Multi-head Temporal Attention-Augmented Bilinear Network for Financial time series prediction

1st Mostafa Shabani
dept. Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
mshabani@ece.au.dk

2nd Dat Thanh Tran
Unit of Computing Sciences
Tampere University
Tampere, Finland
thanh.tran@tuni.fi

3rd Martin Magris
dept. Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
magris@ece.au.dk

4th Juho Kanniainen
Unit of Computing Sciences
Tampere University
Tampere, Finland
juho.kanniainen@tuni.fi

5th Alexandros Iosifidis
dept. Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
ai@ece.au.dk

***Abstract*—Financial time-series forecasting is one of the most challenging domains in the field of time-series analysis. This is mostly due to the highly non-stationary and noisy nature of financial time-series data. With progressive efforts of the community to design specialized neural networks incorporating prior domain knowledge, many financial analysis and forecasting problems have been successfully tackled. The temporal attention mechanism is a neural layer design that recently gained popularity due to its ability to focus on important temporal events. In this paper, we propose a neural layer based on the ideas of temporal attention and multi-head attention to extend the capability of the underlying neural network in focusing simultaneously on multiple temporal instances. The effectiveness of our approach is validated using large-scale limit-order book market data to forecast the direction of mid-price movements. Our experiments show that the use of multi-head temporal attention modules leads to enhanced prediction performances compared to baseline models.**

***Index Terms*—Deep learning, Attention mechanism, Limit Order Book, Financial Time-series**

I. Introduction

Time-series analysis has been significantly improved by recent machine learning and deep learning approaches. One of the most challenging domains in time-series analysis is financial time-series classification and prediction. The complex dynamics of financial markets reflect in highly non-stationary and noisy data. This characteristic and the large-scale high-dimensional nature of financial data strongly affect the analysis of financial time-series data. To tackle challenges in financial time-series analysis, many approaches have been proposed based on econometric, machine learning, and deep learning techniques.

In recent years, the accessibility to large-scale datasets and the improvements in computational capabilities have enabled deep Learning to excel in a variety of domains such as computer vision and natural language processing. Popular neural network designs for financial time-series include Recurrent Neural Networks (RNN) [1], of which the Long-Short Term Memory (LSTM) [2] and the Gated Recurrent Unit (GRU) [3] are the most widely used recurrent cells. Convolutional Neural Network (CNN) [4], which was originally designed for visual data, is nowadays also a popular choice for time-series data.

Recently, neural networks that are designed using multilinear operations have also shown competitive performance in time-series analysis tasks compared to recurrent or convolutional networks [5]. The Temporal Attention-Augmented Bilinear (`TABL`) is a neural network layer based on bilinear projection and attention mechanism that adaptively learns to mask out irrelevant time instances [5]. A new architectural design called Transformer [6], which heavily employs multiple attention modules, has emerged as a state-of-the-art model in language understanding tasks [7], as well as vision understanding tasks [8].

In this paper, inspired by the recent success of multi-head attention design, we propose an extension of the `TABL` network with multi-head attention design. The new design enables a bilinear mapping with the ability to simultaneously learn to focus on different temporal instances in the input time-series. As a result, more discriminative features can be extracted using our neural layer design, which leads to performance improvements compared to the original `TABL` networks. The remainder of this paper is organized as follows. In Section II, we provide a literature review on deep learning research for financial time-series forecasting. In Section III, we describe the proposed multi-head attention design for bilinear mapping. In Section IV, experimental protocols and empirical results are presented. Section V concludes our paper.

II. Related works

The complex dynamics of financial data and the existence of large-scale datasets have fostered the use of deep learning models in financial applications. Among those, analysis tasks derived from high-frequency Limit-Order Book (LOB) data have attracted great attention from the community due to its

unique capability in tracking market dynamics. A comprehensive description of LOBs can be found in [9].

Since our work focuses on analyzing LOB data, here we review related works in LOB research. There have been several works using LOB data. For example, the spatial distribution in LOB has been studied in [10] via deep neural networks. The LOB data is generally highly non-stationary and requires great attention in terms of pre-processing. Adaptive data normalization schemes have been proposed recently to tackle such challenges [11], [12]. Designing suitable neural network architectures for time-series derived from LOB has also been the focus of several works, including both manually [13], [14] and automatically generated network architectures [15]. Besides recurrent networks and TABL networks, neural networks constructed from Bag-of-Feature layers [16] have also demonstrated a great fit for variable-length sequences.

Among the human expert designs, the attention module has shown a consistent ability to enhance the baseline models. The main idea of an attention unit is to learn to focus on relevant parts of the input while masking out unimportant parts of it. Attention computation in neural networks was first introduced for machine translation tasks by the work of [17]. Incorporation of attention mechanism is also popular among time-series analysis community [5], [14], [18]–[20]. Our work relies on a computationally fast and efficient design called Temporal Attention-augmented Bilinear Layer (TABL) network [5], which has been shown to achieve excellent performance in both computational cost and modeling capacity. To have a better understanding of our proposed multi-head attention design in Section III, the working mechanism of a TABL is described next.

In TABL, the bilinear projection incorporating a temporal attention mechanism produces an output matrix $\mathbf{Y} \in \mathbb{R}^{D' \times T'}$ given an input matrix $\mathbf{X} \in \mathbb{R}^{D \times T}$. $\mathbf{X}$ is a multivariate time-series in which each column represents the D features at a certain time instance, for a series of length T. A TABL layer performs five computational steps to transform the input $\mathbf{X}$ to the output $\mathbf{Y}$ as follows:

$$\bar{\mathbf{X}} = \mathbf{W}_1 \mathbf{X}. \tag{1}$$

$$\mathbf{E} = \bar{\mathbf{X}} \mathbf{W}, \tag{2}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}, \tag{3}$$

$$\tilde{\mathbf{X}} = \lambda(\bar{\mathbf{X}} \odot \mathbf{A}) + (1 - \lambda)\bar{\mathbf{X}}, \tag{4}$$

$$\mathbf{Y} = \phi\left(\tilde{\mathbf{X}} \mathbf{W}_2 + \mathbf{B}\right). \tag{5}$$

III. Temporal Multi-head Attention Bilinear Layer

Our proposed neural layer is constructed based on the structure of TABL. The main idea of our design is to augment the bilinear mapping with multiple attention computation units (otherwise called attention heads), which are calculated independently (in parallel). By using multiple attention heads, we hypothesize that for certain input series, the salient features can appear in pairs, triplets, or larger subsets, which cannot be captured by a single attention head. Thus, by extending the number of attention heads, we might be able to detect more relevant features that lie within the input data. To reach this goal, the intermediate output in the TABL layer after going through the linear transformation in the first dimension (output of Eq. (1)) is used as the input of multiple soft attention heads, generating multiple attended features $\tilde{\mathbf{X}}$ as in Eq. (4) for each attention head. Therefore, if we consider K attention heads, each of which is associated with a weight matrix $\{\mathbf{W}^{(\mathrm{k})}\}, k = \{1, ..., K\}$. The output of all attention heads must be combined based on a strategy, which can be, e.g., summation or concatenation. In this paper, we investigate concatenation for combining the outputs of attention mechanisms as the outputs of each of the attention mechanisms are used without any processing and losing information.

The computational steps of our Multi-head Temporal Attention Bilinear Layer (MTABL) with K attention heads are as follows:

- The first step in MTABL is similar to TABL, which projects each temporal slice (column) of the input matrix to a D'-dimensional feature space:
$$\bar{\mathbf{X}} = \mathbf{W}_1 \mathbf{X}. \tag{6}$$
- In the second step, the resulting feature matrix is passed through K parallel attention heads, each of which learns to focus on an important temporal instance:
$$\mathbf{E}_1 = \bar{\mathbf{X}} \mathbf{W}^{(1)}, \mathbf{E}_2 = \bar{\mathbf{X}} \mathbf{W}^{(2)}, ..., \mathbf{E}_K = \bar{\mathbf{X}} \mathbf{W}^{(\mathrm{k})} \tag{7}$$
where all $\mathbf{W}^{(\mathrm{k})} \in \mathbb{R}^{T \times T}$ is the weight matrix to compute attention in the k-th head.
- The un-normalized attention matrices are then normalized by the softmax function in a row-wise manner, similar to Eq. (3), generating the attention masks $\{\mathbf{A}_{(\mathrm{k})}\}, k = \{1, ..., K\}$
- The final attended features $\tilde{\mathbf{X}}$ for each attention head are computed by combining the original and masked-out features using the attention mask $\{\mathbf{A}_{(\mathrm{k})}\}$ and λ:
$$\begin{aligned} \tilde{\mathbf{X}}_{(1)} &= \lambda(\bar{\mathbf{X}} \odot \mathbf{A}_{(1)}) + (1 - \lambda)\bar{\mathbf{X}}. \\ \tilde{\mathbf{X}}_{(2)} &= \lambda(\bar{\mathbf{X}} \odot \mathbf{A}_{(2)}) + (1 - \lambda)\bar{\mathbf{X}}. \\ &\vdots \\ \tilde{\mathbf{X}}_{(\mathrm{K})} &= \lambda(\bar{\mathbf{X}} \odot \mathbf{A}_{(\mathrm{K})}) + (1 - \lambda)\bar{\mathbf{X}}. \end{aligned} \tag{8}$$
λ, which is constrained to have a value between $[0, 1]$, represents the fraction of original information that is relevant and should be allowed to flow through the network along with the attended features. For this reason, it is more intuitive to use a single value of λ for all attention heads.
- All K attended features $\tilde{\mathbf{X}}_{(\mathrm{k})}$ are combined together as a single matrix. To end this, the concatenation is used to combine $\{\tilde{\mathbf{X}}_{(\mathrm{k})}\}$. For the concatenation scheme, even though all attended features have the exact same size, it is counterintuitive to concatenate $\{\tilde{\mathbf{X}}_{(\mathrm{k})}\}$ on the second

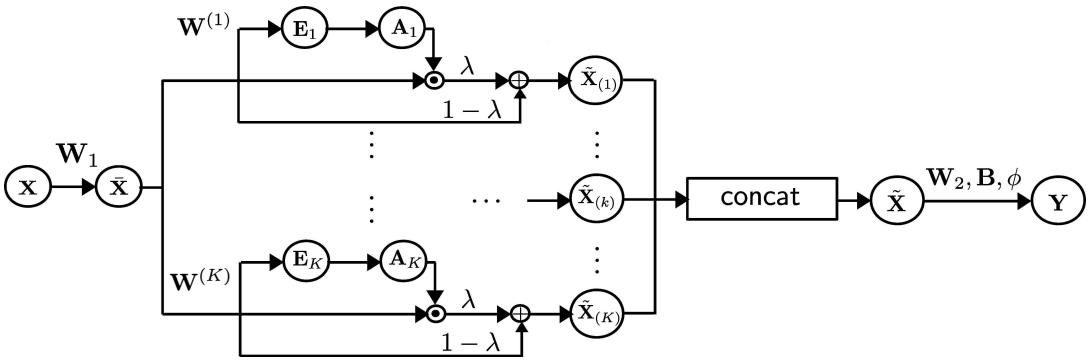

Fig. 1. Schematic illustration of the proposed MTABL layer

dimension i.e., the temporal dimension, since this means that multiple features of a sequence are concatenated to form a much longer sequence, therefore breaking its temporal coherence. Thus, our formulation of the method concatenates $\{\tilde{\mathbf{X}}_{(\mathrm{k})}\}$ on the feature dimension, then combines all the features of a given temporal instance by linearly projecting them back to D'-dimensional space:

$$\tilde{\mathbf{X}} = \tilde{\mathbf{W}}_1 \begin{bmatrix} \tilde{\mathbf{X}}_{(1)} \\ \vdots \\ \tilde{\mathbf{X}}_{(\mathrm{K})} \end{bmatrix} \in \mathbb{R}^{D' \times T} \tag{9}$$

where $\tilde{\mathbf{W}}_1 \in \mathbb{R}^{D' \times (D' \cdot K)}$ is a weight matrix that is learned to combine the contenated features.

- In the final step, similar to TABL, MTABL computes the output sequence $\mathbf{Y}$.

Fig. 1 illustrates the structure of MTABL.

The complexity of TABL is $O(D'DT + D'TT' + 2D'T' + D'T^2 + 3D'T)$ [5]. Due to the additional attention heads and the combination of their respective outputs, the MTABL is of greater computational complexity. In particular, for a MTABL with K attention heads, the number of additional multiplications involved in Eq. (7) w.r.t Eq. (2) is $K-1$. Furthermore, an additional complexity term of order $O(D'(D' \cdot K)T)$ is implied by the multiplications in (Eq. (9)). The overall complexity of our proposed method is thus $O(D'DT + D'TT' + 2D'T' + KD'T^2 + 3D'T + D'(D' \cdot K)T)$.

IV. EXPERIMENTS

The performance of our model is evaluated on the mid-price movement prediction task using the publicly available FI-2010 dataset [21]. We used the first 40 dimensions of the feature vectors, which correspond to the top ten bid and ask prices and volumes of the LOB. For each feature vector, the authors in [21] derived the labels for future movements of the mid-price in the next $H = \{10, 20, 30, 50, 100\}$ order events, which are referred to as prediction horizons.

To evaluate the performance of MTABL in comparison to TABL, we used the same experimental protocol of TABL used in [5]. We trained all the networks to predict the future movements of mid-price in the next 10 order events, i.e., the target label corresponding to $H = 10$. Three network topologies proposed in [5] were used in our experiments as the baseline models. The topology A consists of one TABL layer, the topology B consists of one BL layer and one TABL layer, and the topology C consists of two BL layer and one TABL layer. In these architectures, the last layer is a TABL layer and all other layers are BL layers. We evaluated MTABL networks with a varying number of attention heads, from 2 to 5.

Table I reports the corresponding experiment results for network topologies with concatenation as the attention aggregation strategy to combine attention mechanisms' outputs. Due to the stochastic nature of the optimizer, we report the mean and standard deviation between four independent runs. The following metrics were used to measure the performance of each model: accuracy, precision, recall, and F1-Score. Since the FI-2010 dataset has a skewed distribution of labels with the majority of samples having the stationary label, the average F1 score, which reflects the trade-off between precision and recall, is used as the main metric to compare between models. The column "Layer" indicates which type of output layer was used in the network architecture. The number of attention heads used in each MTABL layer is indicated by the last number in the notation, that is (MTABL-3) denotes a MTABL layer using three attention heads. The whole row corresponds to the model with the best performance for each network topology based on F1-Score is highlighted in boldface. The results show improved performances of the multi-head attention configuration in for all network topologies. This suggests that using multiple attentions can help the output layer to detect and focus on crucial elements of data more accurately and improve the prediction performance.

The interpretation of the results from Table I that correspond to network topologies that use concatenation to combine the outputs of all attention heads in each layer is straightforward. MTABL show the major improvements over the original TABL with five attention heads. In the first instance, this can be interpreted as a considerable ($K = 5$) amount of relevant attention that is neglected in TABL that MTABL's increased number of attention layers captures. When additional BL layers are considered in topologies B and C the best performances

TABLE I
PERFORMANCES OF MULTI-HEAD ATTENTION MODELS USING CONCATENATION (MTABL−C) (MEAN ± STD.)

Topology	Layer	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
A	TABL	67.21±0.045	53.76±0.039	55.47±0.015	54.25±0.03
	MTABL−C-2	69.78±0.029	56.64±0.029	59.58±0.026	57.81±0.029
	MTABL−C-3	72.45±0.009	59.03±0.009	60.41±0.001	59.66±0.005
	MTABL−C-4	72.30±0.007	59.25±0.007	61.60±0.005	60.28±0.006
	MTABL−C-5	**72.57±0.003**	**59.63±0.003**	**62.68±0.005**	**60.90±0.004**
B	TABL	78.56±0.002	67.55±0.003	71.07±0.004	69.10±0.002
	MTABL−C-2	77.68±0.004	66.44±0.004	70.56±0.007	68.18±0.004
	MTABL−C-3	78.13±0.007	67.04±0.009	71.39±0.004	68.89±0.007
	MTABL−C-4	77.63±0.005	66.48±0.006	70.89±0.003	68.35±0.005
	MTABL−C-5	**78.22±0.012**	**67.4±0.017**	**71.52±0.004**	**69.16±0.012**
C	TABL	83.52±0.009	75.12±0.013	77.02±0.006	76.01±0.009
	MTABL−C-2	83.69±0.005	75.21±0.008	77.74±0.004	76.39±0.006
	MTABL−C-3	81.64±0.014	72.16±0.022	75.17±0.015	73.54±0.019
	MTABL−C-4	**83.71±0.01**	**75.37±0.015**	**77.63±0.006**	**76.42±0.011**
	MTABL−C-5	82.63±0.004	73.66±0.005	76.93±0.008	75.16±0.006

are achieved under $K = 4$ and $K = 5$ respectively but the improvement is minor. This can indicate that the five attention heads in topology A the attention-relevant information are highly useful to improve the prediction performance. On the other hand, in topologies B and C when additional BL layers are introduced, the best MTABL performance is still observed with a higher number of layers (4 and 5) but the improvement is not as much as topology A. This shows that there is little temporal attention discarded in TABL that MTABL captures.

V. CONCLUSION

In this paper, a new neural layer based on the structure of TABL and the idea of multi-head attention is proposed for financial time-series analysis. We proposed a formulation of the TABL layer that utilizes multiple attention units to focus on different temporal importances. Our MTABL design stands out as a suitable neural layer for addressing numerous forecasting problems over a wide class of time-series characterized by complex and time-varying dynamics.

Extensive experiments in forecasting the direction of mid-price movements using limit-order book data show that the proposed MTABL design is indeed capable of unveiling additional layers of relevant predictive significance lodged in the data. The improved MTABL performance is generally achieved for different combination schemes of the attention heads' outputs, for a different number of attention heads, and under different network topologies.

ACKNOWLEDGMENT

The research received funding from the Independent Research Fund Denmark project DISPA (Project Number: 9041-00004).

REFERENCES

[1] Mandic, D. & Chambers, J. Recurrent neural networks for prediction: learning algorithms, architectures and stability. (Wiley,2001)
[2] Fischer, T. & Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal Of Operational Research*. **270**, 654-669 (2018)
[3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv Preprint ArXiv:1406.1078*. (2014)
[4] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of The IEEE*. **86**, 2278-2324 (1998)
[5] Tran, D., Iosifidis, A., Kanniainen, J. & Gabbouj, M. Temporal Attention-Augmented Bilinear Network for Financial Time-Series Data Analysis. *IEEE Transactions On Neural Networks And Learning Systems*. **30** pp. 1407-1418 (2017)
[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention is all you need. *ArXiv Preprint ArXiv:1706.03762*. (2017)
[7] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)
[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. & Others An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*. (2020)
[9] Bouchaud, J., Mézard, M. & Potters, M. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*. **2**, 251-256 (2002)
[10] Sirignano, J. Deep learning for limit order books. *Quantitative Finance*. **19**, 549-570 (2019)
[11] Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Deep Adaptive Input Normalization for Time Series Forecasting. *IEEE Transactions On Neural Networks And Learning Systems*. **31**, 3760-3765 (2020)
[12] Tran, D., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Data Normalization for Bilinear Structures in High-Frequency Financial Time-series. *International Conference On Pattern Recognition (ICPR)*. (2020)
[13] Zhang, Z., Zohren, S. & Roberts, S. DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. *IEEE Transactions On Signal Processing*. **67** pp. 3001-3012 (2019)
[14] Tran, D., Passalis, N., Tefas, A., Gabbouj, M. & Iosifidis, A. Attention-based Neural Bag-of-Features Learning for Sequence Data. *ArXiv Preprint ArXiv:2005.12250*. (2020)

[15] Tran, D., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Data-driven Neural Architecture Learning For Financial Time-series Forecasting. *ArXiv*. abs/1903.06751. (2019)

[16] Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Temporal Bag-of-Features Learning for Predicting Mid Price Movements Using High Frequency Limit Order Book Data. *IEEE Transactions On Emerging Topics In Computational Intelligence*. pp. 1-12 (2018)

[17] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*. (2014)

[18] Mäkinen, Y., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Forecasting jump arrivals in stock prices: new attention-based network architecture using limit order book data. *Quantitative Finance*. **19**, 2033-2050 (2019)

[19] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G. & Cottrell, G. A dual-stage attention-based recurrent neural network for time series prediction. *ArXiv Preprint ArXiv:1704.02971*. (2017)

[20] Shabani, M. & Iosifidis, A. Low-Rank Temporal Attention-Augmented Bilinear Network for financial time-series forecasting. *2020 IEEE Symposium Series On Computational Intelligence (SSCI)*. pp. 2156-2161 (2020)

[21] Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. & Iosifidis, A. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal Of Forecasting*. **37** pp. 852-866 (2018)