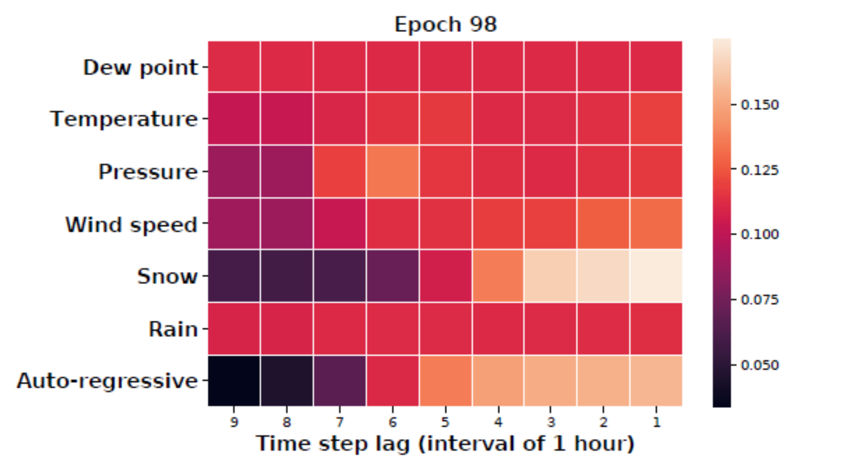
ETH zürich

Results

- Variable importance
  - Learned during the training
  - The higher the value, the more important



- Variable-wise temporal importance
  - The lighter the color, the more important



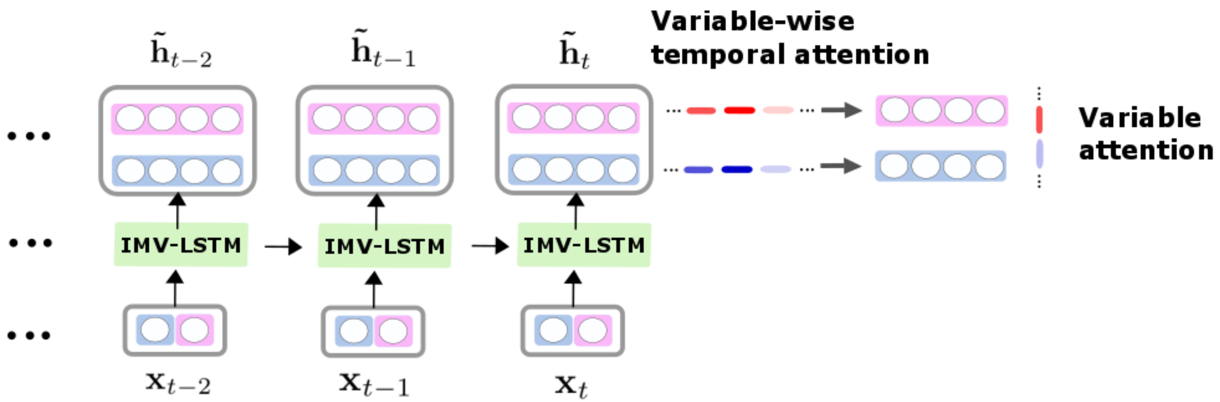
Conclusion

- Explored the internal structures of LSTMs to enable variable-wise hidden states.

- Developed mixture attention and associated learning procedure to quantify variable importance and variable-wise temporal importance w.r.t. the target.

- Extensive experiments provide insights into achieving superior prediction performance and importance interpretation for LSTM.

Backup



Network architecture:

$$\tilde{\mathbf{j}}_t = \tanh\left(\mathcal{W}_j \circledast \tilde{\mathbf{h}}_{t-1} + \mathcal{U}_j \circledast \mathbf{x}_t + \mathbf{b}_j\right)$$

$$\begin{bmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \end{bmatrix} = \sigma\left(\mathbf{W}\,[\mathbf{x}_t \oplus \mathrm{vec}(\tilde{\mathbf{h}}_{t-1})] + \mathbf{b}\right)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathrm{vec}(\tilde{\mathbf{j}}_t)$$

$$\tilde{\mathbf{h}}_t = \mathrm{matricization}(\mathbf{o}_t \odot \tanh(\mathbf{c}_t))$$

Mixture attention to model generative process of the target:

$$\begin{aligned} p(y_{T+1}|\mathbf{X}_T) &= \sum_{n=1}^{N} p(y_{T+1}|z_{T+1}=n, \mathbf{X}_T) \cdot p(z_{T+1}=n|\mathbf{X}_T) \\ &= \sum_{n=1}^{N} p(y_{T+1}\,|\,z_{T+1}=n, \mathbf{h}_1^n, \cdots, \mathbf{h}_T^n) \cdot p(z_{T+1}=n\,|\,\tilde{\mathbf{h}}_1, \cdots, \tilde{\mathbf{h}}_T) \\ &= \sum_{n=1}^{N} p(y_{T+1}\,|\,z_{T+1}=n, \underbrace{\mathbf{h}_T^n \oplus \mathbf{g}^n}_{\text{variable-wise temporal attention}}) \cdot \underbrace{p(z_{T+1}=n\,|\,\mathbf{h}_T^1 \oplus \mathbf{g}^1, \cdots, \mathbf{h}_T^N \oplus \mathbf{g}^N)}_{\text{overall variable attention}} \end{aligned}$$