

Transformers for limit order books

James Wallbridge*

March 3, 2020

Abstract

We introduce a new deep learning architecture for predicting price movements from limit order books. This architecture uses a causal convolutional network for feature extraction in combination with masked self-attention to update features based on relevant contextual information. This architecture is shown to significantly outperform existing architectures such as those using convolutional networks (CNN) and Long-Short Term Memory (LSTM) establishing a new state-of-the-art benchmark for the FI-2010 dataset.

Contents

1	Introduction	1
2	Experiments	4
3	Architecture	5
4	Results	8
5	Discussion	10
A	Training curves	15
B	Attention distributions	15

1 Introduction

Understanding high-frequency market micro-structure in time-series data such as limit order books (LOB) is complicated by a large number of factors including high-dimensionality,

*Correspondence to james.wallbridge@gmail.com

trends based on supply and demand, order creation and deletion around price jumps and the overwhelming relative percentage of order cancellations. It makes sense in this inherently noisy environment to take an agnostic approach to the underlying mechanisms inducing this behavior and construct a network which learns to uncover the relevant features from raw data. This removes the bias contained in models using hand-crafted features and other market assumptions such as those in autoregressive models VAR [42] and ARIMA [1].

Arguably the most successful architecture used to extract features is the convolutional neural network [20] which makes use of translation equivariance, present in many domains including time-series applications. For time-series however, further inductive biases prove to be beneficial. Convolutional neural networks with a causal temporal bias were introduced in [24] to encode long-range temporal dependencies in raw audio signals. Here convolutions are replaced by dilated causal convolutions controlled by a dilation rate. The dilation rate is the number of input values skipped by the filter, thereby allowing the network to act with a larger receptive field. In this work, features from our architecture will come from the output of multiple such dilated causal convolutional layers connected in series.

Once we have a collection of features, we would like to do computations with these learned representations to enable context dependent updates. Historically, attention networks were introduced in [3] to improve existing long-short term memory (LSTM) [16, 15] models for neural machine translation by implementing a “soft search” over neighboring words enabling the system to focus only on words relevant to the generation of the next target word. This early work combined attention with RNNs. Shortly after, CNNs were combined with attention in [39] and [6] for image captioning and question-answering tasks respectively.

In [37], self-attention was introduced as a stand alone replacement for LSTMs on a wide range of natural language processing tasks leading to state-of-the-art results [10, 26] which included masked word prediction. Introducing self-attention can be thought of as incorporating an inductive biases into the learning architecture to exploit relational structure in the task environment. This amounts to learning over a graph neural network [28, 5] where nodes are entities given by the learned features which are then updated through a message passing procedure along edges. Results in various applications show that self-attention can better capture long range dependencies in comparison to LSTMs [9].

More precisely, [37] introduced the transformer architecture which consists of an encoder and decoder for language translation. Both the encoder and decoder contain the repetition of modules which we refer to as *transformer blocks*. Each transformer block consists of a multi-head self-attention layer followed by normalization, feedforward and residual connections. This is described in detail in Section 3.

Combining transformer blocks with convolutional layers for feature extraction is a powerful combination for various tasks. In particular, for complex reasoning tasks in various strategic game environments, the addition of these transformer modules significantly

enhanced performance and sample efficiency compared with existing non-relational baselines [40, 38, 8]. In this work we combine the causal convolutional architecture of [24] with multiple transformer blocks. Moreover, our transformer blocks contain masked multi-head self-attention layers. By applying a mask to our self-attention functions, we ensure that the ordering of events in our time-series is never violated at each step, ie. entities can only attend to entities in its causal past.

We train and test our model on the publicly available FI-2010 data-set¹ which is a LOB of five instruments from the Nasdaq Nordic stock market for a ten day period [23]. We show that our algorithm outperforms other common and previously state-of-the-art architectures using standard model validation techniques.

In summary, inspired by the wavenet architecture of [24] where dilated causal convolutions were used to encode long-range temporal dependencies, we use these causal convolutions to build a feature map for our transformer blocks to act on. We refer to our specific architecture as TransLOB. It is a composition of differentiable functions that process and integrate both local and global information from the LOB in a dynamic relational way whilst respecting the causal structure.

There are a number of advantages to our architecture outside of the significant increases in performance. Firstly, in spite of the $O(N^2)$ complexity of the self-attention component, our architecture is substantially more sample efficient than existing LSTM architectures for this task. Secondly, the ability to analyse attention distributions provides a clearer picture of internal computations within the model compared with these other methods leading to better interpretability.

Related work

There is now a substantial literature applying deep neural networks to time-series applications, and in particular, limit order books (LOB). Convolutional neural networks (CNN) have been explored in LOB applications in [12, 34]. To capture long-range dependencies in temporal behavior, CNNs have been combined with recurrent neural networks (RNN) (typically long-short term memory (LSTM)) which improve on earlier results [36, 41]. Some modifications to the standard convolutional layer have been used in attempts to infer local interactions over different time horizons. For example, [41] uses an inception module [32] after the standard convolutional layers for this inference followed by an LSTM to encode relational dynamics. Stand-alone RNNs have been used extensively in market prediction [11, 13, 4] and have been shown to outperform models based on standard multi-layer perceptrons, random forests and SVMs [35].

For time-series applications, recent work [33, 25] uses attention and [18, 22, 29] in combination with CNNs. However, there are relatively few references which combine CNNs with transformers to analyse time-series data. We mention [30] which uses a CNN plus multi-head self-attention to analyse clinical time-series behaviour and [21] which

¹The “MNIST” for limit order books.

became aware to us during the final write-up of this paper which uses a similar architecture to our own and applied to univariate synthetic and energy sector datasets. As far as we are aware, ours is the first work applying this class of architectures to the multivariate financial domain, with the various subtleties arising in this particular application.

2 Experiments

A limit order book (LOB) at time t is the set of all active orders in a market at time t . These orders consist of two sides; the bid-side and the ask-side. The bid-side consists of buy orders and the ask-side consists of sell orders both containing price and volume for each order. Our experiments will use the LOB from the publicly available FI-2010 dataset². A general introduction to LOBs can be found in [14].

Let $\{p_a^i(t), v_a^i(t)\}$ denote the price (resp. volume) of sell orders at time t at level i in the LOB. Likewise, let $\{p_b^i(t), v_b^i(t)\}$ denote the price (resp. volume) of buy orders at time t at level i in the LOB. The bid price $p_b^1(t)$ at time t is the highest stated price among active buy orders at time t . The ask price $p_a^1(t)$ at time t is the lowest stated price among active sell orders at time t . A buy order is executed if $p_b^1(t) > p_a^1(t)$ for the entire volume of the order. Similarly, a sell order is executed if $p_a^1(t) < p_b^1(t)$ for the entire volume of the order.

The FI-2010 dataset is made up of 10 days of 5 stocks from the Helsinki Stock Exchange, operated by Nasdaq Nordic, consisting of 10 orders on each side of the LOB. Event types can be executions, order submissions, and order cancellations and are non-uniform in time. We restrict to normal trading hours (no auction). The general structure of the LOB is contained in Table 1.

$(p_a^{10}(t), v_a^{10}(t))$	$(p_a^{10}(t+1), v_a^{10}(t+1))$	$(p_a^{10}(t+10), v_a^{10}(t+10))$
⋮	⋮	⋮
$(p_a^1(t), v_a^1(t))$	$(p_a^1(t+1), v_a^1(t+1))$	$(p_a^1(t+10), v_a^1(t+10))$
$(p_b^1(t), v_b^1(t))$	$(p_b^1(t+1), v_b^1(t+1))$	$(p_b^1(t+10), v_b^1(t+10))$
⋮	⋮	⋮
$(p_b^{10}(t), v_b^{10}(t))$	$(p_b^{10}(t+1), v_b^{10}(t+1))$	$(p_b^{10}(t+10), v_b^{10}(t+10))$
Event t	Event $t + 1$	Event $t + 10$

Table 1: Structure of the limit order book.

The data is split into 7 days of training data and 3 days of test data. Preprocessing consists of normalizing the data x according to the z -score

$$\bar{x}_t = \frac{x_t - \bar{y}}{\sigma_{\bar{y}}}$$

²The dataset is available at <https://etsin.fairdata.fi/dataset/73eb48d7-4dbc-4a10-a52a-da745b47a649>

where \bar{y} (resp. $\sigma_{\bar{y}}$) is the mean (resp. standard deviation) of the previous days data. Since the aim of this work is to extract the most amount of possible latent information contained in the LOB, we do not include any of the hand-crafted features contained in the FI-2010 dataset. For a detailed description of this dataset we refer the reader to [23].

We aim to predict future movements from the (virtual) mid-price. Price direction of the data is calculated using the following smoothed version of the mid-price. This amounts to adjusting for the average volatility of each instrument. The virtual mid-price is the mean

$$p(t) = \frac{p_a^1(t) + p_b^1(t)}{2}$$

between the bid-price and the ask-price. The mean of the next k mid-prices is then

$$m_k^+(t) = \frac{1}{k} \sum_{n=0}^k p(t+n).$$

The direction of price movement for the FI-2010 dataset is calculated using the percentage change of the virtual mid-price according to

$$r_k(t) = \frac{m_k^+(t) - p(t)}{p(t)}.$$

There exist other more sophisticated methods to determine the direction of price movement at a given time. However, for fair comparison to other work, we utilize this definition and leave other methods for future work. The direction is up (+1) if $r_k(t) > \alpha$, down (-1) if $r_k(t) < -\alpha$ and neutral (0) otherwise, according to a chosen threshold α . For the FI-2010 dataset, this has been set to $\alpha = 0.002$.

We consider the following four test cases $k \in \{10, 20, 50, 100\}$ for the denoising horizon window. The 100 most recent events are used as input to our model.

3 Architecture

In this section we give a detailed account of our architecture. The main two components are a convolutional module and a transformer module. They contain multiple iterations of dilated causal convolutional layers and transformer blocks respectively. A transformer block consists of a specific combination of multi-head self-attention, residual connections, layer normalization and feedforward layers. We took seriously the causal nature of the problem by implementing both causality in the convolutional module and causality in the transformer module through masked self-attention to accurately capture temporal information in the LOB. Our resulting architecture will be referred to as TransLOB.

Since each order consists of a price and volume, a state $x_t = \{p_a^i(t), v_a^i(t), p_b^i(t), v_b^i(t)\}_{i=1}^{10}$ at time t is a vector $x_t \in \mathbb{R}^{40}$. Events are irregularly spaced in time and the 100 most recent events are used as input resulting in a normalized vector $X \in \mathbb{R}^{100 \times 40}$.

We apply five one-dimensional convolutional layers to the input X , regarded as a tensor of shape $[100, 40]$ (ie. an element of $\mathbb{R}^{100} \otimes \mathbb{R}^{40}$). All layers are dilated causal convolutional layers with 14 features, kernel size 2 and dilation rates 1, 2, 4, 8 and 16 respectively. This means the filter is applied over a window larger than its length by skipping input values with a step given by the dilation rate with each layer respecting the causal order. The first layer with dilation rate 1 corresponds to standard convolution. All activation functions are ReLU.

The full size of the channel filter is used to allow the weights in the filter to infer the relative importance of each level on each side of the mid-price. It is expected that higher weights will be allocated to shallower levels in the LOB since those levels are most indicative of future activity. The output of the convolutional module is a tensor of shape $[100, 14]$.

This output then goes through layer normalization [2] to stabilize dynamics before each feature vector is concatenated with a one-dimensional temporal encoding resulting in a tensor X of shape $[100, 15]$. We will refer to $N = 100$ as the number of *entities* and $d = 15$ as the model dimension. We denote these entities by e_i , $1 \leq i \leq N$, where $e_i \in E = \mathbb{R}^d$. These entities are then updated through learning in a number of steps.

First we introduce an inner product space $H = \mathbb{R}^d$ with dot product pairing $\langle h, h' \rangle = h \cdot h'$. We employ a multi-head version of self-attention with C channels. Therefore, we choose a decomposition $H = H_1 \oplus \dots \oplus H_C$ and apply a linear transformation

$$T = \bigoplus_{a=1}^C T_a : E \rightarrow \bigoplus_{a=1}^C H_a^{\oplus 3}$$

with H_a each of dimension d/C . The vectors $(q_{i,(a)}, k_{i,(a)}, v_{i,(a)}) = T_a(e_i)$ are referred to as *query*, *key* and *value* vectors respectively. We arrange these vectors into matrices Q_a , K_a and V_a respectively with N -rows and d -columns. In other words, $Q_a = XW_a^Q$, $K_a = XW_a^K$ and $V_a = XW_a^V$ for weight matrices W_a^Q , W_a^K and W_a^V which are vectors in $\mathbb{R}^{d \times d/C}$.

Next we apply the masked scaled dot-product self-attention function

$$\text{head}_a = V'_a = \text{Softmax} \left(\text{Mask} \left(\frac{Q_a K_a^T}{\sqrt{d}} \right) \right) V_a$$

resulting in a matrix of refined value vectors for each entity. Here Mask substitutes infinitesimal values to entries in the upper right triangle of the applied matrix which forces queries to only pay attention to keys in its causal history via the softmax function. The heads are then concatenated and a final learnt linear transformation is given leading to the multi-head self-attention operation

$$\text{MultiHead}(X) = \left(\bigoplus_{a=1}^C \text{head}_a \right) W^O$$

where $W^O \in \mathbb{R}^{d \times d}$.

We next add a residual connection and apply layer normalization resulting in

$$Z = \text{LayerNorm}(\text{MultiHead}(X) + X).$$

This is followed by a feedforward network MLP consisting of a ReLU activation between two affine transformations applied identically to each position, ie. individually to each row of Z . The inner layer is of dimension $4 \times d = 60$. Finally, a further residual connection and final layer normalization is applied to arrive at our updated matrix of entities

$$\text{TransformerBlock}(X) = \text{LayerNorm}(\text{MLP}(Z) + Z).$$

The output of the transformer block is the same shape $[N, d]$ as the input. Our updated entities are $e'_i \in \mathbb{R}^{15}$, $1 \leq i \leq N$.

After multiple iterations of the transformer block, the output is then flattened and passed through a feedforward layer of dimension 64 with ReLU activation and L2 regularization. Finally, we apply dropout followed by a softmax layer to obtain the final output probabilities. A schematic of the TransLOB architecture is given in Figure 1.

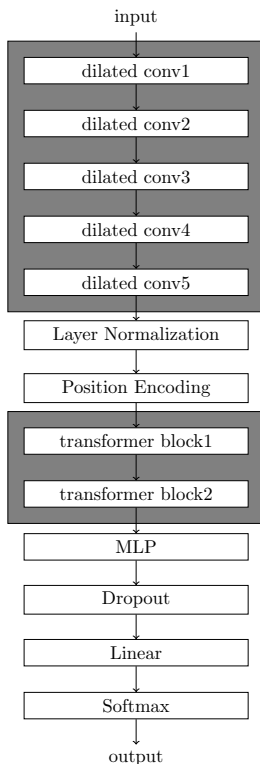


Figure 1: Architecture schematic with enclosed convolutional and transformer modules.

For the FI-2010 dataset, we employ two transformer blocks with three heads and with the weights shared between iterations of the transformer block. The hyperparameters are contained in Table 2. No dropout was used inside the transformer blocks.

Hyperparameter	Value
Batch size	32
Adam β_1	0.9
Adam β_2	0.999
Learning rate	1×10^{-4}
Number of heads	3
Number of blocks	2
MLP activations	ReLU
Dropout rate	0.1

Table 2: Hyperparameters for the FI-2010 experiments.

4 Results

Here we record our experimental results for the FI-2010 dataset. The first 7 days were used to train the model and the last 3 days were used as test data. Training was done with mini-batches of size 32. Our metrics include accuracy, precision, recall and F1. All training was done using one K80 GPU on google colab.

To be consistent with earlier works using the same dataset, we train and test our model on the horizons $k = \{10, 20, 50, 100\}$. All models were trained for 150 epochs, although convergence was achieved significantly earlier. See Figure 3 of Appendix A for an example.

The following models were used as comparison. An LSTM was utilized and compared to a support vector machine (SVM) and multi-layer perceptron (MLP) in [35] with favourable results. Results using a stand-alone CNN were reported in [34]. This model was reproduced and trained for use as our baseline for the horizon $k = 100$. The baseline training and test curves are shown in Figure 4 of Appendix A. In [36] a CNN was combined with an LSTM resulting in the architecture denoted CNN-LSTM. An improvement over the CNN-LSTM architecture, named DeepLOB, was achieved in [41] by using an inception module between the CNN and LSTM together with a different choice of convolution filters, stride and pooling. Finally, the architecture C(TABL) refers to the best performing implementation of the temporal attention augmented bilinear network of [33].

Our results are shown in Table 3, Table 4, Table 5 and Table 6 for each of the horizon choices respectively. The training and test curves with respect to accuracy for $k = 100$ are shown in Figure 3 of Appendix A.

For inspection of our model, we plot the attention distributions for all three heads in the first transformer block. A random sample input was chosen from the horizon $k = 10$ test set. Pixel intensity has been scaled for ease of visualization. The vertical axes represent the query index $0 \leq i \leq 100$ and the horizontal axes represent the key index $0 \leq j \leq 100$. Queries are aware of the distance to keys through the position embedding layer and entities are only updated with memory from the past owing to the attention mask. As can be seen in Figure 2, and Figure 5 and Figure 6 of Appendix B, the different

Model	Accuracy	Precision	Recall	F1
SVM [35]	-	39.62	44.92	35.88
MLP [35]	-	47.81	60.78	48.27
CNN [34]	-	50.98	65.54	55.21
LSTM [35]	-	60.77	75.92	66.33
CNN-LSTM [36]	-	56.00	45.00	44.00
C(TABL) [33]	84.70	76.95	78.44	77.63
DeepLOB [41]	84.47	84.00	84.47	83.40
TransLOB	87.66	91.81	87.66	88.66

Table 3: Prediction horizon $k = 10$.

Model	Accuracy	Precision	Recall	F1
SVM [35]	-	45.08	47.77	43.20
MLP [35]	-	51.33	65.20	51.12
CNN [34]	-	54.79	67.38	59.17
LSTM [35]	-	59.60	70.52	62.37
CNN-LSTM [36]	-	-	-	-
C(TABL) [33]	73.74	67.18	66.94	66.93
DeepLOB [41]	74.85	74.06	74.85	72.82
TransLOB	78.78	86.17	78.78	80.65

Table 4: Prediction horizon $k = 20$.

Model	Accuracy	Precision	Recall	F1
SVM [35]	-	46.05	60.30	49.42
MLP [35]	-	55.21	67.14	55.95
CNN [34]	-	55.58	67.12	59.44
LSTM [35]	-	60.03	68.58	61.43
CNN-LSTM [36]	-	56.00	47.00	47.00
C(TABL) [33]	79.87	79.05	77.04	78.44
DeepLOB [41]	80.51	80.38	80.51	80.35
TransLOB	88.12	88.65	88.12	88.20

Table 5: Prediction horizon $k = 50$.

Model	Accuracy	Precision	Recall	F1
CNN [34]	63.06	63.29	63.06	62.97
TransLOB	91.62	91.63	91.62	91.61

Table 6: Prediction horizon $k = 100$.

heads learn to attend to different properties of the temporal dynamics. A majority of the queries pay special attention to the most recent keys which is sensible for predicting the next price movement. This is particularly clear in heads two and three.

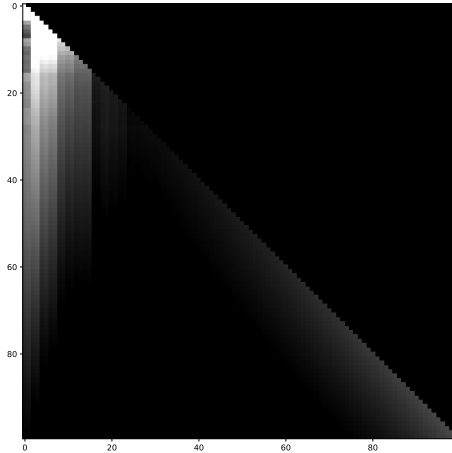


Figure 2: First head of the first transformer block.

5 Discussion

We have shown that the limit order book contains informative information to enable price movement prediction using deep neural networks with a causal and relational inductive bias. This was shown by introducing the architecture TransLOB which contains both a dilated causal convolutional module and a masked transformer module. This architecture was tested on the publicly available FI-2010 dataset achieving state-of-the-art results. We expect further improvements using more sophisticated proprietary additions such as the inclusion of sentiment information from news, social media and other sources. However, this work was developed to exploit only the information contained in the LOB and serves as very strong baseline from which additional tools can be added.

Due to the limited nature of the FI-2010 dataset, significant time was spent tuning hyperparameters of our model to negate overfitting. In particular, our architecture was notably sensitive to the initialization. However, due to the very strong performance of the model, together with the flexibility and sensible inductive biases of the architecture, we expect robust results on larger LOB datasets. This is an important second step and will be addressed in future work. In particular, this will allow us to explore the generalization capabilities of the model together with the optimization of important parameters such as the horizon k and threshold α . Nevertheless, based on these initial results we argue that further investigation of transformer based models for financial time-series prediction tasks is warranted.

The efficiency of our algorithm is another important property which makes it amenable to training on larger datasets and LOB data with larger event windows. In spite of the $O(N^2)$ complexity of the self-attention component, our architecture is significantly more sample efficient than existing LSTM architectures for this task such as [35, 36, 41].

However, moving far beyond the window size of 100, to the territory of LOB datasets on the scale of months or years, it would be interesting to explore sparse and compressed representations in the transformer blocks. Implementations of sparsity and compression can be found in [7, 31, 19, 21] and [17, 27] respectively.

Looking forward, similar to recent advances in natural language processing, the next generation of financial time-series models should implement self-supervision as pretraining [10, 26]. Finally, it would be interesting to consider the influence of higher-order self-attention [8] in LOB and other financial time-series applications.

Acknowledgements

The author would like to thank Andrew Royal and Zihao Zhang for correspondence related to this project.

References

- [1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112. IEEE, 2014.
- [2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [4] W. Bao, J. Yue, and Y. Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7), 2017.
- [5] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- [6] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.
- [7] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.
- [8] J. Clift, D. Doryn, D. Murfet, and J. Wallbridge. Logic and the 2-simplicial transformer. In *Proceedings of the International Conference on Learning Representations*, 2020.

- [9] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11] M. Dixon. Sequence classification of the limit order book using recurrent neural networks. *Journal of computational science*, 24:277–286, 2018.
- [12] J. Doering, M. Fairbank, and S. Markose. Convolutional neural networks applied to high-frequency market microstructure forecasting. In *2017 9th Computer Science and Electronic Engineering (CEECE)*, pages 31–36. IEEE, 2017.
- [13] T. Fischer and C. Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.
- [14] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.
- [15] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.
- [16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [17] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *Proceedings of the International Conference on Learning Representations*, 2020.
- [18] G. Lai, W.-C. Chang, Y. Yang, and H. Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.
- [19] G. Lample, A. Sablayrolles, M. Ranzato, L. Denoyer, and H. Jégou. Large memory layers with product keys. In *Advances in Neural Information Processing Systems*, pages 8546–8557, 2019.
- [20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [21] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*, pages 5244–5254, 2019.

- [22] Y. Mäkinen, J. Kannianen, M. Gabbouj, and A. Iosifidis. Forecasting jump arrivals in stock prices: new attention-based network architecture using limit order book data. *Quantitative Finance*, 19(12):2033–2050, 2019.
- [23] A. Ntakaris, M. Magris, J. Kannianen, M. Gabbouj, and A. Iosifidis. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8):852–866, 2018.
- [24] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [25] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.
- [26] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.
- [27] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap. Compressive transformers for long-range sequence modelling. In *Proceedings of the International Conference on Learning Representations*, 2020.
- [28] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.
- [29] S.-Y. Shih, F.-K. Sun, and H.-y. Lee. Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, 108(8-9):1421–1441, 2019.
- [30] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [31] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin. Adaptive attention span in transformers. *arXiv preprint arXiv:1905.07799*, 2019.
- [32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [33] D. T. Tran, A. Iosifidis, J. Kannianen, and M. Gabbouj. Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE transactions on neural networks and learning systems*, 30(5):1407–1418, 2018.

- [34] A. Tsantekidis, N. Passalis, A. Tefas, J. Kannianen, M. Gabbouj, and A. Iosifidis. Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 1, pages 7–12. IEEE, 2017.
- [35] A. Tsantekidis, N. Passalis, A. Tefas, J. Kannianen, M. Gabbouj, and A. Iosifidis. Using deep learning to detect price change indications in financial markets. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2511–2515. IEEE, 2017.
- [36] A. Tsantekidis, N. Passalis, A. Tefas, J. Kannianen, M. Gabbouj, and A. Iosifidis. Using deep learning for price prediction by exploiting stationary limit order book features. *arXiv preprint arXiv:1810.09965*, 2018.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [38] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.
- [39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [40] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, M. Shanahan, V. Langston, R. Pascanu, M. Botvinick, O. Vinyals, and P. Battaglia. Deep reinforcement learning with relational inductive biases. In *Proceedings of the International Conference on Learning Representations*, 2019.
- [41] Z. Zhang, S. Zohren, and S. Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019.
- [42] E. Zivot and J. Wang. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-Plus*, pages 385–429, 2006.

A Training curves

We plot the training and validation history with respect to accuracy for both our TransLOB architecture in Figure 3 and the baseline CNN architecture of [34] in Figure 4.

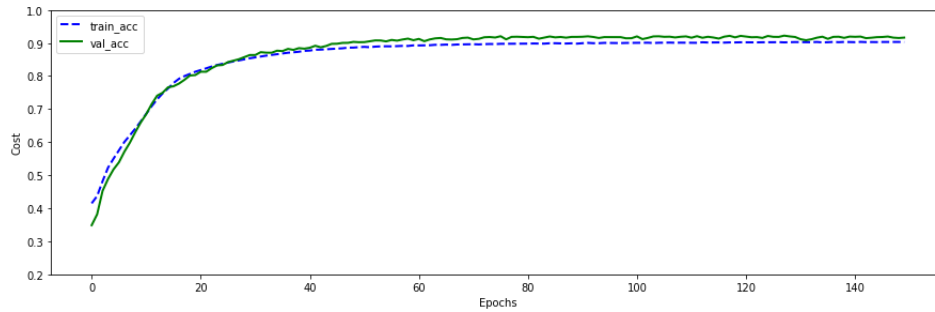


Figure 3: Training and validation accuracy for TransLOB for $k = 100$.

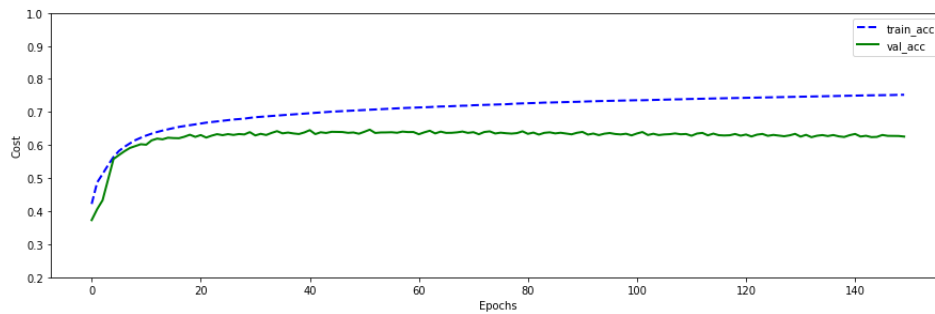


Figure 4: Training and validation accuracy for baseline CNN for $k = 100$.

B Attention distributions

We include here the remaining visualizations of the attention output of our learned model in the first transformer block. Input is a random sample for the horizon $k = 10$.

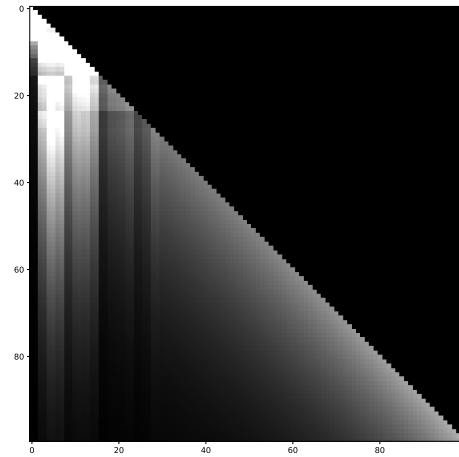


Figure 5: Second head of the first transformer block.

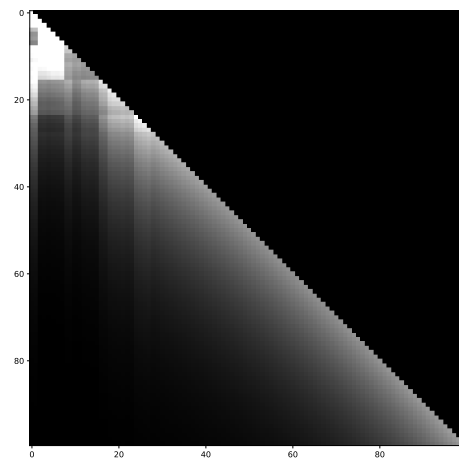


Figure 6: Third head of the first transformer block.