AT-LSTM: An Attention-based LSTM Model for Financial Time Series Prediction

To cite this article: Xuan Zhang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **569** 052037

IOP Conf. Series: Materials Science and Engineering **569** (2019) 052037 doi:10.1088/1757-899X/569/5/052037

AT-LSTM: An Attention-based LSTM Model for Financial Time Series Prediction

Xuan Zhang, Xun Liang*, Aakas Zhiyuli, Shusen Zhang, Rui Xu, and Bo Wu

Information School, Renmin University of China, Beijing, 100872, China

*Corresponding author's e-mail: xunliangruc@163.com

Abstract. This paper proposes an attention-based LSTM (AT-LSTM) model for financial time series prediction. We divide the prediction process into two stages. For the first stage, we apply an attention model to assign different weights to the input features of the financial time series at each time step. In the second stage, the attention feature is utilized to effectively select the relevant feature sequences as input to the LSTM neural network for the prediction in the next time frame. Our proposed framework not only solves the long-term dependence problem of time series prediction effectively, but also improves the interpretability of the time series prediction methods based on the neural network. In the end of this paper, we conducted experiments on financial time series prediction task with three real-world data sets. The experimental results show that our framework for time series pre-diction is state-of-the-art against the baselines.

1. Introduction

Financial market prediction [1] is a classic research problem in quantitative finance and neural networks research area. Financial time series often present characteristics such as volatility [2], non-stationarity [3], periodicity [4], nonlinearity [5] and long-term dependence [6]. Traditional statistical models, including the multiple regression method, the Autoregressive integrated moving average (ARIMA) model [7] and the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model [8], can accurately capture the volatility and periodicity of financial series and are widely used in the real world. The statistical models are also highly interpretable. But it is difficult for them to analyse non-stationary sequences and capture the nonlinear relationships of financial time series. The neural network-based Recurrent neural networks (RNNs) model [9] has a timing concept and is very flexible in dealing with non-stationary sequences and capturing nonlinear relationships. However, the traditional RNN models have a vanishing gradient problem [10], and few of them can capture the long-term dependence of time series appropriately. In addition, the neural network model often uses raw time series as input [11], which makes it difficult to explain what role each input feature sequence plays in the prediction.

To solve the aforementioned problem, we propose an attention model [12] to assign different weights to each input financial feature sequence to replace the raw time series. These weights are prediction, thereby increasing the interpretability of the model. The essence of the attention mechanism is to train a model to selectively learn the input and associate the output sequence with it. The attention model was first used in image recognition tasks [13] in computer vision and was subsequently applied to graphic transformation. Now attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of

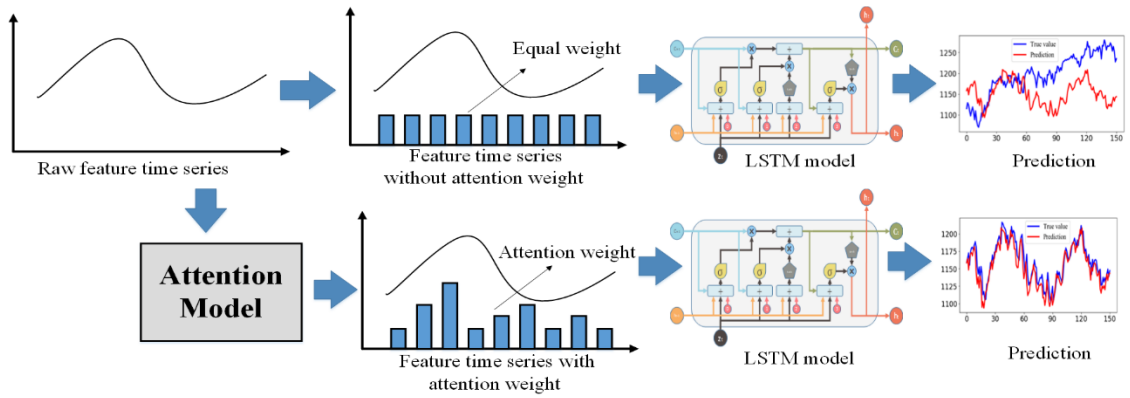

Figure 1. Comparison of pipelines in traditional LSTM-based model and AT-LSTM. The traditional LSTM-based financial time series prediction framework directly takes the raw feature sequences as input, with all input feature treated equally. AT-LSTM gives important features a higher weight and reduce the weight of redundant features to guarantee good prediction accuracy.

dependencies without regard to their distance in the input or output sequences [14].

We use the LSTM [15] model instead of the RNN model to solve the long-term dependence problem in time series prediction. LSTM is a special RNN structure. What distinguishes it from RNN is that it adds a processor called memory cell to the algorithm to judge whether the information is useful or not. Three gates are placed in a memory cell, called forget gate, input gate and output gate. A message enters the LSTM network and can be judged according to the rules. Only the information that complies with the algorithm's certification will be retained, and information that does not match will be forgotten through the forget gate. This makes it possible to sharply capture the long-term dependence of the financial time series.

Therefore, the main contributions of this paper mainly include the following: we introduce the attention model to efficiently extract the features of financial time series and use them as input to LSTM deep learning model [16]. Compared with traditional statistical models, AT-LSTM can efficiently process non-stationary sequences and capture nonlinear relationships. Compared with deep learning models such as RNN, AT-LSTM can avoid long-term dependence problems and have better interpretability. The attention mechanism in the model makes it easy to understand how the information in the input sequence affects the final generated sequence during the model output process. This may help us explore the internal operation mechanism of the model and debug some specific input-output. In addition, the experimental results on three real world datasets demonstrate that AT-LSTM performs better than baseline models.

2. AT-LSTM model

Our attention-based LSTM (AT-LSTM) model for financial time series prediction, consists of two parts: the attention model and the LSTM deep learning model. For the attention model section, the input sequence is a raw feature time series describing the financial market history information. The attention mechanism can adaptively select the most relevant input features and give higher weights to the corresponding original feature sequence. Then we use the output of the attention model as the LSTM deep learning model’s input to predict the financial market. That is using the previous information with attention weight to predict the closing price of the next trading day. The comparison of pipelines in traditional LSTM model and AT-LSTM for financial time series prediction is shown in figure 1. Since the encoding of the attention model requires the LSTM units, we will introduce the LSTM neural network first and then discuss the attention model in the following sections.

2.1. Raw time series

For a given input raw time series, *i.e.*, $\mathbf{X}=(\mathbf{x}^1,\mathbf{x}^2,...,\mathbf{x}^n)^{\mathrm{T}}=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m)\in\mathbf{R}^{(n\times m)}$, n is the number of feature sequences, m is the length of the window. $\mathbf{x}^k=(x_1^k,x_2^k,...,x_m^k)^{\mathrm{T}}\in\mathbf{R}^m$ is used to represent a sequence (vector) of length m. For financial time series prediction, this sequence can be transaction volume, opening price and so on. We also employ $\mathbf{x}_t=(x_t^1,x_t^2,...,x_t^n)^{\mathrm{T}}\in\mathbf{R}^n$ to represents a set vector of n features at time t.

2.2. LSTM model

The Long Short-Term Memory model is defined as follows. Let $\mathbf{x}_t$, $\mathbf{h}_t$ and $\mathbf{C}_t$ be the input, control state, and cell state at timestep t. Given a sequence of inputs $(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m)$, the LSTM computes the h-sequence $(\mathbf{h}_1,\mathbf{h}_2,...,\mathbf{h}_m)$, and the C-sequence $(\mathbf{C}_1,\mathbf{C}_2,...,\mathbf{C}_m)$ as follows:

$$\mathbf{f}_t=\sigma(\mathbf{W}_f\cdot[\mathbf{h}_{t-1},\mathbf{x}_t]+\mathbf{b}_f) \tag{1}$$

$$\mathbf{f}_t=\sigma(\mathbf{W}_f\cdot[\mathbf{h}_{t-1},\mathbf{x}_t]+\mathbf{b}_f) \tag{2}$$

$$\mathbf{c}_t=\tanh(\mathbf{W}_C\cdot[\mathbf{h}_{t-1},\mathbf{x}_t]+\mathbf{b}_C) \tag{3}$$

$$\mathbf{C}_t=\mathbf{f}_t*\mathbf{C}_{t-1}+\mathbf{i}_t*\mathbf{c}_t \tag{4}$$

$$\mathbf{o}_t=\sigma(\mathbf{W}_o\cdot[\mathbf{h}_{t-1},\mathbf{x}_t]+\mathbf{b}_o) \tag{5}$$

$$\mathbf{h}_t=\mathbf{o}_t*\tanh(\mathbf{C}_t) \tag{6}$$

Where σ is a logistic sigmoid function, $*$ is an element-wise multiplication, and $\mathbf{C}t$ is the state of the cell that needs to be updated. $\mathbf{W}_f$,$\mathbf{W}_i$,$\mathbf{W}_C$,$\mathbf{W}_o$ and $\mathbf{b}_f$, $\mathbf{b}_i$, $\mathbf{b}_C$, $\mathbf{b}_o$ are parameters of the model and can be learned. In addition, $\mathbf{f}_t$, $\mathbf{i}_t$ and $\mathbf{o}_t$ are also called forget gate, input gate and output gate. Each LSTM unit has a memory cell with state $\mathbf{C}_t$ at time t, which is controlled by the above three gates

2.3. Attention model

An important feature of human perception is that it does not immediately deal with all inputs from the outside world. Instead, humans will first focus on the important parts to get the information they need. Similarly, the importance of various trading information in financial markets is also different, some information is redundant [17], and the other may be critical [18]. In finance index prediction, we also need to focus on key features first and discard redundant features. Therefore, the effective information of different time periods can be used to establish a financial prediction model and guide the decision.

Inspired by the above information, we propose an attention model to optimize the input feature sequence in financial time prediction. An attention mechanism [19] can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

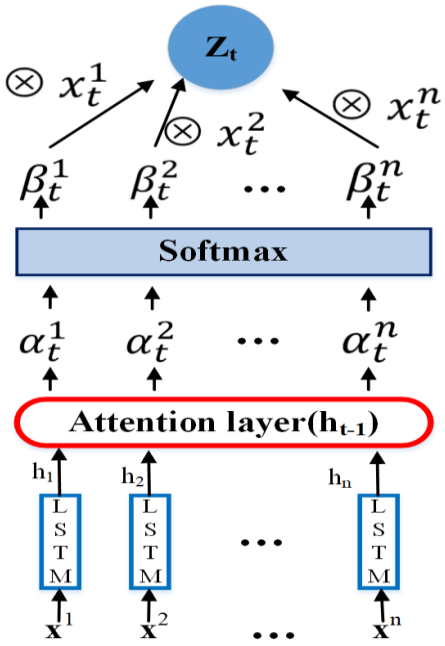

Figure 2. Illustration of the architecture of the attention model.

The process of the generation of attention weights and new input features based on attention is summarized in figure 2. In the first part, we map $\mathbf{x}_t$ to $\mathbf{h}_t$ with

$$\mathbf{h}_t = f_1(\mathbf{h}_{t-1}, \mathbf{x}_t) \tag{7}$$

Where f_1 is a non-linear activation function and $\mathbf{h}_t \in \mathbf{R}^s$ is the hidden state at time t, s is the size of the hidden state. LSTM is adopted as $f1$ in order to avoid the long-term dependence problem which occurs in time series prediction.

In the second part, we construct an attention mechanism through a deterministic attention model. For a certain feature sequence $\mathbf{x}^k = (x_1^k, x_2^k, ..., x_m^k)^{\mathrm{T}} \in \mathbf{R}^m$, by referring to the previous hidden state $\mathbf{h}_{t-1}$ and cell state $\mathbf{C}_{t-1}$ in the LSTM unit, we define:

$$\alpha_t^k = \mathbf{v}^{\mathrm{T}} \tanh(\mathbf{W}_1 \cdot [\mathbf{h}_{t-1}, \mathbf{C}_{t-1}] + \mathbf{W}_2 \mathbf{x}^k) \tag{8}$$

$$\beta_t^k = softmax(\alpha_t^k) = \frac{\exp(\alpha_t^k)}{\sum_{i=1}^{n} \exp(\alpha_t^k)} \tag{9}$$

The vector v and matrices $\mathbf{W}_1$, $\mathbf{W}_2$ are learnable parameters of the model. The vector $\boldsymbol{\alpha}^k$ has length m and its i-th item measures the importance of the k-th input feature sequence at time t. These items are normalized by softmax. $\boldsymbol{\beta}^k$ is an attention weight, which contains a score of how much attention should be put on the k-th feature sequences. We can also get the output of the attention model at time t, *i.e.*, the weighted input feature sequence $\mathbf{z}_t$ as follows:

$$\mathbf{z}_t = (\beta_t^1 x_t^1, \beta_t^2 x_t^2, ..., \beta_t^n x_t^n)^{\mathrm{T}} \tag{10}$$

Then we have $\mathbf{x}_t$ in equation (1)-(7) replaced by the newly computed $\mathbf{z}_t$ to update the attention model and we finally finish the work of converting raw time series to attention-based time series.

Traditional financial time series prediction framework based on recurrent neural networks often uses raw time series as input, and all input feature sequences are treated equivalently. However, the newly obtained $\mathbf{z}_t$ can pay more attention to the specific input feature sequence, extract the key feature sequences effectively and eliminate the influence of the redundant feature sequences because of the attention weight. Theoretically, there will be a better prediction accuracy with $\mathbf{z}_t$ as the input of the LSTM deep learning model.

3. Experiments

In this section, we will do empirical research on three data sets in order to demonstrate the validity of our financial time series prediction framework. First, we will give a basic introduction to the dataset and experimental parameter settings, and then we will compare it with other prediction methods.

3.1. Datasets and settings

To test the performance of different methods, we use three different data sets: Russell 2000, DJIA, and Nasdaq. The Russell 2000 index is an index measuring the performance of approximately 2,000 small-cap companies in the Russell 3000 Index, which is made up of 3,000 of the biggest U.S. stocks, calculated by weighted average method. DJIA (Dow Jones Industrial Average) is the most influential and widely used stock price index in the world, with a representative portion of the company's stock listed on the New York Stock Exchange. The Nasdaq Index is an average stock price index that reflects changes in the Nasdaq Stock Market. In the above dataset, we use the data of 6885 days from January 2, 1991 to April 30, 2018. The data of first 4500 days is used as the training set, and the data of next 1300 days is used as the validation set. The rest of the data is used as the test set. All data can be downloaded from *finance@yahoo.com*.

The attention mechanism of AT-LSTM has two main parameters, the size of the window m and the size of the hidden state s. After repeated experiments on the validation set, the model performs ideally with the following parameters settings: m=10, s=32. For the LSTM deep learning model part, the LSTM layer number is 2, the number of hidden neurons is 8, the number of training time steps is 20, that is, the information of the past 4 weeks is used to predict the closing price of the next trading day, the batch size and epoch is set to be 50 and 5000, respectively.

3.2. Evaluation metric

In order to compare the effectiveness of different methods for financial time series prediction, we have adopted mean absolute percentage error (MAPE) as evaluation metric. MAPE is a measure of prediction accuracy of a forecasting method in statistics and is more persuasive when comparing the performance of the model on different data sets, since it not only considers the deviation between the predicted value and the true value, but also considers the ratio between them. MAPE is defined by equation (11), where Y_i and y_i are true values and predicted values, and N is the size of the test set.

$$\mathbf{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - y_i}{Y_i}\right| \times 100\% \tag{11}$$

3.3. Results and discussion

To demonstrate the effectiveness of our method, we compared AT-LSTM with ARIMA and LSTM. ARIMA and LSTM are classical methods for time series prediction of traditional statistical models and neural network models, respectively. The results are shown in Table 1. All MAPE values were selected from the optimal experimental results under the above parameter settings. It is not difficult to see that the performance of ARIMA is worse than LSTM and AT-LSTM on all three datasets, because ARIMA can only capture linear relationships substantially, and LSTM-based models can effectively capture nonlinear relationships as well as long-term dependence in time series. By comparing the prediction results of the LSTM and AT-LATM models, it can be seen that AT-LSTM's performance is better. This suggests that attention mechanism can adaptively select the most relevant input features and effectively improve the accuracy of the prediction to some extent. It also implies that input feature sequences (transaction volume, opening price, closing price, etc.) are not equally important, which is consistent with our financial knowledge.

Table 1. Comparison with baseline methods

Models	Russell 2000	DJIA	Nasdaq
ARIMA	0.01863	0.01420	0.01850
LSTM	0.00838	0.00625	0.00805
AT-LSTM	0.00550	0.00486	0.00545

To present and analyse the predictions more intuitively, we show the comparison of Russell 2000 index and prediction of our method over a long time-period in figure 3.In addition, we also present some specific shorter time-periods where the market is behaving very differently including fluctuating (figure 3-b), bearish (figure 3-c) and bullish (figure 3-d). It is not difficult to see that our prediction fits the real-word data curve very well. But there is still a time lag in the prediction and our method sometimes is not sensitive for extreme conditions (seen at the bottom right of figure 3-c), which may be a direction for our future work.

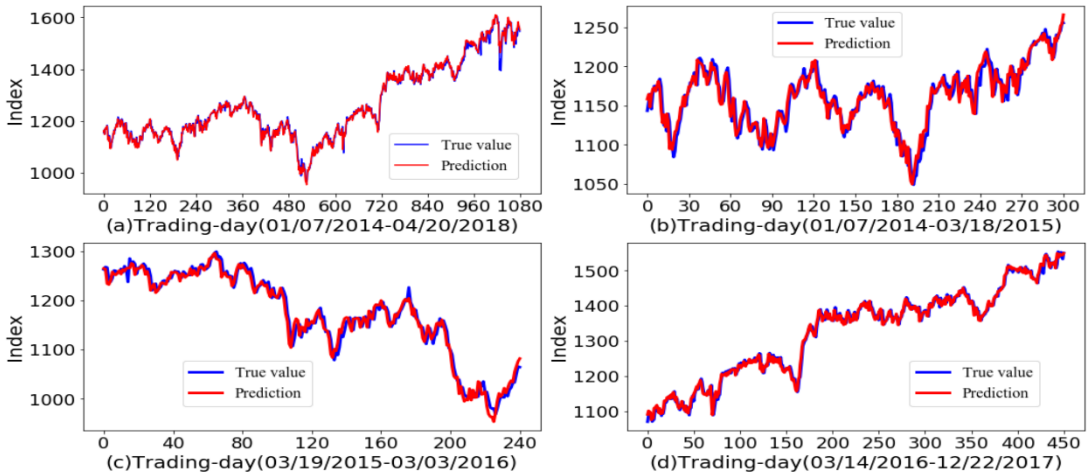

Figure 3. Russell 2000 vs. AT-LSTM. (a), (b), (c), (d) shows the performance of AT-LSTM over a long time-period, a fluctuating time-period, a bearish time-period and a bullish time-period of Russell 2000, respectively.

4. Conclusion

In this work, we proposed a financial time series prediction framework, namely AT-LSTM, based on the attention mechanism. The attention mechanism can adaptively assign different weight to each input feature sequence to automatically choose the most relevant features. Thus, our model can effectively capture long-term dependence in the time series. Experiments on three real-world datasets show that our method is state-of-the-art against all baselines.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 71531012, 71271211), the Natural Science Foundation of Beijing (Grant No. 4172032). The corresponding author can be contacted by *xunliangruc@163.com*.

References

[1] Chen C H. “Neural networks for financial market prediction,” IEEE World Congress on Computational Intelligence, pp.1199-1202,2002.
[2] Michael McAleer. “Automated inference and learning in modeling financial volatility,” Econometric Theory, pp.232-261,2005.
[3] Simonds R R, Lamotte L R, Mcwhorter A. “Testing for nonstationarity of market: An exact test and power considerations,” Journal of Financial & Quantitative Analysis, pp.209-220,1986.
[4] Andersen T G, Bollerslev T. “Intraday periodicity and volatility persistence in financial markets,” Journal of Empirical Finance, pp.115-158, 1997.
[5] So M K P, Chen C W S, Chen M T. “A Bayesian threshold nonlinearity test for financial time series,” Journal of Forecasting, pp.61-75, 2005.
[6] Greene M T, Fielitz B D. “Long-term dependence in common stock returns,” Journal of Financial Economics, pp.339-349, 1977.

[7] Dimitros Asteriou and Stephen G Hall. “Arima models and the box-jenkins methodology,” Applied Econometrics, pp.265–286, 2011.
[8] Engle R. GARCH 101: “The use of ARCH/GARCH models in applied econometrics,” Journal of Economic Perspectives, pp.157-168, 2001.
[9] Jeffrey L Elman. “Distributed representations, simple recurrent networks, and grammatical structure,” Machine learning, pp.195–225, 1991.
[10] Sepp Hochreiter. “The vanishing gradient problem during learning recurrent neural nets, and problem solutions,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp.107-116, 1998.
[11] Khandelwal I, Satija U, Adhikari R. “Forecasting seasonal time series with functional link artificial neural network,” International Conference on Signal Processing and Integrated Networks. IEEE, pp.725-729,2015.
[12] Bahdanau D, Cho K, Bengio Y. “Neural machine translation by jointly learning to align and translate,” Computer Science, 2014.
[13] Fu J, Zheng H, Mei T. “Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition,” IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society pp.4476-4484, 2017.
[14] Vaswani A, Shazeer N, Parmar N. “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.
[15] Hochreiter S, Schmidhuber J. “Long short-term memory,” Neural Computation, pp.1735-1780, 1997.
[16] Heaton J B, Polson N G, Witte J H. “Deep learning in finance,” 2016.
[17] Rosas-Romero R, Etcheverry G. “Forecasting of stock return prices with sparse representation of financial time series over redundant dictionaries,” Expert Systems with Applications, pp.37-48, 2016.
[18] Chen W S, Du Y K. “Using neural networks and data mining techniques for the financial distress prediction model,” Expert Systems with Applications, pp.4075-4086, 2009.
[19] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M.Rush. “Structured attention networks,” International Conference on Learning Representations, 2017.