Multimodal Block Transformer for Multimodal Time Series Forecasting

Sungho Park[1]
[1]Dept. of Applied Artificial Intelligence, Korea University

cuzai@naver.com

Abstract

Time series forecasting can be enhanced by integrating various data modalities beyond the past observations of the target time series. This paper introduces the Multimodal Block Transformer, a novel architecture that incorporates multivariate time series data alongside multimodal static information, which remains invariant over time, to improve forecasting accuracy. The core feature of this architecture is the Block Attention mechanism, designed to efficiently capture dependencies within multivariate time series by condensing multiple time series variables into a single unified sequence. This unified temporal representation is then fused with other modality embeddings to generate a non-autoregressive multi-horizon forecast. The model was evaluated on a dataset containing daily movie gross revenues and corresponding multimodal information about movies. Experimental results demonstrate that the Multimodal Block Transformer outperforms state-of-the-art models in both multivariate and multimodal time series forecasting.

1. Introduction

Multivariate time series data can be processed using various methods. One common approach involves concatenating the time series variables along the feature dimension, expressed as:

$$[X_1, \dots, X_n] \in \mathbb{R}^{t\times nd},\;\; X_n = \left[\boldsymbol{x}_{n,1}, \dots, \boldsymbol{x}_{n,t}\right]^T,\;\; \boldsymbol{x}_{n,t} \in \mathbb{R}^d, \tag{1}$$

where X_n represents the n-th time series variable with t time steps and d dimensions. However, this approach assumes that all variables contribute equally over time, neglecting potential variations in the importance of each variable at different time steps. For instance, in a multivariate time series forecasting scenario involving product price and the day of the week as variables for predicting sales, the importance of each variable may change depending on the context. On days when discounts are offered, price might be the dominant factor, whereas on weekends, the day of the week could have a larger impact due to increased customer traffic.

The attention mechanism provides a promising solution for this problem by capturing the dynamic importance of variables over time. However, applying attention to multivariate time series requires an alternative concatenation strategy such as:

$$[X_1^T, X_2^T, \dots, X_n^T]^T \in \mathbb{R}^{tn\times d}, \tag{2}$$

to enable attention computation across both different time series variables and time steps. A key challenge with this approach is the computational complexity of the attention mechanism, which scales quadratically with sequence length. Applying attention to the concatenated sequence requires $(tn)^2$ operations, meaning that for each additional sequence element, $n^2(2t+1)$ additional operations are needed:

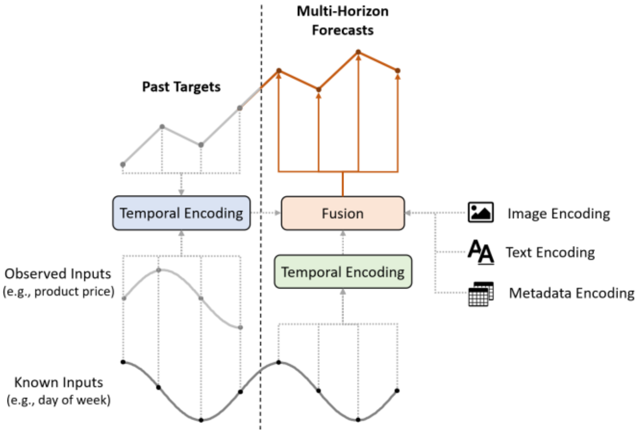
$$(t+1)^2 n^2 = t^2 n^2 + n^2(2t+1). \tag{3}$$

Similarly, introducing an additional time series variable adds $t^2(2n+1)$ operations:

$$t^2(n+1)^2 = t^2 n^2 + t^2(2n+1). \tag{4}$$

These equations highlight the need for a more sophisticated approach to manage attention in multivariate time series, as naïve concatenation becomes computationally expensive, particularly for large-scale datasets.

Another important consideration is the forecasting strategy. The traditional Transformer [1] employs an autoregressive approach, where each prediction relies on previous outputs. This autoregressive nature can lead to error accumulation in time series forecasting, which is more problematic compared to language transduction tasks, for which the Transformer was originally designed.



(Figure 1) Process overview of multimodal time series forecasting.

This paper proposes the Multimodal Block Transformer, a novel architecture that integrates multivariate time series data with multimodal static data, such as images, text, and tabular metadata that remain consistent over time. As illustrated in

(Figure 1), the Multimodal Block Transformer employs an encoder-decoder structure, with the encoder processing past multivariate time series alongside static multimodal data. The decoder fuses these encodings with future-known multivariate time series, which can be predetermined such as the day of the week, to produce a non-autoregressive multi-horizon forecast that predicts multiple future time steps simultaneously.

Using a dataset collected from the web, which includes movie gross revenues and detailed movie-related information, this paper demonstrates that the Multimodal Block Transformer significantly outperforms existing benchmarks in both multivariate and multimodal times series forecasting.

2. Related Works

2.1 Temporal Fusion Transformer

The Temporal Fusion Transformer (TFT) [2] is a state-of-the-art deep learning model designed specifically for interpretable multivariate time series forecasting. TFT is capable of processing heterogeneous data, including both known and observed time-varying inputs, as well as time-independent static features. It effectively addresses the complexity of multivariate time series forecasting through its novel architecture, which incorporates a Gating mechanism, Variable Selection Networks, Static Covariate Encoders, and Temporal Processing layers.

Despite its sophisticated design, TFT is primarily suited for tabular, structured data and does not extend its functionality to multimodal data.

2.2. GTM-Transformer

The GTM-Transformer [3] is designed for multimodal time series forecasting, particularly in scenarios where historical sales data is unavailable, such as predicting sales for new products. It leverages Google Trends data from the period before a product launch, along with visual and metadata information about the product. While the GTM-Transformer effectively handles both multivariate time series and multimodal data, its design focuses primarily on new product forecasting, which limits its ability to utilize historical time series information. Additionally, the GTM-Transformer employes a simple concatenation strategy during the early fusion stage, where features from different modalities are merely concatenated at the input level. This approach constrains the model's capacity to capture complex relationships and interactions between the various features.

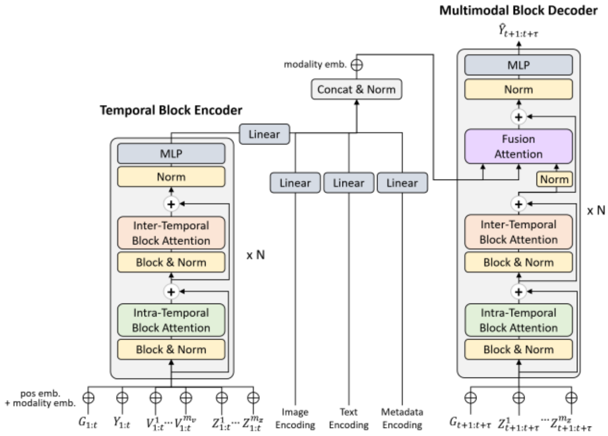
3. Model Architecture

The Multimodal Block Transformer processes multivariate time series data in a similar manner to the Temporal Fusion Transformer. It categorizes time series variables into three types: the target time series Y, the observed time series V^{m_v}, which can only be measured at each time step, and the known time series Z^{m_z}, which can be predetermined such as the day of the week at time step t. The variables Y, V^{m_v} and Z^{m_z} are used to encode historical temporal information, while only Z^{m_z} is used for encoding future temporal information. Additionally, a learnable sequence of embeddings G, referred to as the global temporal embedding, is introduced to aggregate multiple time series variables into a single sequence of vectors.

Each time series variable is represented as a matrix in $\mathbb{R}^{(j-i+1)\times d}$. For example, the observed time series variable can be expressed as:

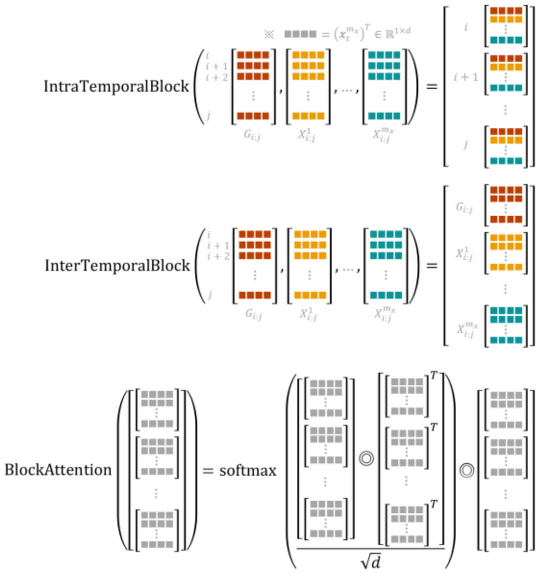
$$V_{i:j}^{m_v} = \left[\boldsymbol{v}_i^{m_v}, \boldsymbol{v}_{i+1}^{m_v}, \dots, \boldsymbol{v}_j^{m_v}\right]^T, \quad \boldsymbol{v}_t^{m_v} \in \mathbb{R}^d, \qquad (5)$$

for each time step $t = i, \dots, j$, and positional encoding is applied before being fed into the model.



(Figure 2) Architecture of Multimodal Block Transformer

3.1 Temporal Block Encoder



(Figure 3) Components of the Temporal Block Encoder

As illustrated in (Figure 3), the Temporal Block Encoder comprises three main operations:

Intra-Temporal Block Operation captures relationships between different time series variables within each single time step. It combines multiple time series variables into a block matrix representation, which can be divided into submatrix blocks. The resulting matrix can be expressed as:

$$\mathrm{IntraTimeBlock}\left(G_{\mathrm{i:j}}, X_{i:j}^{1}, \ldots, X_{i:j}^{m_x}\right) \\ = \left[\left[\boldsymbol{g}_i, \boldsymbol{x}_i^1 \ldots, \boldsymbol{x}_i^{m_x}\right]^T; \ldots; \left[\boldsymbol{g}_j, \boldsymbol{x}_j^1, \ldots, \boldsymbol{x}_j^{m_x}\right]^T\right]. \tag{6}$$

Inter-Temporal Block Operation focuses on capturing relationships across time steps. It concatenates the time series across the sequence from time step i to j, forming a block matrix that reflects temporal dependencies across time steps:

$$\mathrm{InterTimeBlock}\left(G_{\mathrm{i:j}}, X_{i:j}^{1}, \ldots, X_{i:j}^{m_x}\right) \\ = \left[\left[\boldsymbol{g}_i, \boldsymbol{g}_{i+1} \ldots, \boldsymbol{g}_j\right]^T; \ldots; \left[\boldsymbol{x}_i^{m_x}, \boldsymbol{x}_{i+1}^{m_x}, \ldots, \boldsymbol{x}_j^{m_x}\right]^T\right]. \tag{7}$$

Block Attention mechanism applies attention across block matrix $B = [S_1, S_2, \ldots]$ and its transpose $B^T = [S_1^T, S_2^T, \ldots]$, where S_i represents the submatrix blocks produced through either the Intra-Temporal or Inter-Temporal Block Operation. Each submatrix captures a specific aspect of the temporal and multivariate relationships. The Block Attention mechanism computes attention scores using a block-wise dot product operation, allowing the model to focus both on important temporal and multivariate relationships within the time series data:

$$\mathrm{BlockAttention}(B) = \mathrm{softmax}\left(\frac{B \circledcirc B^T}{\sqrt{d}}\right) \circledcirc B, \tag{8}$$

where $\circledcirc$ denotes the block-wise dot product operation.

The Temporal Block Encoder requires $tn^2 + t^2 n$ operations for computing the attention score as the number of operations required for Inter-Temporal and Intra-Temporal Block Attention are tn^2 and $t^2 n$ respectively. It is $\frac{t+n}{tn}$ of the computations required for naïve concatenation strategy. As t and n increase, this approach significantly reduces computational cost compared to equations (3) and (4). This reduction not only alleviates the computational burden but also enhances the model's ability to learn meaningful patterns across multivariate time series.

3.2 Multimodal Encodings

Before integrating multimodal data with the output of the Temporal Block Encoder, the data undergoes processing to enhance self-representation.

The image encoding module utilizes the Vision Transformer (ViT) architecture [4] to extract image patch embeddings $I \in \mathbb{R}^{m_i \times d}$, where $m_i = \frac{HW}{P^2}$ with (H, W) representing the resolution of the original image and (P, P) the resolution of each image patch. Thus, m_i denotes the number of image patches.

The text encoding module uses the BERT architecture [5] to extract language token embeddings $T \in \mathbb{R}^{m_t \times d}$, where m_t denotes the number of tokens in the text.

The metadata encoding module relies on the standard Transformer architecture and is represented as:

$$M = \mathrm{Transformers}([m^1, m^2, \ldots, m^{m_m}]^T + E_{\mathrm{mod}}), \tag{9}$$

where $E_{\mathrm{mod}} \in \mathbb{R}^{m_m \times d}$ represents the modality embeddings and $m^{m_m} \in \mathbb{R}^d$ represents the metadata inputs.

3.3 Multimodal Block Decoder

The fusion process integrates the future-known temporal encodings with other modality encodings using the cross-attention mechanism. As illustrated in (Figure 2), the global temporal embedding $G_{t+1:t+\tau}$ and future-known inputs $Z_{t+1:t+\tau}^{1}, \ldots, Z_{t+1:t+\tau}^{m_z}$ are processed using Intra-Temporal and Inter-Temporal Block Attention. This is then fused with concatenated modality encodings through the cross-attention, enabling the model to attend to and integrate information from multiple modalities. After processing this through the MLP layer, the model produces multi-horizon forecasts for multiple future time steps $\hat{Y}_{t+1:t+\tau} = \{\hat{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_{t+2}, \ldots, \hat{\boldsymbol{y}}_{t+\tau}\}$, effectively mitigating the error accumulation problem commonly associated with regression tasks.

4. Experiment

4.1 Data

To evaluate the performance of the Multimodal Block Transformer, a dataset was compiled, consisting of movie gross revenues and detailed movie information, including the main poster, synopsis, directors, and casts for each movie. This data was collected from the web, covering box office results in North America from 2005 to 2019. This time range was specifically chosen to avoid distortions in patterns caused by the COVID-19 pandemic.

Rather than augmenting the time series data using the traditional window-shifting method, which involves sliding a fixed-length window across the time steps, a method of selecting random sequence lengths was employed. Since each movie has a different length of sales sequence, using window-shifting could result in movies with longer sequences dominating the dataset, potentially misleading the model during training. By applying random length selection, similar to the concept of random cropping in image augmentation, the distribution of movies remains consistent while allowing the model to observe various input lengths during training epochs. Unlike random cropping in images, where the crop is taken from a random position, the sequence always starts from the first time step, with only the length varying. This approach enables the model to forecast time steps of varying lengths based on different lengths of historical data inputs.

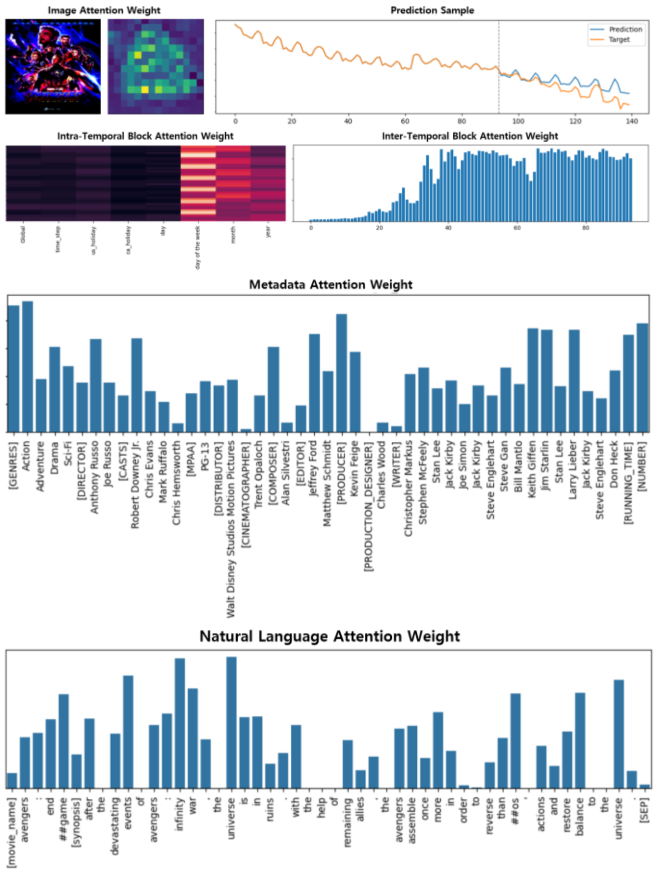
4.2 Results

As shown in <Table 1>, the Multimodal Block Transformer outperformed other models in terms of Mean Squared Error (MSE). While the proposed model with full modality achieved the best performance, the version using only temporal information also performed exceptionally well, achieving better result than both the TFT and the GTM-Transformer. The strong performance of the temporal-only model highlights the effectiveness of the Multimodal Block Transformer architecture, particularly its ability to capture complex temporal dependencies through the Block Attention mechanism.

<Table 1> Performance comparison with Multimodal Block Transformer and benchmarks

Models		MSE
Multimodal Block Transformer (proposed)	**<u>Full modality</u>**	**<u>0.1193</u>**
	Temporal only	0.1259
	Temporal + Metadata	0.1239
	Temporal + Text	0.1247
	Temporal + Image	0.1273
TFT	-	0.3710
GTM-Transformer	-	0.4458

4.3 Interpretation of Attention Score



(Figure 4) Visualization of Sample Attention Weights

(Figure 4) illustrates how the model attends to different aspects of each modality. The image attention weights identify the parts of the image that carry the most semantic importance. The Intra-Temporal Block Attention weights highlight that date-related information, such as day of the week, month and year, plays a more significant role compared to holiday features, with the day of the week showing a clear weekly pattern. This pattern is also observed in the Inter-Temporal Attention weights, which not only emphasize the importance of recent historical data over older data but also reveal fluctuations that follow a weekly cycle. Additionally, the metadata attention weights indicate that the model focuses on key elements within each attribute, such as the genre “action”, the presence of a symbolic actor like “Robert Downey Jr.”, or a major production studio like “Walt Disney”. Similarly, the natural language attention weights demonstrate that the model attends to relevant language tokens associated with the movie’s content.

This interpretability is especially valuable in scenarios where factors like product design are critical. For instance, in the case of a fashion product, the model can capture which specific design patterns of a shirt contribute to higher sales compared to others, providing insights into the product design that attract more customer interest.

5. Conclusion

This study introduced the Multimodal Block Transformer, a novel neural network architecture designed to effectively integrate multivariate time series data with multimodal data, such as images, natural language, and metadata, to enhance time series forecasting. This architecture not only demonstrates improved forecasting performance but also provides valuable insights through the interpretability of attention weights. The fusion process in the proposed model leverages multimodal attention at each time step, offering a detailed understanding of how different aspects of multimodal data influence forecasting outcomes over time. In scenarios where factors like product design are critical, the different attention patterns across time steps can provide deeper insights into the dynamics of the forecasting process, suggesting potential applications in other domains where multimodal data is essential for predictive modeling.

References

[1] Vaswani, A. *et al.* (2017) ‘Attention is all you need’, *Advances in Neural Information Processing Systems,* 2017–December, pp. 5999–6009.

[2] Lim, B. *et al.* (2021) ‘Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting’, *International Journal of Forecasting*, 37(4), pp. 1748–1764.

[3] Skenderi, G. *et al.* (2024) ‘Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends’, *Journal of Forecasting*, 43(6), pp. 1982–1997.

[4] Dosovitskiy, A. *et al.* (2021) ‘An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale’, *ICLR 2021 - 9th International Conference on Learning Representations* [Preprint].

[5] Devlin, J. *et al.* (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, pp. 4171–4186.