Exploring Interpretable LSTM Neural Networks over Multi-Variable Data

Tian Guo [1] **Tao Lin** [2] **Nino Antulov-Fantulin** [1]

1. Appendix

1.1. Interpretable Multi-Variable LSTM

Proof of Lemma 3.3

Proof. For simplicity, we ignore the data instance index m in the following proof.

The log-likelihood of the target conditional on input variables is defined as:

$$\begin{aligned}
&\log p(y_{T+1}\,|\mathbf{X}_T) \\
&= \log \sum_{n=1}^{N} p(y_{T+1}, z_{T+1}=n|\mathbf{X}_T) \\
&\geq \sum_{n=1}^{N} q^n \log p(y_{T+1}|z_{T+1}=n,\mathbf{X}_T)\Pr(z_{T+1}=n,\mathbf{X}_T) \\
&- q^n \log q^n \\
&= \sum_{n=1}^{N} q^n [\log p(y_{T+1}|z_{T+1}=n,\mathbf{X}_T) + \\
&\log \Pr(z_{T+1}=n,\mathbf{X}_T)] + q^n \log q^n - 2q^n \log q^n
\end{aligned} \tag{1}$$

Based on Gibbs inequality, we can have

$$\sum_{n=1}^{N} q^n \log q^n \geq \sum_{n=1}^{N} q^n \log \Pr(z_{T+1}=n|\mathbf{I}) \tag{2}$$

Since $\mathbf{I} \in R^N_{\geq 0}$, $\sum_{n=1}^{N} \mathrm{I}_n = 1$, it can parameterize a categorical distribution on z_{T+1}.

Then introducing Eq. 2 into Eq. 1, we can obtain

$$\begin{aligned}
&\log p(y_{T+1}\,|\mathbf{X}_T) \geq \sum_{n=1}^{N} q^n [\log p(y_{T+1}|z_{T+1}=n,\mathbf{X}_T) \\
&+ \log \Pr(z_{T+1}=n,\mathbf{X}_T)] + q^n \log \Pr(z_{T+1}=n|\mathbf{I}) - 2q^n \log q^n \\
&\approx \mathbb{E}_{q^n}[\log p(y_{T+1,}\,|\,z_{T+1,}=n,\mathbf{h}^n_T \oplus \mathbf{g}^n)] + \\
&\mathbb{E}_{q^n}[\log \Pr(z_{T+1,}=n\,|\,\mathbf{h}^1_T \oplus \mathbf{g}^1,\cdots,\mathbf{h}^N_T \oplus \mathbf{g}^N)] \\
&\quad + \mathbb{E}_{q^n}[\log \Pr(z_{T+1,}=n\,|\,\mathbf{I}) - 2q^n \log q^n
\end{aligned} \tag{3}$$

During the EM process, after the E-step, $2q^n \log q^n$ will be a constant and is not involved in the optimization process. In the M-step, minimizing the negative log-likelihood amounts to minimize the loss function as follows:

$$\begin{aligned}
\mathcal{L}(\Theta,\mathbf{I}) = -\sum_{m=1}^{M} &\mathbb{E}_{q^n_m}[\log p(y_{T+1,m}\,|\,z_{T+1,m}=n,\mathbf{h}^n_T \oplus \mathbf{g}^n)] \\
&- \mathbb{E}_{q^n_m}[\log \Pr(z_{T+1,m}=n\,|\,\mathbf{h}^1_T \oplus \mathbf{g}^1,\cdots,\mathbf{h}^N_T \oplus \mathbf{g}^N)] \\
&- \mathbb{E}_{q^n_m}[\log \Pr(z_{T+1,m}=n\,|\,\mathbf{I})]
\end{aligned}$$

□

1.2. Experiments

In this part, we provide complementary experiment results as well as the insights from the results.

NASDAQ is the dataset from (**?**). It contains 81 major corporations under NASDAQ 100, as exogenous time series. The index value of the NASDAQ 100 is the target series. The frequency of the data collection is minute-by-minute. The first 35,100, the following 2,730 and the last 2,730 data points are respectively used as the training, validation and test sets.

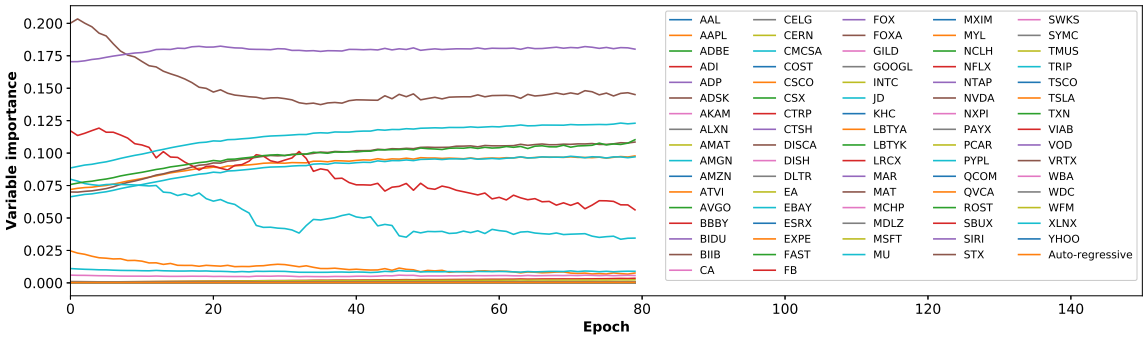
1.2.1. Prediction performance analysis

1.2.2. Model Interpretation

In the following Table 2, 3, and 4, we list the full ranking of variables of the datasets by each approach. Variables associated with the importance or attention values are ranked in decreasing order.

[1]ETH, Zürich, Switzerland [2]EPFL, Switzerland. Correspondence to: Tian Guo <tian.guo@gess.ethz.ch>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



(a) Variable importance w.r.t. epochs.

(b) Variable-wise temporal importance at different epochs.

Figure 1: IMV-LSTM on NASDAQ dataset. (Best viewed in color)

Table 1: RMSE and MAE with std. errors

Dataset	NASDAQ
STRX	$0.41 \pm 0.01, 0.35 \pm 0.02$
ARIMAX	$0.34 \pm 0.02, 0.23 \pm 0.03$
RF	$0.31 \pm 0.02, 0.27 \pm 0.03$
XGT	$0.28 \pm 0.01, 0.23 \pm 0.02$
ENET	$0.31 \pm 0.03, 0.21 \pm 0.01$
DUAL	$0.31 \pm 0.003, 0.21 \pm 0.002$
RETAIN	0.12 ± 0.07 , 0.11 ± 0.06
IMV-Full	$0.27 \pm 0.01, 0.23 \pm 0.01$
IMV-Tensor	$\mathbf{0.09 \pm 0.005}, \mathbf{0.07 \pm 0.004}$

1.3. Discussion

In this part, we summarize the insights from the experiments.

Prediction performance For multi-variable data, capturing individual variable's behaviors and their interaction is the key for both prediction and interpretation. Conventional hidden states in standard LSTMs consume the data from all input variables at each step, while our IMV-LSTM family decomposes the hidden states by defining variable data flows for each hidden state element.

In the experiments, IMV-Full and IMV-Tensor outperform baselines using the traditional hidden states. Multi-variable data potentially carries different dynamics. Conventional hidden states mix the data of all input variables, thereby failing to explicitly capture individual dynamics. In the multi-variable setting, these opaque hidden states are a burden to both prediction and interpretation.

On the contrary, IMV-Tensor models individual variables and then uses mixture attention to capture the variable interaction by variable-wise hidden states. It achieves superior prediction performance and enables the interpretability on both temporal and variable levels.

Effectiveness of importance values For LSTM networks on multi-variable data, importance values inherently learned by the network are more suitable for retaining useful variables for predicting.

By choosing the variables based on the learned importance value, IMV-LSTM family mostly retains the prediction performance and presents lower prediction errors on two datasets. The importance value in IMV-LSTM is derived during the training and therefore it is able to effectively identify the variables used by IMV-LSTM to minimize the loss function, i.e. maximize the prediction accuracy.

Pearson correlation variable selection leads to the quality loss in prediction performance, i.e. higher errors. Pearson correlation measures the linear correlation and pre-selecting variables based on it neglects the potential non-linear correlation in data indispensable for LSTMs to capture.

Table 2: Variable importance ranking by IMV-LSTM on NASDAQ dataset.

Dataset	Method	Rank of variables according to importance
NASDAQ	IMV-LSTM	'ADSK', 0.00023858716, 'PAYX', 0.00023869322, 'AAL', 0.00023993119, 'MYL', 0.00024015515, 'CA', 0.00024144033], 'FOX', 0.00024341498, 'EA', 0.00024963205], 'BIDU', 0.00025009923, 'MCHP', 0.00025015706, 'QVCA', 0.00025018162, 'NVDA', 0.00025088928, 'WBA', 0.00025147066, 'LRCX', 0.00025165512, 'TSCO', 0.00025247637, 'CTSH', 0.00025284023, 'CSX', 0.00025417344, 'COST', 0.00025498777, 'BIIB', 0.00025547648, 'LBTYA', 0.00025680827, 'SIRI', 0.00025686354, 'ADBE', 0.00025687047, 'MDLZ', 0.00025788756, 'LBTYK', 0.00025885308, 'INTC', 0.00025894548, 'TSLA', 0.0002592771, 'WFM', 0.00025941888, 'SBUX', 0.00025953245, 'AVGO', 0.00026012328], 'CTRP', 0.00026024296, 'AMZN', 0.00026168497, 'ALXN', 0.00026173133, 'AMGN', 0.0002617908, 'GILD', 0.0002619058, 'VOD', 0.00026195042, 'ROST', 0.00026237246, 'NXPI', 0.0002624988, 'KHC', 0.0002625609, 'ADP', 0.0002626155, 'WDC', 0.00026269013, 'QCOM', 0.00026288, 'TMUS', 0.00026333777, 'AMAT', 0.00026334616, 'AKAM', 0.00026453246, 'PCAR', 0.00026510606, 'CERN', 0.00026535543, 'VRTX', 0.00026579297, 'MU', 0.00026719182, 'MAR', 0.00026789604, 'TXN', 0.00026821258, 'GOOGL', 0.0002684545, 'ESRX', 0.00026995668, 'ATVI', 0.0002703378, 'STX', 0.0002708045, 'FAST', 0.00027182887, 'EXPE', 0.0002747627, 'CELG', 0.00027897576, 'PYPL', 0.00027971127, 'MXIM', 0.0002802631, 'NFLX', 0.00028330996, 'BBBY', 0.00028975168, 'SYMC', 0.0002932911, 'CMCSA', 0.00031882498, 'SWKS', 0.00034903747, 'DLTR', 0.0004099159, 'YHOO', 0.0004359138, 'VIAB', 0.00046212596, 'Auto-regressive', 0.0004718905, 'MAT', 0.0008193875, 'MSFT', 0.002350653, 'ADI', 0.0035426863, 'DISH', 0.0056709386, 'AAPL', 0.007597621, 'EBAY', 0.008922806, 'JD', 0.03449823, 'FB', 0.056254942, 'XLNX', 0.09711476, 'CSCO', 0.09782402, 'DISCA', 0.108503476, 'NCLH', 0.11029968, u'TRIP', 0.12302372, 'FOXA', 0.14510903, 'NTAP', 0.18010232

Table 3: Variable importance ranking by DUAL and RETAIN methods on NASDAQ dataset.

Dataset	Method	Rank of variables according to importance
NASDAQ	DUAL	'NXPI', 0.003557, 'QCOM', 0.003564, 'FOX', 0.003566, 'NTAP', 0.003566, 'CELG', 0.003566, 'FOXA', 0.003567, 'PAYX', 0.003567, 'AAPL', 0.003567, 'WFM', 0.003567, 'ADSK', 0.003567, 'SBUX', 0.003567, 'STX', 0.003567, 'AKAM', 0.003567, 'DISH', 0.003567, 'AVGO', 0.003567, 'XLNX', 0.003567, 'AAL', 0.003567, 'FAST', 0.003567, 'TMUS', 0.003567, 'LRCX', 0.003567, 'NCLH', 0.003567, 'MCHP', 0.003567, 'MSFT', 0.003567, 'MU', 0.003567, 'NFLX', 0.003567, 'NVDA', 0.003567, 'PCAR', 0.003567, 'SIRI', 0.003567, 'MAR', 0.003567, 'TXN', 0.003567, 'ROST', 0.003567, 'CMCSA', 0.003567, 'ADI', 0.003567, 'ADP', 0.003567, 'DISCA', 0.003567, 'AMAT', 0.003567, 'WDC', 0.003567, 'CSX', 0.003567, 'WBA', 0.003567, 'GOOGL', 0.003622, 'COST', 0.003678, 'INTC', 0.003712, 'CTSH', 0.003908, 'BBBY', 0.004027, 'TRIP', 0.004881, 'MAT', 0.004956, 'ATVI', 0.005121, 'LBTYK', 0.00523, 'CERN', 0.00524, 'CTRP', 0.005283, 'ALXN', 0.00536, 'VOD', 0.005369, 'VRTX', 0.005433, 'LBTYA', 0.005445, 'MXIM', 0.00554, 'BIIB', 0.005554, 'EBAY', 0.005555, 'BIDU', 0.005605, 'FB', 0.005654, 'VIAB', 0.005685, 'GILD', 0.005695, 'AMGN', 0.005716, 'MYL', 0.005737, 'YHOO', 0.006166, 'KHC', 0.006555, 'AMZN', 0.006605, 'CSCO', 0.007836, 'ESRX', 0.010614, 'SWKS', 0.012777, 'MDLZ', 0.017898, 'CA', 0.02198, 'EXPE', 0.024373, 'QVCA', 0.026462, 'EA', 0.027808, 'TSLA', 0.043082, 'ADBE', 0.043829, 'JD', 0.071079, 'SYMC', 0.081596, 'PYPL', 0.087612, 'DLTR', 0.119737, 'TSCO', 0.122887
	RETAIN	'DLTR', 0.000866, 'QVCA', 0.001128, 'TSLA', 0.00119, 'PYPL', 0.00128, 'EA', 0.001439, 'EXPE', 0.001502, 'CA', 0.001713, 'TSCO', 0.001737, 'SYMC', 0.002334, 'ADBE', 0.00252, 'JD', 0.002607, 'AMZN', 0.003367, 'CSCO', 0.003543, 'KHC', 0.003996, 'CTSH', 0.004695, 'NXPI', 0.004865, 'EBAY', 0.004963, 'SWKS', 0.005011, 'MXIM', 0.005135, 'MYL', 0.005541, 'COST', 0.006052, 'BIDU', 0.006534, 'GOOGL', 0.006906, 'INTC', 0.007153, 'GILD', 0.007212, 'ESRX', 0.007512, 'NTAP', 0.007695, 'QCOM', 0.008037, 'CELG', 0.008168, 'MDLZ', 0.008829, 'AMGN', 0.008998, 'FOX', 0.009943, 'VIAB', 0.010123, 'AAPL', 0.010157, 'FB', 0.010359, 'YHOO', 0.010744, 'PAYX', 0.010899, 'BBBY', 0.01117, 'AKAM', 0.012054, 'BIIB', 0.012069, 'NFLX', 0.012266, 'ADSK', 0.012319, 'DISH', 0.012338, 'LBTYA', 0.012697, 'FOXA', 0.01282, 'MCHP', 0.012833, 'WFM', 0.012869, 'STX', 0.012887, 'VRTX', 0.013318, 'SBUX', 0.013458, 'VOD', 0.013798, 'ALXN', 0.013878, 'CTRP', 0.013963, 'SIRI', 0.01475, 'CERN', 0.014777, 'LBTYK', 0.014799, 'ATVI', 0.015651, 'AVGO', 0.016382, 'CMCSA', 0.016531, 'TXN', 0.016977, 'LRCX', 0.017131, 'AMAT', 0.017378, 'ROST', 0.017399, 'MU', 0.018045, 'TRIP', 0.018236, 'MAT', 0.018297, 'Auto-regressive', 0.018626, 'WDC', 0.019083, 'DISCA', 0.019233, 'FAST', 0.019392, 'CSX', 0.019734, 'WBA', 0.019984, 'AAL', 0.021188, 'ADI', 0.021215, 'NCLH', 0.022932, 'NVDA', 0.022994, 'TMUS', 0.024187, 'MSFT', 0.026354, 'ADP', 0.028515, 'MAR', 0.028783, 'PCAR', 0.029459, 'XLNX', 0.03248

Table 4: Variable importance ranking on PLANT and SML datasets.

Dataset	Method	Rank of variables according to importance
PLANT	IMV-LSTM	'Dew-point', 0.040899094, 'Wind-bearing', 0.04476319, 'Pressure', 0.06180005, 'P-temperature', 0.07244386, 'Auto-regressive', 0.1083069, 'Temperature', 0.11868146, 'Irradiance', 0.12043289, 'Humidity', 0.13192631, 'Cloud-cover', 0.14283147, 'Wind-speed', 0.15791483
	DUAL	'Irradiance', 0.06128826, 'Dew-point', 0.066655099, 'Temperature', 0.071131147, 'Wind-speed', 0.094427079, 'Wind-bearing', 0.106529392, 'P-temperature', 0.115000054, 'Pressure', 0.115962856, 'Cloud cover', 0.144996881, 'Humidity', 0.224009201
	RETAIN	'Dewpoint', 0.031317, 'Temperature', 0.037989, 'Wind-bearing', 0.044226, 'Wind-speed', 0.052027, 'P-temperature', 0.053034, 'Cloud cover', 0.138427, 'Irradiance', 0.142899, 'Auto-regressive', 0.143269, 'Humidity', 0.172893, 'Pressure', 0.183919
SML	IMV-LSTM	'Outdoor temp.', 0.008530081, 'Outdoor humidity', 0.0120737655, 'Sun irradiance', 0.012943255, 'CO2 dining', 0.01563413, 'Sunlight in south', 0.01569774, 'Sun dusk', 0.015769556, 'Wind', 0.015868865], 'Forecast temp.', 0.015990425, 'Sunlight in west', 0.01609429, 'Lighting dining', 0.016338758, 'Humid. dining', 0.016379833, 'Sunlight in east', 0.016386982, 'Auto-regressive', 0.016530316, 'Temp. dining', 0.01663947, 'Lighting room', 0.18322693, 'CO2 room', 0.26715645, 'Humid. room', 0.33873916
	RETAIN	'Humid. dining', 0.012169, 'Humid. room', 0.014563, 'Sunlight in south', 0.018446, 'Lighting room', 0.018732, 'Outdoor humidity', 0.019388, 'Sunlight in west', 0.02219, 'Sunlight in east', 0.036744, 'CO2 room', 0.036864, 'CO2 dining', 0.037174, 'Sun dusk', 0.040011, 'Sun irradiance', 0.04075, 'Wind', 0.041191, 'Lighting dining', 0.054166, 'Forecast temp.', 0.133079, 'Outdoor temp.', 0.144314, 'Auto-regressive', 0.164673, 'Temp. dining', 0.165543