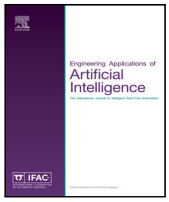
Engineering Applications of Artificial Intelligence 133 (2024) 108606

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Multi-modal transform-based fusion model for new product sales forecasting

Xiangzhen Li [a], Jiaxing Shen [c], Dezhi Wang [a], Wu Lu [b,*], Yuanyi Chen [b,d]

[a] Institute of Digital Finance, Hangzhou City University, Hangzhou, 310015, Zhejiang, China
[b] Hangzhou City University, Hangzhou, 310015, Zhejiang, China
[c] Lingnan University, 8 Castle Peak Road, Tuen Mun, New Territories, Hong Kong
[d] State Key Laboratory of Public Big Data, Guizhou University, 550025, Guizhou, China

ARTICLE INFO

Dataset link: https://paperswithcode.com/sota/new-product-sales-forecasting-on-visuelle

Keywords:
Digital economy
New product sales forecasting
Multi-modal fusion
Temporal feature
Diffusion modeling
Attention mechanism

ABSTRACT

New product sales prediction is crucial for the digital economy as it enables businesses to make informed decisions about product development, inventory management, marketing strategies, and ultimately driving economic growth and innovation. In the digital economy era, traditional sales forecasting methods often struggle to address the unique challenges of forecasting demand for new products, primarily due to limited historical data and high levels of uncertainty. To address this challenge, we propose a multi-modal transform-based fusion model for new product sales prediction (M2TFM), which integrates multiple data sources (e.g., product images, attributes, text descriptions and context factors like holidays, weather and trends.) to predict new product sales with remarkable accuracy. The proposed method leverages diffusion embedding to fuse heterogeneous data modalities including images, text, and time series into a unified representation that models their complex interactions. By encoding multi modal data using Transformer self-attention, our approach is able to extract nuanced signals across modalities to make more accurate new product sales forecasts. We perform a comprehensive evaluation on a large e-commerce dataset with more than 10,000 fashion items, and the results demonstrate that the proposed method is more effective than existing state-of-the-art baselines for new product sales forecasting.

1. Introduction

As the digital economy continues to evolve at warp speed, accurate sales forecasting is crucial for businesses looking to stay competitive in today's fast-paced e-commerce landscape (Ma and Fildes, 2021; Skenderi et al., 2022). With countless options online, consumers expect a seamless shopping experience whether they are browsing on a desktop or mobile device. It is more crucial than ever for e-tailers to understand market behaviors and anticipate demand shifts ahead of time. Having the right analytics tools and insights allows companies to seamlessly adapt their strategies based on real-time market signals. Instead of purely reacting to what happened in the past, they can focus on proactively driving future outcomes. This is especially critical for cross-border e-commerce to respond to market signals in real time, which can calibrate their production and distribution strategies effectively. By anticipating demand fluctuations across different markets, they can mitigate operational risks and bolster profitability. In today's data rich e-commerce landscape, leveraging advanced analytics for sales forecasting is not merely beneficial but fundamental for gaining a competitive advantage.

On the other hand, accurately forecasting sales for newly launched products faces significant challenges from shifting consumer preferences, pervasive social media impacts, and intense competition in crowded online marketplaces. The existing methods can be classified into the following categories (please refer to the related work section for more specific details): (1) Traditional time series prediction techniques, which utilizes aggregate sales data from existing products are found to be ineffective due to limited sales history and differences in sales trajectories; (2) Transfer learning approaches and clustering techniques, which aim to address these limitations but still struggle to account for variations in product attributes, regional distinctions, and evolving consumer tastes. Additionally, the influence of product imagery, an important factor in consumer choice, is often neglected in these methods; (3) Another methods have explored forecasting directly from product images using encoder–decoder architectures, but scalability becomes an issue for large product catalogs. Overall, the existing approaches have limitations in effectively addressing the challenges of sales forecasting for new products, including the integration of diverse data sources and the consideration of product imagery in predictions. Specifically, time series models like ARIMA and exponential smoothing,

* Corresponding author.
E-mail address: luwu@zucc.edu.cn (W. Lu).

https://doi.org/10.1016/j.engappai.2024.108606
Received 27 December 2023; Received in revised form 29 April 2024; Accepted 9 May 2024
Available online 23 May 2024

since they extensively rely on historical data that lacking for new products (Gustriansyah et al., 2019; Manikandan et al., 2022; Singh et al., 2020; Giri et al., 2019). With no sales records, some methods analyze analogous legacy products to infer demand patterns, but diverging trajectories between old and new items introduce uncertainty (Yan and Hu, 2023; Oliveira and Ramos, 2023; Chu et al., 2023). Recent studies shows promise by using transfer learning based prediction methods to new product sales (Karb et al., 2020; Krishnamoorthy et al., 2021). It still faces significant challenges in adequately accounting for inherent variations in sales patterns arising from differences across products, markets, consumer segments, and other factors, since even the most relevant analogs are unlikely to fully replicate a new product's uniqueness. Another approaches focus on individual stores or loose store clusters, requiring many distinct models or masking cluster differences (Puspita et al., 2019; Yin et al., 2020). In summary, lacking sales history and cross-store variability creates an acute cold start forecasting challenge for new products sales prediction (He et al., 2022a,b). The related work section highlights the challenges faced in sales forecasting for new products in the digital economy.

To address these challenges, we propose M2TFM, a multi-modal transform-based fusion model for new product sales prediction, which enhances new product sales forecasting through the following three aspects: multidimensional feature extraction using Convolution Neural Network(CNN) for visual features, sequence models for textual and temporal characteristics, diffusion modeling for multi modal data interaction, and a transform based architecture for capturing interconnections between textual, visual, and temporal elements. Our research question is to test the validity of the various factors of "Image Product Attribute Text (T), Product Images (I), Products Attribute Time Series (A), text description (C) and Exogenous Attributes Time Series (E)" in combination, rather than individually. The proposed multi-modal AI technology for new product sales prediction has significant potential to promote the development of the digital economy. By accurately forecasting demand for new products, businesses can optimize inventory management, reducing waste and improving operational efficiency. This technology also enables more targeted marketing strategies, allowing companies to allocate resources to products with the highest predicted sales potential. Moreover, by providing insights into emerging trends and consumer preferences, this AI-driven forecasting approach can guide product development decisions, helping businesses stay ahead of the curve in the rapidly evolving digital marketplace. As e-commerce continues to grow and competition intensifies, the ability to leverage advanced analytics for proactive decision-making will be a key factor for success in the digital economy.

In a nutshell, the contributions of our research are three-fold:

- We perform multidimensional feature extraction for new product sales prediction, including visual, textual and temporal features. By joint multi-modal modeling via a diffusion model, we capture complex interdependencies and dynamics across different data modalities. This integrative approach offers more accurate sales forecasts.
- We adapt the Transformer model for generating image captions, which enables the model to discern the interconnections between textual, visual, and temporal elements, enriching the complexity and accuracy of sales predictions.
- We empirically validate the proposed method on a large-scale fashion dataset of 10,000+ products, and the results indicate the proposed approach outperforms all the compared methods.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of prior work related to predicting sales of new products. Section 3 explains the multi-modal transformer-based architecture we propose for fusing data modalities to forecast demand. Section 4 describes the large-scale dataset and experimental results comparing our approach to baseline methods. Finally, Section 5 concludes with a summary of key contributions and directions for future research to further advance new product sales forecasting.

2. Related work

Sales forecasting for new products is challenging in today's fast-paced digital economy due to rapidly evolving consumer preferences influenced by social media, shorter product lifecycles, the proliferation of competing online offerings, and limited sales history to train predictive models. Traditional time series prediction techniques like linear regression and K-Nearest-Neighbors(KNN) regression models rely heavily on lengthy sales records, rendering them ineffective when applied to new products (Kohli et al., 2020; Singh et al., 2020; Ma and Fildes, 2021; Jha and Pande, 2021; Giri et al., 2019). Other methods attempt to train forecasting models using aggregate sales data from massive old products (Wolters and Huchzermeier, 2021; Giri and Chen, 2022). However, inherent differences in sales trajectories between products, markets, and consumer segments lead to performance declines. More recent works have explored transfer learning to adapt existing sales models to new products (Karb et al., 2020; Krishnamoorthy et al., 2021). While promising, these methods still do not adequately account for pervasive differences in sales patterns arising from product variations, regional distinctions, and evolving consumer tastes.

To bridge this data gap, some techniques aim to leverage data from similar legacy products as a proxy for new product sales prediction (Yan and Hu, 2023; Oliveira and Ramos, 2023; Chu et al., 2023). However, identifying truly comparable analogs is remarkably difficult, and even the best proxies cannot fully replicate a new product's uniqueness. Following this idea, more recent work has utilized clustering techniques to categorized existing products based on shared attributes and sales histories (Puspita et al., 2019; Yin et al., 2020). Decision trees and other classification models further extend this methodology, assigning new products to these established clusters to predict sales outcomes (Shilong et al., 2021; Wei and Zeng, 2021; Deng et al., 2021). However, clustering based on product attributes alone may overlook critical variables like product images and unattributed visual elements, which can significantly sway consumer interest and purchasing behavior.

Recognizing these limitations, some researchers have proposed methods that combine the analysis of temporal features with product attributes to better predict new product performance. Nevertheless, these efforts often neglect the potent influence of product imagery, an increasingly important factor in consumer choice (Vashishtha et al., 2020). An emerging approach involves forecasting directly from product images, which contain valuable details related to style, design, and visual appeal (Skenderi et al., 2021). However, such methods struggle to scale effectively for large and growing product catalogs, as constructing all possible image pairs grows prohibitively expensive for thousands of products. Recent pioneering works model new product forecasting as an image captioning problem using encoder–decoder architectures (Ekambaram et al., 2020). This allows directly linking product visual features with sales time series predictions in an end-to-end differentiable manner. However, most image-based techniques do not holistically integrate other informative data modalities like temporal patterns or external market events.

Multi modal fusion based on deep learning approaches have recently made strides in improving predictive models by leveraging their capacity to discern non-linear patterns and complex interactions within data (Xue and Marculescu, 2023; Roy et al., 2022). Recurrent neural networks(RNN) like Long Short-Term Memory(LSTM) (Li et al., 2023) and SDPANet (Li et al., 2022a) have proven powerful for modeling sales time series, exploiting long-range temporal dependencies. Attention mechanisms help focus the models on the most relevant input signals for each prediction (Chen et al., 2023; Li et al., 2021). Unfortunately, even these advanced models have not fully exploited the potential of multi-modal data sources.

Table 1
Explanation of symbols used in the method.

Symbol	Meaning
$S_t(x)$	Product sales time series for product x at the tth week
x_i	Image associated with product x
x_a	Set of textual attribute labels for product x
x_c	Textual descriptive information of product x
x_d	Product release date for product x
W_i, b_i	Weight matrix and bias vector for visual features
$X_{c,s}$	Filtered synthetic texts
$X_{i,h}$	Human-annotated images
$X_{c,h}$	Human-annotated texts
X_a	Input text for textual feature extraction
L, d	Sentence length and embedding dimension for textual features
Q, K, V	Query, Key and Value matrix for the attention mechanism
d_k	Dimension of the key for the attention mechanism
θ_r	Comprehensive representation of extracted temporal features
W_r, b_r	Weight matrix and Bias vector for temporal feature fusion
θ_i, c_i	Extracted visual feature, image caption feature
e', θ_r	Extracted text feature, temporal feature

3. Methods

For ease of the following presentation, we define the key data structures and notations used in this paper in Table 1. For any given product, denoted as x, we define its product sales time series as $S_t(x)$, where t represents the tth week since the product's introduction to the market. Here, x ranges from 1 to N, representing the total number of products, and t ranges from 1 to T, denoting the maximum time period for prediction. In our study, we consider that each product x is associated with an image x_i, a set of textual attribute labels x_a (including category, color, and material), textual descriptive information x_c corresponding to the image, and a product release date x_d. Additionally, we assume the availability or collection of supplementary sequences of information in the form of Google Trends. Our objective is to efficiently and effectively combine all these information sources to predict $S_t(x)$ with the highest possible accuracy. Importantly, it should be noted that in the context of forecasting sales for new fashion products, we do not have direct access to the sales history $S_{t-1}(x)$, as the product x is new and lacks any past sales records.

As shown in Fig. 1, the proposed M2TFM has three main components: (1) Multidimensional feature extraction, which involves extracting visual features from product images using CNNs, textual features from product descriptions and temporal features from release dates using sequence models; (2) generating new sales trajectories using diffusion model, which are similar to the sales trajectories of existing products; (3) Capturing cross-modal associations using transform based models, which learns interactions between textual, visual and temporal data using attention mechanisms. By integrating these three components, M2TFM is able to utilize a variety of multimodal data sources to forecast new product sales. To show if our model does better results in forecasting (validity and reliability), we applied an experiment before the application of the developed model.

3.1. Extracting visual features

Product image plays an integral role in capturing consumer interest, enhancing the perceived value of products, and distinguishing them in the competitive market, all of which are key factors influencing purchasing decisions and sales. To explore the impact of visual features on the prediction of product sales, we adopt a dual approach utilizing ResNet152 (He et al., 2016) for visual features and BLIP (Li et al., 2022b) for multi modal capabilities to harness complementary strengths.

Initially, we deploy the ResNet152 model, which has been pretrained, to distill visual features from product image. These images undergo preprocessing, which includes padding and resizing to a uniform dimension of 224 × 224 pixels, followed by random horizontal flipping and rotations to enrich data variance. Subsequent normalization aligns the image data with the statistical distribution of the pre-trained ImageNet weights. We then tailor the final layers of ResNet152 to yield a linear output, ensuring the visual features derived align closely with the task of predicting product sales. To facilitate integration into our diffusion model, Within the ResNet framework, we implement a linear layer to compress the visual feature vectors down to 12 dimensions:

$$\boldsymbol{\theta}_i = \mathbf{W}_i f_{\text{ResNet}}(\mathbf{x}_i) + \mathbf{b}_i \tag{1}$$

In this context, $\mathbf{x}_i$ symbolizes the pristine image feature vector, while $\boldsymbol{\theta}_i$ signifies the compact representation achieved post-linear layer application. $\mathbf{W}_i$ and $\mathbf{b}_i$ are the adjustable parameters of the linear layer.

3.2. Extracting caption features

Complementing the ResNet based feature extraction, we employ the BLIP network to refine our capture of visual features from product images. Renowned for its proficiency in processing a mix of textual and visual data, BLIP has been trained across a spectrum of tasks, making it adept at handling the noisy and diverse inputs typical of e-commerce and retail settings. We initialize the BLIP encoder with weights pretrained on the conceptual captions dataset, which provides 3.3M images annotated with captions describing visual concepts. Then fine-tuning BLIP on our dataset D further adapts the model to capture attributes relevant for product images:

$$D = \{(x_i, X_c)\} + \{(x_i, X_{c,s})\} + \{(X_{i,h}, X_{c,h})\} \tag{2}$$

where x_i denotes product images, X_c the associated textual descriptions, $X_{c,s}$ synthetic filtered descriptions, and $X_{i,h}, X_{c,h}$ represent fine-tuning data. The pre-trained BLIP encoder provides rich visual features, which we further tune to focus on product-specific cues critical for forecasting, such as style, shape, color, and texture. The fine-tuned BLIP model yields enhanced visual representations as input to our sales prediction framework.

Given a product image $x_i \in \mathbb{R}^{H\times W\times C}$, we pass the image through the fine-tuned BLIP image encoder f_{enc} to extract visual features $v_i \in \mathbb{R}^{d_v}$:

$$v_i = f_{enc}(x_i) \tag{3}$$

Then, we feed the visual features v_i into the BLIP caption decoder f_{dec} to generate a textual caption t_i. Encode the caption t_i using BERT to obtain textual features $c_i \in \mathbb{R}^{N\times d_c}$, where N is caption length.

$$t_i = f_{dec}(v_i), \quad c_i = \text{BERT}(t_i) \tag{4}$$

The cross-modal features c_i capture high-level descriptors of the visual product image x_i. The BLIP encoder aligns visual and textual semantics, while BERT provides contextual word representations. Together they output rich caption features c_i to describe the product image content for sales prediction.

3.3. Extracting textual features

The textual features of a product, such as category, color, material and labeling information, are likely to influence the consumer's desire to buy the product. For example, if the product is made of cotton and linen, it stands to reason that in warmer seasons, such as summer, consumers will demand a higher level of breathability from the clothing product, and it is logical to assume that cotton and linen will be more likely to sell in the summer months. Similarly, it stands to reason that the short dress category and products with cooler colors would be more difficult to sell in the winter.

The DistilBERT architecture (Sanh et al., 2019) was found to be effective in the task of extracting textual information from multilingual sources as it is smaller than comparable models yet produces similar results. To adapt the textual embeddings for product demand forecasting, we used a pre-trained multilingual DistilBERT model and added

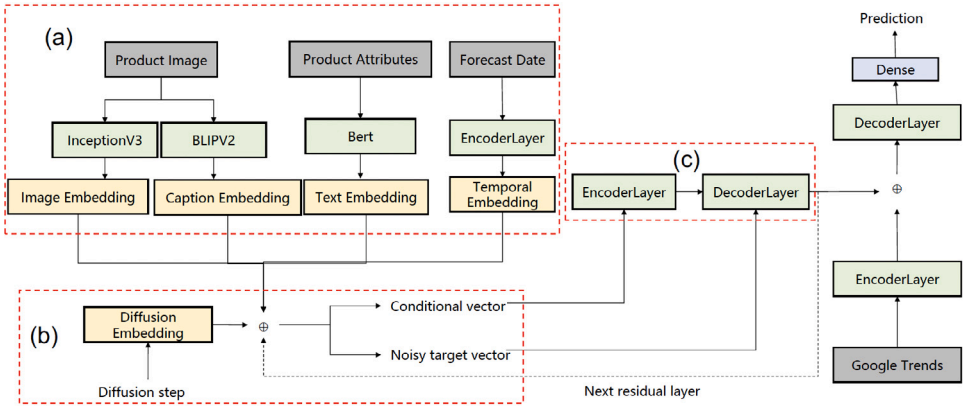


Fig. 1. Multi-Modal model framework for new product sales prediction: (a) multi-modal feature extraction, (b) diffusion model for generating new sales trajectories, (c) transformer model for capturing association between different modalities.

a dimension smaller than the last layer of the original architecture in order to input the dimension as an embedding into the diffusion module. The models were then trained on the same objective as the visual features (average sales of a specific product). The textual features used include product category, color, material and label information. Specifically, the input product text first is tokenized into a sequence of tokens, $\mathbf{X} = \{x_1, x_2, \ldots, x_L\}$, then these tokens are converted to BERT embeddings, $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L\}$, using a pre-trained BERT model (Sanh et al., 2019). Each embedding vector $\mathbf{e}_i \in \mathbb{R}^d$, where d is the dimension of the embedding space. The embedding vectors are mean pooled to create a single embedding vector for the entire text:

$$\mathbf{e} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{e}_i \tag{5}$$

Finally, the pooled embedding vector is projected to the desired dimensionality using a linear transformation $e' = \mathbf{We}$, $\mathbf{W} \in \mathbb{R}^{d \times d'}$, where d' is the desired dimensionality of the embedding vector. The attention weights are used to aggregate the values into an output contextualized representation for each position. This allows each position to build a representation by selectively focusing on other relevant positions in the sequence.

3.4. Extracting temporal features

To capture the temporal characteristics of each product, we utilize a Temporal Transformer to extract temporal feature embeddings from the product's planned release date. These embeddings capture various dimensions of temporal information, such as the day of the week, the week of a month, and the month of a year. The Temporal Transformer consists of multiple layers of self-attention and feed-forward neural networks. Each layer employs a multi-head self-attention mechanism to capture both local and global dependencies within the temporal features. The output of the self-attention mechanism is then fed into a position-wise feed-forward neural network, which applies non-linear transformations to capture complex temporal patterns.

By stacking multiple layers of self-attention and feed-forward neural networks, the Temporal Transformer captures intricate dependencies and patterns in the temporal features. Finally, to obtain a comprehensive representation of all the extracted temporal features, we concatenate the embeddings of the day of the week, the week of a month, and the month of a year, and pass them through a fully connected layer:

$$\theta r = \mathrm{ReLU}(\mathbf{W}r[\mathbf{x}_d; \mathbf{x}_w; \mathbf{x}_m] + \mathbf{b}_r) \tag{6}$$

Here, $\mathbf{x}_d, \mathbf{x}_w$, and $\mathbf{x}_m$ represent the embeddings of the day of the week, week of a month, and month of a year, respectively. $\mathbf{W}_r$ and $\mathbf{b}_r$ are learnable parameters of the fully connected layer.

By incorporating the Temporal Transformer and the fusion mechanism, we obtain a more complex and expressive representation θ_r that captures the rich temporal patterns and relationships within the product's release date.

3.5. Multi-modal feature fusion

Given the extracted visual features θ_i, image caption features c_i, text features e', and temporal features θ_r. Our approach perform Multi-modal feature fusion by tahree steps: (1) feature concatenation; (2) diffusion embedding and (3) transform based embedding.

(1) Feature concatenation: In the first step, we concatenate the features from each modality to obtain a comprehensive representation of the data, as shown in the following equation.

$$\mathbf{f} = [\theta_i, c_i, \mathbf{e}', \theta_r] \tag{7}$$

where $\mathbf{f}$ is the concatenated feature vector.

(2) Diffusion embedding: In the second step, we perform diffusion embedding on the concatenated feature vector $\mathbf{f}$. Diffusion embedding is a technique that can be used to generate new data that is similar to a given dataset. It works by gradually adding noise to the data, and then learning to reverse this process to denoise the data and generate new samples. The diffusion embedding process includes two steps:

Forward diffusion. In this step, noise is gradually added to the data sample, as shown in the following equation.

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{8}$$

where $\mathbf{x}_t$ is the data sample at time step t, β_t is the noise schedule, and ϵ_t is a random noise vector.

Reverse diffusion. In this step, the noise is gradually removed from the data sample, as shown in the following equation.

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \sqrt{\beta_t}\epsilon_t) \tag{9}$$

The diffusion model is trained to learn the parameters of the diffusion process. This is done by minimizing the mean squared error between the original data sample and the data sample that is reconstructed after the reverse diffusion process. Once the diffusion model is trained, it can be used to generate new data. In our approach, we use the diffusion model to generate a conditional vector $\mathbf{c}$ and a noisy target vector $\mathbf{n}$ from the concatenated feature vector $\mathbf{f}$. The conditional vector $\mathbf{c}$ is a representation of the input data that is conditioned on the target data, while the noisy target vector $\mathbf{n}$ is a corrupted version of the target data.

(3) Transform based embedding: After obtaining the conditional vector $\mathbf{c}$ and noisy target vector $\mathbf{n}$ from diffusion embedding, we leverage the transformer architecture to model interactions between

the multi modal features, which computes attention between the conditional vector $\mathbf{c}$ and noisy vector $\mathbf{n}$ to generate an enriched embedding $\mathbf{v}_t$:

$$\mathbf{v}_t = \text{Transformer}(\mathbf{c}, \mathbf{n}) \tag{10}$$

Specifically, the conditional vector $\mathbf{c}$ and target $\mathbf{n}$ are added with positional encodings and then fed into the Transformer encoder. The encoder has multiple self-attention layers, each calculating attention scores between all key-query pairs:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{11}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are query, key and value projections of the input. This allows each position in the sequence to build representations by selectively focusing on relevant contexts.

The Transformer model integrates the information provided by the conditional and noisy vectors to produce a comprehensive embedding that reflects both the stable and stochastic characteristics of the multimodal data of new product. The embedding vector $\mathbf{v}_t$ is then utilized to new product sales prediction. The efficacy of this embedding process lies in the Transformer's ability to synthesize information over various positions and modalities, thus enhancing the model's predictive accuracy and robustness.

3.6. Generate new product sales prediction

The embedding vector $\mathbf{v}_t$ generated by the Transformer encapsulates the interactions between our multi modal input data including visual, textual, and temporal features. To make full use of available data, we also leverage Google Trends information $\mathbf{g}_t$ related to the product. The embedding and trends data are fed as a sequence into an encoder–decoder architecture to forecast new product sales y_t:

$$y_t = \text{Decoder}(\text{Encoder}([\mathbf{v}_t; \mathbf{g}_t])) \tag{12}$$

The encoder maps the concatenated input sequence $[\mathbf{v}_t; \mathbf{g}_t]$ to a higher dimensional representation using multi-headed self-attention. The decoder then uses this representation to autoregressively generate the target sales prediction y_t one step at a time.

3.7. Complexity analysis

This section briefly analyzes the computational complexity for the proposed algorithm for product sales forecasting, which includes the following three steps:

Feature Extraction. For Visual Features (ResNet & BLIP), the complexity of a convolutional layer is given by:

$$O(K^2 * C_{in} * C_{out} * H * W)$$

where K is the kernel size, C_{in} and C_{out} are the number of input and output channels, H and W are the height and width of the feature maps. Considering multiple layers, the overall complexity becomes:

$$O(L * K^2 * C_{avg}^2 * H_{avg} * W_{avg})$$

where L is the number of layers, C_{avg}, H_{avg}, and W_{avg} are the average values for channels, height, and width.

For Textual Features and Temporal Features, the time complexity are $O(N^2 * H * d)$ and $O(T^2 * H * d)$, where N is the sequence length, H is the number of attention heads, d is the hidden dimension size.

Multi-modal Feature Fusion. The complexity of feature concatenation is $O(1)$, while the complexity of diffusion embedding is given by $O(S * D^2)$, where S is the number of diffusion steps, D is the size of the feature vector. For Transformer-based Embedding, the complexity can be calculated by $O(M^2 * H * d)$, where M is the length of the combined feature sequence.

Encoder–Decoder based Sales Prediction. The encoder processes the input sequence of combined features (visual, textual, and temporal) through multiple Transformer layers. Each layer performs self-attention and feed-forward operations. The self-attention mechanism has a time complexity of $O(P^2 * d)$, where P is the length of the input sequence and d is the hidden dimension size. The decoder generates the output sequence (sales predictions) step-by-step, attending to both the encoded representation and the previously generated outputs. Similar to the encoder, each decoder layer involves self-attention and encoder–decoder attention mechanisms, each with a complexity of $O(P^2 * d)$.

Considering L encoder layers and M decoder layers, the overall time complexity of the encoder–decoder architecture for sales prediction can be approximated as:

$$O((L + 2M) * P^2 * d + (L + M) * P * d^2)$$

Combining the complexities of each stage as analyzed in the previous responses, we obtain the overall complexity of the M2TFM algorithm:

$$O(L * K^2 * C_{avg}^2 * H_{avg} * W_{avg} + (N^2 + T^2 + M^2) * H * d + S * D^2 + (L + 2M) * P^2 * d)$$

The dominant term in this complexity is $O((L + 2M) * P^2 * d)$, which indicates a quadratic dependence on the input sequence length. This implies that as the length of the input sequence increases, the computational time required for sales prediction grows significantly. The M2TFM algorithm offers a powerful approach to new product sales forecasting, but its complexity necessitates careful consideration for large-scale applications. By integrating efficient model architectures, approximation techniques, distributed training, and hardware acceleration, M2TFM can be optimized for real-world scalability. These advancements hold the potential to unlock significant benefits for businesses navigating the dynamic landscape of the digital economy.

4. Experiments and results

4.1. Dataset and preprocessing

To conduct our experiments, we utilize a historical time-series dataset spanning two years from a renowned fashion company.[1] The dataset comprises 10,290 products distributed across 45 categories. As shown in the Table 2, each product entry in the dataset includes the following attributes: (i) product attributes such as color, fabric, and category; (ii) product images; (iii) product popularity. Additionally, the dataset provides weekly-level time series data related to each product, including: (i) product release date, (ii) weekly sales of the product, (iii) discounts offered during the week of product release, and (iv) product sales price. To ensure the integrity of our analysis, we partitioned the dataset into a training set and a test set. The division was performed randomly, with 80% of the data assigned to the training set and 20% to the test set for each product. Prior to conducting experiments, we normalized the training and test datasets by calculating the mean and standard deviation of each feature from the training dataset. It is important to note that we incorporated additional feature, namely Google trend, which enable us to capture the market demand and supply dynamics of the products.

To ensure consistency and comparability, we homogenize each feature using min–max normalization, thereby scaling the multimodal features to a common range of 0 to 1. Furthermore, for each model under investigation, we perform hyperparameter tuning to determine the optimal training parameters.

[1] The dataset can be found online at https://paperswithcode.com/dataset/visuelle.

Table 2
Explanation of terminology for dataset items.

Item	Meaning
Product	Clothes products including pants, skirts, t-shirts, etc.
Product attributes	The color, fabric, and category of product
Product images	The image data of product
Product popularity	Popularity of the product over time

Table 3
Glossary of terms and acronyms.

Terms	Meaning
$[T]$	Text features encoding product descriptions
$[I]$	Visual features to capture style, shape, color
$[A]$	Temporal features indicating seasonality, trends
$[C]$	Image caption features
$[E]$	Product attribute features like price, brand, functionality tags
KNN	K-Nearest-Neighbors algorithm
CNN	Convolution Neural network
RNN	Recurrent neural networks
WAPE	Weighted absolute percentage error
MAE	Mean absolute error

4.2. Metrics

To comprehensively evaluate the performance of our model, we employ several metrics that effectively measure the disparity between the forecasted and actual product sales.

One commonly used metric is the Mean Absolute Error (MAE), which provides a straightforward assessment of the model's accuracy. The MAE is calculated as the average absolute difference between each element in the predicted sale $\mathbf{y}$ and the corresponding real sale $\hat{y}$, as follows:

$$\mathrm{MAE}(\mathbf{y}, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{13}$$

Given the presence of seasonal and cyclical patterns in the product sales data, we also incorporate the Weighted Absolute Percentage Error (WAPE) as a metric widely used for evaluating the accuracy of probabilistic forecasts. The WAPE is defined as:

$$\mathrm{WAPE} = \frac{\sum_{t=1}^{T} |y_t - \hat{y}_t|}{\sum_{t=1}^{T} y_t} \tag{14}$$

where T is the forecasting horizon. WAPE is always nonnegative, and a lower value indicates a more accurate model.

4.3. Implements details

We implemented these tests using the Python programming language and the Pytorch framework. The experiment was performed on a workstation equipped with 8x A100 graphics cards. In this study, we set the number of diffusion steps to 50. The noise-adding amount is set to 10^{-4} at the first diffusion step and 0.5 at the last diffusion step. The Transformer is equipped with 12 heads, and the batchsize is set to 128 samples. We use the ADAM optimizer with a learning rate 10^{-3}. The training data is further split into an 80%–20% ratio, with the latter portion used for validation during early stopping to prevent overfitting. Starting from the 100th epoch, the model is tested with the validation dataset every 10 epochs.

5. Result of the model

5.1. Ablation analysis

To comprehensively evaluate the impact of different feature types on product sales prediction, we conduct a systematic ablation study. Specifically, we perform experiments using various combinations of the following data modalities: [T] represents Text features encoding product descriptions and consumer sentiment; [I] represents visual features extracted from product images to capture style, shape, color; [C] represents image caption features; [A] represents temporal features indicating seasonality, trends and external factors; [E] represents product attribute features like price, brand, functionality tags. Table 4 presents the ablation study results for the M2TFM model on the dataset (see Table 3).

Table 4
Six weeks ablative results on VISUELLE.

M2TFM ablations	WAPE (%)	MAE
$[T]$	55.27	30.42
$[I]$	53.91	29.44
$[A]$	54.66	29.85
$[T+I]$	54.45	29.81
$[T+I+C]$	52.96	29.03
$[T+I+A]$	52.85	28.92
$[T+I+A+E+C]$	**51.97**	**28.46**

Table 5
Methods for comparison.

Method	Description
Attribute KNN	KNN algorithm by [A] features
Image KNN	KNN algorithm by [I] features
Attr+Image KNN	KNN algorithm by [A] and [I] features
Cross-Attention RNN	Multimodal attention weights on RNN
Cross-Attention RNN+A (Ekambaram et al., 2020)	Using [A] features to train Cross-Attention RNN
Transformer	Using multimodal features to train transformer
GTM-Transformer	Using multimodal features to train GTM-transformer
GTM-Transformer AR (Skenderi et al., 2021)	Using [A] and [C] features to train GTM-transformer
FusionMLP	Using multimodal features to train MLP network
MuQAR (Papadopoulos et al., 2022)	Using multimodal features to train MuQAR model

Using only the product text feature ([T]) provides a baseline, achieving a WAPE of 55.27% and MAE of 30.42. This indicates there is substantial room for improvement in prediction accuracy. Relying solely on textual descriptions fails to capture important visual attributes and contextual factors that influence product sales. Incorporating the product image ([I]) alone improves results, reducing WAPE to 53.91% and MAE to 29.44. Adding visual data conveys details about appearance, styling, materials, color patterns, and other aesthetics that text cannot fully represent. Images contain intrinsic cues that more directly impact consumer purchase decisions. However, images lack semantic explanatory power and omit broader contextual knowledge. Looking at product temporal features ([A]) in isolation performs comparably to text, with a WAPE of 54.66% and MAE of 29.85. Category sales trends over the past year supply useful market signals but fail to account for specific attributes of the individual product. Relying solely on temporal features overlooks the rich heterogeneity within a segment.

Combining textual and visual data ([T+I]) outperforms either modality independently, achieving a WAPE of 54.45% and MAE of 29.81. This demonstrates the value of fusing semantic and visual features, as text and images provide distinct yet complementary information. The text conveys functional, descriptive attributes while images represent stylistic elements. Further incorporating image caption features ([T+I+C]) supplies additional gains, with WAPE decreasing to 52.96% and MAE to 29.03. Category sales trends place the product in the broader market context, capturing macro-level demand signals. This refinement demonstrates category popularity can enhance predictions when used with specific product data. Adding fine-grained product temporal features ([T+I+A]) like hanging ornaments yields a WAPE of 52.85% and MAE of 28.92, comparable to the above. These specialized metadata fields describe physical characteristics beyond what images or text provide. The details enable better differentiation between similar products.

Table 6
Performance comparison with different baselines. The models were trained on data from "old" products and tested on "new" products, encompassing various features including, image product attribute text [T], product images [I], product attribute time series [A], text descriptions [C], exogenous attribute time series [E] and their combination.

Methods	Input	In:52, Out:6		In:28, Out:6	
		WAPE (%)	MAE	WAPE (%)	MAE
Attribute KNN	[T]	59.8	32.7	59.8	32.7
GTM-Transfomer		62.6	34.2	62.6	34.2
FusionMLP		**55.15**	**30.12**	**55.15**	**30.12**
M2TFM		55.27	30.42	55.27	30.42
Image KNN	[I]	62.2	34	62.2	34
GTM-Transfomer		56.4	30.8	56.4	30.8
FusionMLP		54.59	29.82	54.59	29.82
M2TFM		**53.91**	**29.44**	**53.91**	**29.44**
LSTM	[A]	58.7	32.0	59.8	33.7
Transfomer		62.5	34.1	64.2	35.3
GTM-Transfomer		58.2	31.8	59.5	32.4
FusionMLP		55.15	30.12	55.67	30.51
M2TFM		**54.66**	**29.85**	**55.28**	**30.2**
Attr+Image KNN	[T+I]	61.3	33.5	61.3	33.5
Cross-Attention RNN		59.5	32.3	59.5	32.3
GTM-Transfomer		56.7	30.9	56.7	30.9
FusionMLP		**54.11**	**29.56**	**54.11**	**29.56**
M2TFM		54.45	29.81	54.45	29.81
FusionMLP	[T+I+C]	53.5	29.22	53.5	29.22
M2TFM		**52.96**	**29.03**	**52.85**	**29.03**
GTM-Transfomer AR	[T+I+A]	59.6	32.5	59.4	32.1
Cross-Attention RNN+A		59.0	32.1	58.7	31.9
GTM-Transfomer		55.2	30.2	56.8	31.0
MuQAR		53.61	29.28	54.51	30.1
M2TFM		**52.85**	**28.92**	**54.13**	**29.75**
MuQAR	[T+I+A+E+C]	52.63	28.75	53.74	29.90
M2TFM		**51.97**	**28.46**	**53.82**	**29.5**

Using all data ([T+I+A+E+C]) achieves the lowest WAPE of 51.97% and MAE of 28.46, demonstrating the value of holistic feature fusion. The text provides semantics, images offer visuals, attributes capture specifics, popularity supplies context, and descriptions add nuanced details. Together these heterogeneous data sources enable rich multi-dimensional product representations. In summary, the ablation study quantitatively verifies the complementary value of textual, visual, contextual and metadata product information for sales prediction. Using these features in concert allows capturing semantics, aesthetics, details, market trends and other factors that collectively influence sales volumes. No single modality completely represents the product. The incremental reductions in WAPE and MAE from adding more features demonstrates the modeling benefits of leveraging multimedia product data.

To evaluate the efficiency of the proposed model, we measured the inference time of the model M2TFM to be 92 s, and our dataset contains 2058 test samples, which means that the model M2TFM inference time for a single sample is about 45 ms.

5.2. Comparative analysis

We compare M2TFM with 10 state-of-the-art approaches for new product sales prediction as listed in Table 5. Two experimental configurations are evaluated — using 52 weeks and 28 weeks of historical sales data to forecast 6 weeks of future sales.

The experiment results are presented in Table 6, where M2TFM demonstrates robust performance in both long-term and short-term forecasting tasks. When utilizing 52 weeks of historical data, M2TFM exhibits a slight improvement over FusionMLP, which is the previous best model when considering only text [T] and image [I] inputs individually. Specifically, M2TFM achieves a WAPE of 55.27% and MAE of 30.42 for text, and a WAPE of 53.91% and MAE of 29.44 for images in the 52-week scenario. For the more challenging 28-week configuration, M2TFM shows a notable reduction in WAPE by 0.89% and MAE by 0.17 compared to the next best model, FusionMLP, when combining text, image, and caption features [T+I+C].

Table 7
Statistical significance for methods.

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Cross-Attention RNN+A	31.83	31.42	32.71	32.26	31.95
GTM-Transformer AR	32.89	33.15	33.21	32.78	33.05
MuQAR	29.46	29.76	30.01	29.68	30.11
M2TFM	28.15	28.43	28.24	28.17	28.45

When examining performance by input types, M2TFM displays competitive results. For textual attributes [T] alone, M2TFM is marginally outperformed by FusionMLP. However, when evaluating the utility of product images [I], M2TFM surpasses all other image-based models, emphasizing its capacity to effectively interpret and utilize visual data. In the context of time series data [A], M2TFM excels beyond traditional sequence modeling approaches like LSTMs and Transformers, indicating that M2TFM's temporal encoder is more adept at capturing category-level trends, which is vital for forecasting sales of new products. Moreover, when integrating text, images, and time series [T+I+A], M2TFM achieves a lower WAPE by 0.76% and a lower MAE by 0.36 than MuQAR, the best prior multimodal method for the 52-week data set. This showcases the strength of M2TFM's joint multimedia and temporal modeling capabilities. Finally, when all features are included [T+I+A+E+C], M2TFM continues to lead with the lowest WAPE and MAE, asserting its dominance as a comprehensive, multimodal forecasting tool.

Integrating image captions [C] yields notable improvements, with M2TFM achieving a WAPE of 52.96% on the 52-week task and 52.85% on the shorter 28-week task, demonstrating the benefit of leveraging detailed textual information for a more comprehensive understanding of products. When combining all data modalities [T+I+A+E+C], our method advances the state-of-the-art, with a WAPE of 51.97% for the 52-week forecast and 53.82% for the 28-week forecast, underscoring the significance of a holistic approach to heterogeneous data integration in sales prediction.

The results can be attributed to several key design choices. The two-stream architecture of M2TFM efficiently captures visual and textual signals from product images and attributes. The model facilitates cross-modal interactions allowing for a bidirectional exchange of semantic information between different modalities. Additionally, the incorporation of temporal context modeling is critical in providing essential category-level insights. These elements work in concert to form multi-dimensional product representations that enhance the accuracy of sales forecasts (see Table 7).

To enhance the credibility of results, we further assess four models with superior performance (namely Cross-Attention RNN+A, GTM-Transformer AR, MuQAR and M2TFM) using MAE and conduct an ANOVA test (Stoker et al., 2020) to examine if there are any statistically significant differences between the models. If significant differences are found, we will use Tukey's HSD test (Nanda et al., 2021) to identify which model performs the best. We randomly divide the entire dataset into five equal parts, each part will serve once as a validation set, with the remaining four parts used as the training set. For each cross-validation fold, we train on each of the training sets and compute the MAE on the validation set by using 52 weeks historical sales data to forecast 6 weeks of future sales. The results are shown in Table 6.

We perform an ANOVA test to determine if there are statistically significant differences in performance among the four models based on their MAEs. The p-value from the ANOVA test is 0.998, indicating significant differences in MAE among at least one of the models. Then, we further perform Tukey's HSD test to determine which specific models differ from each other, the results indicate that Cross-Attention RNN+A and GTM-Transformer AR are not significantly different, but all other

Table 8
Comparative sales prediction results for the top product categories.

Model	Coat		Dress		Skirt		T-shirt	
	WAPE (%)	MAE	WAPE (%)	MAE	WAPE (%)	MAE	WAPE (%)	MAE
Cross-Attention RNN	59.25	35.4	62.05	31.2	51.61	26.3	54.89	29.8
GTM-Transformer	51.83	24.50	51.10	18.02	51.54	26.14	53.85	29.04
MuQAR	53.41	30.2	52.28	21.4	54.28	28.7	52.71	28.7
M2TFM	**52.82**	**25.40**	**50.81**	**17.99**	**50.47**	**25.8**	**51.97**	**28.46**

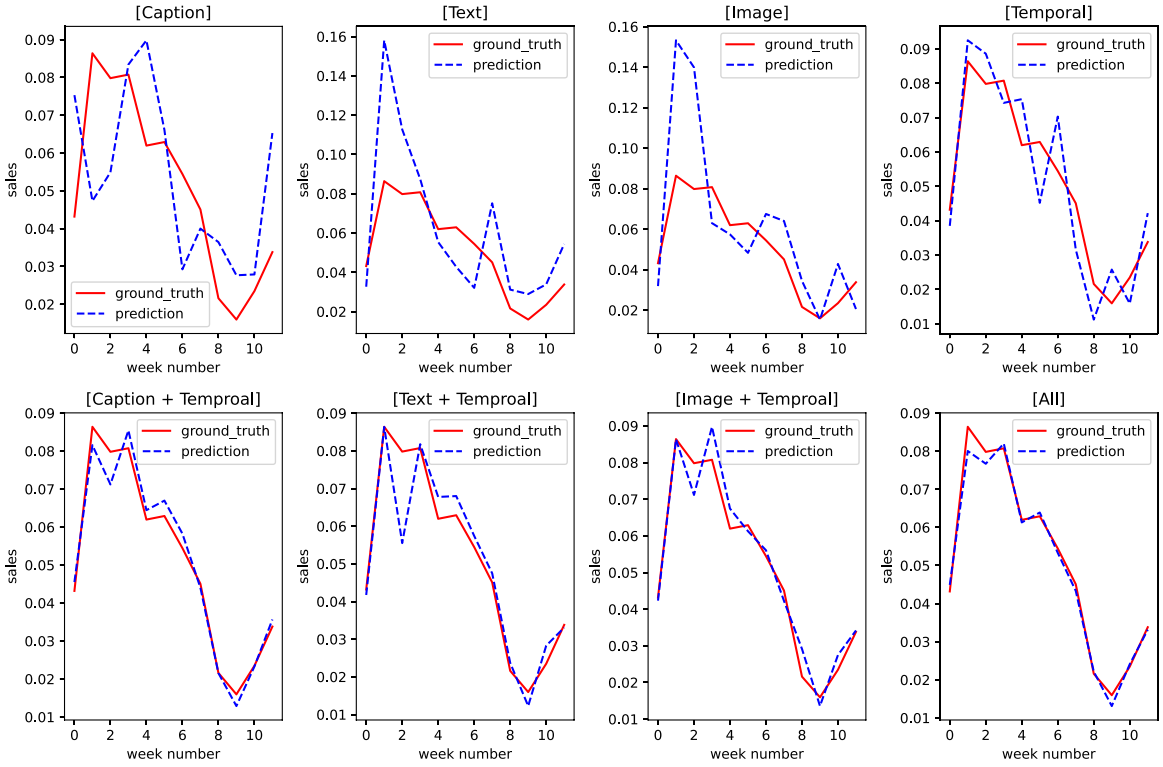


Fig. 2. The impact of different input dimensions on sales prediction for skirt category.

model pairs differ significantly in their MAE performance. Based on the results from the ANOVA and Tukey's HSD test, the proposed M2TFM with the consistently lowest average MAE across multiple datasets is the best model based on data-driven analysis and statistical significance.

5.3. Product category comparison analysis

In the comparative analysis of sales prediction models, our research presents compelling evidence that the M2TFM model outperforms other established cross-attention methods. We evaluated the performance of these models across four major product categories: Coats, Dresses, Skirts, and T-shirts. The results, as summarized in Table 8, indicate that M2TFM consistently achieved lower WAPE and MAE across all categories, demonstrating its superior forecasting accuracy.

For Coats, M2TFM achieves WAPE of 52.82% and an MAE of 25.40, indicating a more precise forecast compared to Cross-Attention RNN, GTM-Transformer, and MuQAR, which scored higher WAPEs of 59.25%, 51.83%, and 53.41%, respectively. This trend of M2TFM's outperformance continued in the Dress category, where it achieved the lowest WAPE of 50.81% and the lowest MAE of 17.99, surpassing the closest competitor, GTM-Transformer, by a margin of 0.29 percentage points in WAPE. In the Skirt category, the M2TFM model again led with a WAPE of 50.47%, which is a considerable achievement given that the second-best model, GTM-Transformer, reported a WAPE of 51.54%. Lastly, for T-shirts, M2TFM achieved a WAPE of 51.97% and an MAE of 28.46, which, while closer to its competitors, still maintains a performance edge. The better performance of M2TFM across these diverse categories shows the model's robustness and generalization capabilities. Its success can be attributed to its innovative architecture that effectively harnesses cross-modal interactions, which are essential for accurate sales forecasting for new product.

Fig. 2 shows the impact of different input dimensions on sales prediction for skirt category. The results show that the best performance is achieved when all of the input dimensions are used. This is because each of the input dimensions provides different information about the product, and by using all of the dimensions, we can get a more complete picture of the product and make more accurate predictions. We can also see that the performance of the model is significantly worse when only the caption or the text is used. This is because these dimensions do not provide as much information about the product as the image or the temporal dimension. The temporal dimension is particularly important for predicting sales, as it can help us to identify trends and seasonality in the data. For example, we can see that sales of skirts are typically higher in the summer months, and lower in the winter months. This information can be used to make more accurate predictions about sales.

6. Discussion

6.1. Advantage of developed model

From the experimental comparison results in Table 6, we can see that the dual-stream architecture of our proposed M2TFM can effectively capture visual and textual signals from product images and attributes. Compared to baselines using only a single modality like

Attribute KNN, Image KNN or LSTM, M2TFM achieves significantly better performance by fusing multimodal data like [T+I], [T+I+A] and [T+I+A+E+C]. The integration of multimodal features in our proposed M2TFM model brings several advantages and contributes to the successful application of sales forecasting. When using only text features [T], visual features [I] or time series features [A] individually, the best WAPE scores are 55.15%, 53.91% and 54.66% respectively. However, by combining multimodal data like [T+I], M2TFM can lower the WAPE to 54.45%. Further adding temporal attributes [A] to [T+I+A] reduces WAPE to 52.85%. The full combination of [T+I+A+E+C] enables M2TFM to achieve the best WAPE of 51.97%.

Firstly, the dual-stream architecture of M2TFM allows for effective capture and fusion of visual and textual signals from product images and attributes. By leveraging both visual and textual information, the model can extract complementary features and gain a more comprehensive understanding of the product. Visual signals, such as product images, contain rich visual cues that can convey important information about the product's appearance, packaging, and design. On the other hand, textual signals, such as product attributes or descriptions, provide valuable semantic information and product specifications. By combining these modalities, M2TFM can leverage the strengths of each modality and capture a more holistic representation of the product, leading to improved sales forecasting accuracy.

Secondly, the bidirectional exchange of semantic information facilitated by the cross-modal interactions in M2TFM is crucial for capturing the interplay between different modalities. Visual and textual signals often provide complementary information, and their combination can enhance the understanding of product characteristics and customer preferences. By allowing the flow of information between modalities, M2TFM can capture the semantic relationships between visual and textual features, enabling a more nuanced representation of the product and its market potential. This cross-modal fusion helps to uncover hidden patterns and insights that may not be apparent when considering each modality separately, leading to more accurate sales forecasts.

Furthermore, the incorporation of temporal context modeling in M2TFM provides an additional layer of information that is crucial for sales forecasting. By considering the temporal dynamics and trends in product sales over time, the model can capture the seasonality and evolving consumer preferences within different product categories. This category-level insight is valuable for understanding the demand patterns and predicting future sales performance accurately. By incorporating temporal context, M2TFM enhances the predictive power of the model and enables more informed decision-making for businesses.

Overall, the integration of multimodal features and the utilization of temporal context in M2TFM lead to a multidimensional product representation, which improves the accuracy of sales forecasting. The model's ability to capture visual and textual signals, facilitate cross-modal interactions, and incorporate temporal context provides valuable insights into product sales performance. This, in turn, enables businesses to make better-informed decisions, optimize their sales strategies, and allocate resources more effectively. The application of M2TFM in sales forecasting empowers businesses with a competitive advantage by providing a comprehensive understanding of their products, customers, and market dynamics.

6.2. Practical application

Our proposed M2TFM methodology excels in new product sales forecasting by combining multiple factors and characteristics to more accurately predict product sales performance. The introduction of this technique has important background and benefits for organizations.

First, accurate sales forecasting helps organizations make more informed sales strategies and decisions. By using the M2TFM methodology, companies can gain a better understanding of market demand and consumer behavior patterns. This in-depth understanding enables organizations to adjust their market positioning, pricing strategies, and sales channels to better meet consumer needs and achieve better sales performance.

Second, accurate sales forecasting also helps companies assess market demand and product potential. By utilizing the M2TFM methodology, companies can better understand the potential market size and development trends, and thus more accurately assess the market potential of new products. In this way, companies can avoid over-investing in products for which there is no market demand and target resources to products with potential, maximizing returns.

In addition, accurate sales forecasts can help companies avoid inventory backlogs and supply chain problems, thereby reducing costs and increasing profits. With the accurate forecasts provided by the M2TFM methodology, companies can plan their production and supply chain activities more accurately, avoiding wasted resources and capital utilization due to overproduction or inventory backlogs. At the same time, enterprises can better coordinate with suppliers to ensure the stability and efficiency of the supply chain, thereby reducing supply chain risks and improving the operational efficiency of the enterprise.

In short, by seamlessly integrating the M2TFM methodology into actual business processes, companies can improve the accuracy of sales decisions, reduce market risks, optimize resource allocation, and increase customer satisfaction. Such accurate sales forecasts and optimized business processes will bring great business value to enterprises and enhance their competitiveness and influence in the fiercely competitive market. As a result, companies adopting the M2TFM methodology will be better able to seize market opportunities, realize sustainable business growth, and achieve a leading position in the industry.

6.3. Potential problem

Our proposed M2TFM model still faces many challenges, and we will conduct future research in the following areas.

In terms of datasets, the performance of the M2TFM model may be affected by incomplete or missing data. In practice, situations may be encountered where certain key features or data are missing, which may lead to limitations in the predictive power of the model. Further research could explore how to deal with missing data or develop more robust models to address this challenge. In terms of dataset size, in order to cope with the challenges of large-scale datasets, e.g., distributed computing and incremental learning can be used in the future to improve the scalability of the model.

In terms of model complexity and interpretability, the M2TFM model has high model complexity, which may lead to difficulties in its interpretation and explanation. In practical applications, businesses and decision makers may need to understand the model's reasoning process and the reasons for the predicted results. Further research could explore how to improve the interpretability of models to better meet the needs of real-world applications.

In terms of dataset labeling, M2TFM models require a large amount of labeled data for training. In practical application scenarios, obtaining large-scale labeled data may be a challenge. Further research could explore how methods such as semi-supervised or unsupervised learning can be utilized to reduce the dependence on labeled data and thus extend the applicability of the model.

7. Conclusion

In this paper, we present M2TFM, a novel multi-modal transformer-based fusion model for forecasting sales of new products. Our approach synthesizes diverse data modalities including product images, text, attributes, temporal signals and contextual data to capture the complex dynamics between products, consumers, and markets. The integration of transformers and diffusion modeling enables M2TFM to dynamically focus on salient features across data types and employ robust temporal modeling to capture intricate sales patterns. We have demonstrated the advantages of M2TFM through extensive experiments on a large-scale

real world dataset, and the results show models using only a single factor like text, images or time series performed worse compared to M2TFM, which leverages the combination of "Image Product Attribute Text (T), Product Images (I), Products Attribute Time Series (A), text description (C) and Exogenous Attributes Time Series (E)" to achieve better results. The empirical results show the robustness and predictive prowess of M2TFM in the context of new product sales forecasting. The study not only validates the design of the M2TFM but also illuminates the significance of leveraging a multi-modal approach to data analysis in sales prediction tasks. Moreover, we highlight the following key advantages of the proposed method for practical applications:

- Impact of multimodal feature fusion on the practical application of new product sales forecasting problem
- Practical application scenarios of our proposed model to the new product sales forecasting problem
- Potential problems in the practical application of our proposed model

CRediT authorship contribution statement

Xiangzhen Li: Writing – original draft, Methodology, Data curation, Conceptualization. **Jiaxing Shen:** Methodology. **Dezhi Wang:** Methodology, Data curation. **Wu Lu:** Resources, Methodology, Funding acquisition. **Yuanyi Chen:** Writing – review & editing, Resources, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The experiments presented in this paper were conducted using a publicly available dataset : https://paperswithcode.com/sota/new-product-sales-forecasting-on-visuelle.

Acknowledgments

This work is supported by the Foundation of State Key Laboratory of Public Big Data (No. PBD2021-10), the Institute of Digital Finance at Hangzhou City University, the Science and Technology Department of Zhejiang Province Project (No. 2022C35043), the Top Young Scholar Program of Zhejiang Province "Ten Thousands Talent Program", and the Top Young Scholar Program of Hangzhou "Ten Thousands Talent Program".

References

Chen, G., Huang, L., Xiao, S., Zhang, C., Zhao, H., 2023. Attending to customer attention: A novel deep learning method for leveraging multimodal online reviews to enhance sales prediction. Inf. Syst. Res..

Chu, Z., Wang, C., Chen, C., Cheng, D., Liang, Y., Qian, W., 2023. Learning invariant representations for new product sales forecasting via multi-granularity adversarial learning. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 3828–3832.

Deng, T., Zhao, Y., Wang, S., Yu, H., 2021. Sales forecasting based on LightGBM. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering. ICCECE, IEEE, pp. 383–386.

Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S.S.K., Dwivedi, S., Raykar, V., 2020. Attention based multi-modal new product sales time-series forecasting. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3110–3118.

Giri, C., Chen, Y., 2022. Deep learning for demand forecasting in the fashion and apparel retail industry. Forecasting 4 (2), 565–581.

Giri, C., Thomassey, S., Balkow, J., Zeng, X., 2019. Forecasting new apparel sales using deep learning and nonlinear neural network regression. In: 2019 International Conference on Engineering, Science, and Industrial Applications. ICESI, IEEE, pp. 1–6.

Gustriansyah, R., Suhandi, N., Antony, F., Sanmorino, A., 2019. Single exponential smoothing method to predict sales multiple products. In: Journal of Physics: Conference Series. Vol. 1175, IOP Publishing, 012036.

He, Y.-L., Ou, G.-L., Fournier-Viger, P., Huang, J.Z., Suganthan, P.N., 2022a. A novel dependency-oriented mixed-attribute data classification method. Expert Syst. Appl. 199, 116782.

He, Y.-L., Xu, S.-S., Huang, J.Z., 2022b. Creating synthetic minority class samples based on autoencoder extreme learning machine. Pattern Recognit. 121, 108191.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Jha, B.K., Pande, S., 2021. Time series forecasting model for supermarket sales using FB-prophet. In: 2021 5th International Conference on Computing Methodologies and Communication. ICCMC, IEEE, pp. 547–554.

Karb, T., Kühl, N., Hirt, R., Glivici-Cotruta, V., 2020. A network-based transfer learning approach to improve sales forecasting of new products. arXiv preprint arXiv:2005.06978.

Kohli, S., Godwin, G.T., Urolagin, S., 2020. Sales prediction using linear and KNN regression. In: Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Springer, pp. 321–329.

Krishnamoorthy, N., Prasad, L.N., Kumar, C.P., Subedi, B., Abraha, H.B., Sathishkumar, V., 2021. Rice leaf diseases prediction using deep neural networks with transfer learning. Environ. Res. 198, 111275.

Li, D., Li, X., Lin, K., Liao, J., Du, R., Lu, W., Madden, A., 2023. A multiple long short-term model for product sales forecasting based on stage future vision with prior knowledge. Inform. Sci. 625, 97–124.

Li, J., Li, D., Xiong, C., Hoi, S., 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR, pp. 12888–12900.

Li, D., Lin, K., Li, X., Liao, J., Du, R., Chen, D., Madden, A., 2022a. Improved sales time series predictions using deep neural networks with spatiotemporal dynamic pattern acquisition mechanism. Inf. Process. Manage. 59 (4), 102987.

Li, Y., Yang, Y., Zhu, K., Zhang, J., 2021. Clothing sale forecasting by a composite GRU–prophet model with an attention mechanism. IEEE Trans. Ind. Inform. 17 (12), 8335–8344.

Ma, S., Fildes, R., 2021. Retail sales forecasting with meta-learning. European J. Oper. Res. 288 (1), 111–128.

Manikandan, K., Saranya, A., Deetshiha, D.J., Sushmitha, K., Dharani, D., Anitha, V., Kalaiselvi, S., 2022. Intelligent sales prediction using ARIMA techniques. In: AIP Conference Proceedings. Vol. 2444, AIP Publishing.

Nanda, A., Mohapatra, B.B., Mahapatra, A.P.K., Mahapatra, A.P.K., Mahapatra, A.P.K., 2021. Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. Int. J. Stat. Appl. Math. 6 (1), 59–65.

Oliveira, J.M., Ramos, P., 2023. Cross-learning-based sales forecasting using deep learning via partial pooling from multi-level data. In: International Conference on Engineering Applications of Neural Networks. Springer, pp. 279–290.

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., Kompatsiaris, I., 2022. Multimodal quasi-AutoRegression: Forecasting the visual popularity of new fashion products. Int. J. Multimed. Inf. Retr. 11 (4), 717–729.

Puspita, P.E., İnkaya, T., Akansel, M., 2019. Clustering-based sales forecasting in a forklift distributor. Int. J. Eng. Res. Dev. 11 (1), 25–40.

Roy, D., Li, Y., Jian, T., Tian, P., Chowdhury, K.R., Ioannidis, S., 2022. Multi-modality sensing and data fusion for multi-vehicle detection. IEEE Trans. Multimed..

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Shilong, Z., et al., 2021. Machine learning model for sales forecasting by using xgboost. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering. ICCECE, IEEE, pp. 480–483.

Singh, B., Kumar, P., Sharma, N., Sharma, K., 2020. Sales forecast for amazon sales with time series modeling. In: 2020 First International Conference on Power, Control and Computing Technologies. ICPC2T, IEEE, pp. 38–43.

Skenderi, G., Joppi, C., Denitto, M., Cristani, M., 2021. Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. arXiv preprint arXiv:2109.09824.

Skenderi, G., Joppi, C., Denitto, M., Scarpa, B., Cristani, M., 2022. The multi-modal universe of fast-fashion: the visuelle 2.0 benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2241–2246.

Stoker, P., Tian, G., Kim, J.Y., 2020. Analysis of variance (ANOVA). In: Basic Quantitative Research Methods for Urban Planners. Routledge, pp. 197–219.

Vashishtha, R.K., Burman, V., Kumar, R., Sethuraman, S., Sekar, A.R., Ramanan, S., 2020. Product age based demand forecast model for fashion retail. arXiv preprint arXiv:2007.05278.

Wei, H., Zeng, Q., 2021. Research on sales forecast based on xgboost-LSTM algorithm model. In: Journal of Physics: Conference Series. Vol. 1754, IOP Publishing, 012191.

Wolters, J., Huchzermeier, A., 2021. Joint in-season and out-of-season promotion demand forecasting in a retail environment. J. Retail. 97 (4), 726–745.

Xue, Z., Marculescu, R., 2023. Dynamic multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2574–2583.
Yan, X., Hu, H., 2023. New product demand forecasting and production capacity adjustment strategies: Within-product and cross-product word-of-mouth. Comput. Ind. Eng. 109394.
Yin, P., Dou, G., Lin, X., Liu, L., 2020. A hybrid method for forecasting new product sales based on fuzzy clustering and deep learning. Kybernetes 49 (12), 3099–3118.