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

    \[ X_T = \{x_1, \cdots, x_T\} \]
    \[ x_t = [x_t^1, \cdots, x_t^{N-1}, y_t] \quad x_t \in \mathbb{R}^N \]

  - Predictive model

    \[ \hat{y}_{T+1} = \mathcal{F}(X_T) \]
Problem formulation

- Weak interpretability of RNNs on multi-variable data

\[ X_T = \{ x_1, \cdots, x_T \} \]

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

\[ x_t = \begin{bmatrix} x_{t1} \\ \vdots \\ x_{tN} \end{bmatrix} \quad \phi(x_t) \quad h_t = \begin{bmatrix} h_{t1} \\ \vdots \\ h_{tD} \end{bmatrix} \quad \hat{y}_{t+1} = \sigma(h_t) \]

⚠️ No correspondence between hidden states and input variables

⚠️ Different dynamics of variables are mingled in hidden states
Problem formulation

- Interpretable prediction model on multi-variable time series
  \[ \hat{y}_{T+1} = \mathcal{F}(X_T) \]

  - 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
  - Mixture attention
  - Learning to interpret and predict
  - Opaque hidden states
  - Temporal and variable level attention
  - Network parameter and importance learning
IMV-LSTM

- IMV-LSTM with variable-wise hidden states

\[
\tilde{h}_t = \begin{bmatrix} h_1^t \\ \vdots \vline \quad \vdots \\ h_1^N \end{bmatrix}^T
\]

\[
h_i^n \in \mathbb{R}^d
\]

- Conventional LSTM with hidden vectors

\[
h_t = \begin{bmatrix} h_1^t \\ \vdots \vline \quad \vdots \\ h_1^D \end{bmatrix}
\]

\[
h_i^d \in \mathbb{R}
\]
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:

\[
\tilde{j}_t = \tanh \left( W_j \otimes \tilde{h}_{t-1} + U_j \otimes x_t + b_j \right)
\]

\[
\begin{bmatrix}
  i_t \\
  f_t \\
  o_t 
\end{bmatrix} = \sigma \left( W \left[ x_t \oplus \text{vec}(\tilde{h}_{t-1}) \right] + b \right)
\]

\[
c_t = f_t \odot c_{t-1} + i_t \odot \text{vec}(\tilde{j}_t)
\]

\[
\tilde{h}_t = \text{matricization}(o_t \odot \tanh(c_t))
\]

Mixture attention to model generative process of the target:

\[
p(y_{T+1} \mid X_T) = \sum_{n=1}^{N} p(y_{T+1} \mid z_{T+1} = n, X_T) \cdot p(z_{T+1} = n \mid X_T)
\]

\[
= \sum_{n=1}^{N} p(y_{T+1} \mid z_{T+1} = n, h_1^1, \cdots, h_T^1) \cdot p(z_{T+1} = n \mid \tilde{h}_1, \cdots, \tilde{h}_T)
\]

\[
= \sum_{n=1}^{N} p(y_{T+1} \mid z_{T+1} = n, \begin{bmatrix} h_1^1 \oplus g_1^n \end{bmatrix} \cdots \begin{bmatrix} h_T^1 \oplus g_1^n \end{bmatrix}) \cdot p(z_{T+1} = n \mid h_T^1 \oplus g_1^1, \cdots, h_T^N \oplus g_1^N)
\]