Journal of Organizational and End User Computing
Volume 36 • Issue 1

Application of Computer Vision on E-Commerce Platforms and Its Impact on Sales Forecasting

Wei-Dong Liu, Dongbei University of Finance and Economics, China*
Xi-Shui She, Fengjia University, China

ABSTRACT

In today's digital age, the e-commerce industry continues to grow and flourish. The widespread application of computer vision technology has brought revolutionary changes to e-commerce platforms. Extracting image features from e-commerce platforms using deep learning techniques is of paramount importance for predicting product sales. Deep learning-based computer vision models can automatically learn image features without the need for manual feature extractors. By employing deep learning techniques, key features such as color, shape, and texture can be effectively extracted from product images, providing more representative and diverse data for sales prediction models. This study proposes the use of ResNet-101 as an image feature extractor, enabling the automatic learning of rich visual features to provide high-quality image representations for subsequent analysis. Furthermore, a bidirectional attention mechanism is introduced to dynamically capture correlations between different modalities, facilitating the fusion of multimodal features.

KEYWORDS

Bidirection Attention Mechanism, BiLSTM, Computer Vision, E-commerce Platform, ResNet-101

1. INDRODUCTION

In today's digital era, the e-commerce industry is continuously expanding (Wang & Chang, 2021), and the widespread application of computer vision technology has brought revolutionary changes to e-commerce platforms (Yang & Liu, 2021). Computer vision is a technology that enables machines to understand and interpret images or videos (Sheela, 2022). Its applications are not limited to the fields of art (Manovich, 2021) and entertainment (Erdelyi, 2019); it also plays a crucial role in the business world (Soni et al., 2020). The application of computer vision technology in e-commerce provides users with a more convenient and intelligent shopping experience. Through image recognition techniques (Mehmood et al., 2019), consumers can use visual searches to find specific products without the need for text descriptions. This intuitive search method not only enhances user satisfaction but also

DOI: 10.4018/JOEUC.336848 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

accelerates the shopping decision-making process, leading to an increase in sales. Furthermore, by analyzing users' shopping history, browsing habits, and preferences, computer vision systems can recommend products that users might be interested in (Zhou, 2020). This personalized recommendation not only increases the likelihood of user purchases but also strengthens the bond between users and e-commerce platforms, encouraging users to visit the platform more frequently. Additionally, computer vision technology can be used to enhance the sales forecasting capabilities of e-commerce platforms (Jain & Wah, 2022). By analyzing product images and videos, the system can identify product features and predict demand trends. This precise sales forecasting helps e-commerce platforms better manage inventory, avoiding situations of surplus or stockouts, thus improving the efficiency of the supply chain and reducing inventory costs. The application of computer vision technology in e-commerce not only enhances the user experience but also strengthens the interaction between e-commerce platforms and users. Simultaneously, it improves sales forecasting and inventory management efficiency. With the continuous innovation and development of computer vision technology, it will continue to bring more opportunities and challenges to the e-commerce industry, driving the entire industry towards a more intelligent and efficient direction.

The extraction of image features from e-commerce platforms using deep learning technology (De la Comble et al., 2022) holds significant importance for predicting product sales. Deep learning models, particularly Convolutional Neural Networks (CNNs) (Liu, 2022), have the capability to automatically learn features from images without the need for manual feature extractors. Consequently, deep learning technology efficiently extracts critical features, including color, shape, texture, among others, from product images, providing more representative and enriched data for sales prediction models. Deep learning models typically exhibit strong generalization abilities when dealing with large-scale data, enabling them to adapt to various product types and market conditions. Thus, models based on deep learning technology can more accurately predict sales, extending beyond specific categories or time periods. Furthermore, product sales are influenced by a multitude of factors (Cheng et al., 2019), often characterized by complex nonlinear relationships. Deep learning models can learn intricate features and relationships within the data, better capturing the complex associations between product sales and image features, thus enhancing prediction accuracy. In addition, product images on e-commerce platforms often exhibit variations in size, angle, lighting, and other characteristics. Traditional feature extraction methods may struggle to handle this diversity. Deep learning models, on the other hand, effectively manage varying image data by using operations such as convolutions, enhancing the model's adaptability to changes in image characteristics. In summary, the extraction of image features from e-commerce platforms using deep learning technology is highly significant for product sales prediction. It not only provides more accurate and diverse feature representations but also handles complex relationships and adapts to diverse image data. It enables the development of end-to-end prediction systems, providing e-commerce businesses with a more precise and efficient tool for sales prediction, aiding in the formulation of market strategies, inventory management, and enhancing sales performance. The deep learning models commonly used in the research on e-commerce platform sales forecasting are as follows:

Convolutional Neural Networks (CNN) (C. Zhang et al., 2021): CNNs are specifically designed deep learning models for handling image data. They can automatically learn features within images (Bhoir & Patil, 2022), eliminating the need for manual feature extractors. In the context of sales forecasting for e-commerce platforms, CNNs can extract features such as color, shape, and texture from product images, providing more representative and rich data. This capability enhances the accuracy of the models.

Recurrent Neural Networks (RNN): RNNs are suitable for processing sequential data and can capture trends in product sales over time (Liu et al., 2021). Considering the seasonal and periodic patterns in sales, particularly in e-commerce, RNNs are adept at capturing these regularities, thereby improving the accuracy of sales forecasts.(Shetty & Buktar, 2022)

Long Short-Term Memory Networks (LSTM): LSTM is a specialized type of RNN equipped with memory units, enabling it to handle long-term dependencies. In sales forecasting, LSTM can capture long-term trends in product sales data and exhibit excellent adaptability to irregular or unstable sales patterns.(Efat et al., 2022)

Transformer Models (Self-Attention Models): Transformer models introduce self-attention mechanisms, allowing them to effectively capture dependencies between different time steps in sequential data. This feature is applicable to nonlinear relationships commonly found in e-commerce sales data. Additionally, their parallel computing capabilities enhance efficiency when training on large-scale datasets. (E. Ning et al., 2024)

Generative Adversarial Networks (GAN): GANs consist of a generator and a discriminator and can generate realistic images. In the context of e-commerce platforms, GANs can be used to generate synthetic data, expanding sales datasets and aiding models in learning features more effectively. Moreover, GANs can generate diverse product images, providing varied training data. (Thivakaran & Ramesh, 2022)

The design philosophy of this study's model is rooted in advanced deep learning techniques, aiming to leverage both images and multimodal data to enhance the accuracy and robustness of sales forecasting on e-commerce platforms. Firstly, ResNet (Durga & Rajesh, 2022) is employed as the image feature extractor, automatically learning rich visual features, and providing high-quality image representations for subsequent analysis. Next, the bidirectional attention mechanism (Li et al., 2022) is introduced to dynamically capture correlations among different modalities, facilitating the fusion of multimodal features (Li et al., 2021). This approach enables the model to comprehensively utilize diverse information sources such as text and images, enhancing the model's ability to model complex and varied e-commerce data. Finally, BiLSTM (Singla et al., 2022) serves as the core model for sales forecasting, effectively capturing temporal dependencies and trends in sales data, enabling more accurate predictions of future sales trends.

This model design demonstrates innovativeness in the following three aspects:

1. This model introduces a bidirectional attention mechanism, facilitating dynamic fusion of multimodal data, including images and text. In contrast to traditional unidirectional fusion methods, the bidirectional attention mechanism simultaneously considers bidirectional relationships among different modalities, enabling a more accurate capture of complex dependencies between various data sources. This innovative data fusion approach enhances the model's ability to comprehensively utilize diverse information sources, offering a fresh perspective for e-commerce sales forecasting tasks.
2. The model seamlessly integrates deep learning features extracted from ResNet with BiLSTM, achieving fusion of platform image features through deep learning and temporal modeling. This dual fusion of deep learning and time series modeling enables the model to accurately capture abstract representations of image features while effectively addressing the temporal dependencies in sales data. Consequently, it enhances the accuracy and robustness of sales forecasting.
3. This model not only conducts overall sales predictions but also achieves fine-grained sales trend prediction. Through BiLSTM's time series modeling, the model captures detailed features in sales data such as seasonality and periodicity. This capability for fine-grained prediction provides e-commerce platforms with more targeted sales strategy formulation and inventory management recommendations.

In the rest of this paper, we will introduce the recently related work in section 2. Section 3 presents the proposed methods: overview, Image feature extraction layer based on ResNet-101, multimodal feature fusion based on bidirectional attention mechanism, platform sales forecasting based on BiLSTM. Section 4 introduces the experimental part, including practical details, comparative experiments, and an ablation study. Section 5 includes a conclusion and an outlook.

2. RELATED WORK

2.1 ResNet Model

ResNet (Residual Network), as a significant innovation in the field of deep learning, exhibits remarkable advantages in the task of extracting image features for e-commerce platforms (Kumar et al., 2023). Firstly, its deep network architecture enables the network to learn image features in greater depth, allowing it to handle diverse and complex product images with precision, capturing visual attributes such as color, shape, and texture more accurately. Secondly, the introduction of residual connections in ResNet ensures efficient information transfer within the network, effectively alleviating the vanishing gradient problem (Yang et al., 2021). This feature enables the training of exceptionally deep networks, preserving and extracting key features from images effectively. This mechanism is particularly vital for e-commerce product images, which often contain intricate details and multi-level features(X. Ning et al., 2024).

Additionally, ResNet possesses outstanding feature extraction capabilities, allowing it to extract features from low-level edges and textures to high-level object parts and holistic features, providing a hierarchical feature representation. This multi-level feature extraction is crucial for e-commerce platforms as different types of products require capturing features at various levels. ResNet's ability to perform multi-level feature extraction enables it to adapt to a wide array of product types, including clothing (Li et al., 2023), electronics (Chen et al., 2023), food (Senapati et al., 2023), and cosmetics (Abdullah & Dawood, 2023), providing accurate feature descriptions, and serving as a reliable foundation for product recognition and recommendation.

Furthermore, ResNet supports transfer learning, allowing pre-trained ResNet models to be fine-tuned for different tasks within e-commerce platforms. This flexibility avoids the time and resource-intensive process of training models from scratch, making ResNet a highly efficient and convenient choice for e-commerce applications. It serves as a powerful image processing tool, enhancing the quality of product recommendation, search, and overall shopping experiences. In summary, ResNet, with its depth, residual connections, and multi-level feature extraction capabilities, stands out as an ideal choice for processing large-scale product images, ushering in new opportunities and possibilities for the development of the e-commerce industry.

2.2 Bidirectional Attention Mechanism

The bidirectional attention mechanism exhibits unique advantages and challenges in the fusion of multimodal data on e-commerce platforms (Y. Zhang et al., 2021). Firstly, its advantage lies in simultaneously considering the bidirectional relationships among different modalities, including images, texts, etc., achieving dynamic fusion of data (Qin, Zhang & You, 2022). This bidirectional attention mechanism enables the model to capture the intricate dependencies more accurately between various data sources, enhancing the model's ability to synthesize information from multiple sources effectively. Particularly in the context of e-commerce platforms, where product information typically includes diverse modalities such as images and textual descriptions, the bidirectional attention mechanism adeptly integrates these modalities, providing a more comprehensive and accurate feature representation, thereby offering reliable support for tasks like product recommendation and search.

However, the bidirectional attention mechanism also presents certain limitations. Firstly, it entails high computational complexity, especially when dealing with large-scale multimodal data, demanding substantial computational resources. Secondly, the model involves a substantial number of parameters, necessitating ample data for training to prevent overfitting issues. Additionally, when handling long-range dependencies between different modalities, the bidirectional attention mechanism may encounter challenges related to accurate information transmission, particularly in the presence of significant disparities or noise between modalities, making it susceptible to interference.

In summary, the bidirectional attention mechanism demonstrates superior information capturing capabilities in the fusion of multimodal data on e-commerce platforms. Nevertheless, further

research and improvements are needed, particularly in terms of computational efficiency and model generalization, to effectively address the challenges encountered in practical applications.

2.3 BiLSTM

The BiLSTM (Bidirectional Long Short-Term Memory) model exhibits distinctive advantages and challenges in predicting sales based on multimodal data features on e-commerce platforms (Zhang & Kim, 2023). Firstly, its strength lies in its ability to handle sequential data, such as sales time series and product description texts, while capturing temporal relationships within the data. This is crucial for predicting changes in sales trends over time. Secondly, BiLSTM incorporates bidirectional memory units, enabling it to capture long-term dependencies effectively, making it suitable for deciphering complex patterns and trends within sales data. Particularly in e-commerce settings, where sales data can be influenced by factors like seasonality and promotional activities, BiLSTM comprehensively understands these influencing factors, enhancing the accuracy of predictions. Furthermore, BiLSTM can handle feature extraction from different modalities, such as images and textual data, providing a comprehensive approach to sales forecasting by learning intricate relationships among multimodal features.

However, the BiLSTM model also presents certain limitations. Firstly, its computational complexity is relatively high, especially when dealing with large-scale multimodal data, requiring substantial computational resources and time. Secondly, BiLSTM is susceptible to challenges posed by excessively long sequential data; when sequences are too lengthy, the model might face issues related to vanishing or exploding gradients, impacting training effectiveness. Additionally, the model has a substantial number of parameters and demands extensive data for training to prevent overfitting, which might limit its applicability in certain e-commerce scenarios.

The BiLSTM model demonstrates excellent capabilities in temporal modeling and learning from multimodal features in sales prediction on e-commerce platforms. However, challenges related to computational efficiency and handling excessively long sequences need to be addressed. To fully leverage its advantages, it is essential to align the model structure and optimization strategies with specific scenarios, ensuring more precise and efficient sales forecasts.

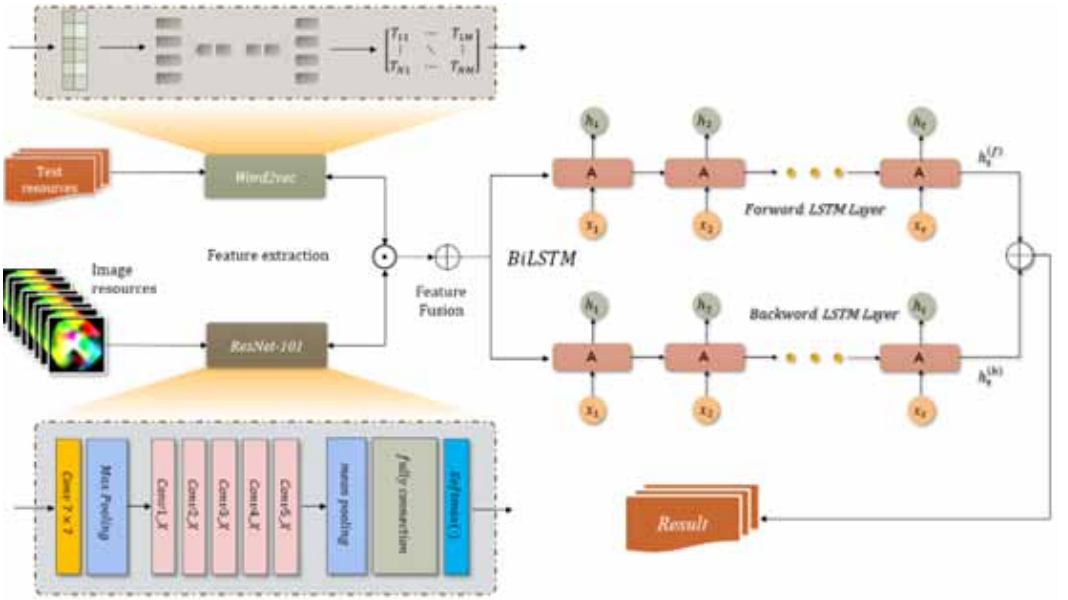
3. METHOD

3.1 Overview

Our research model has developed an innovative and multifaceted deep learning framework tailored to address the complexities of sales forecasting in e-commerce platforms. This model amalgamates several advanced neural network techniques, endowing it with robust capabilities in temporal modeling and multimodal feature learning. At its core, the model incorporates Bidirectional Long Short-Term Memory networks (BiLSTM), renowned for their excellence in handling sequential data, enabling precise capturing of the evolving trends in sales data. To harness the richness of product information, we introduced Residual Networks (ResNet), an efficient feature extraction module adept at extracting intricate visual features from product images, including colors, shapes, and textures. Incorporating a Bidirectional Attention Mechanism enhances the fusion of textual and visual information, enabling the model to dynamically focus on essential features across different modalities. This meticulous data integration approach surpasses mere sales predictions, delving deep into the intrinsic patterns and features within sales data through sophisticated deep learning methodologies. An overview of our framework is shown in the figure 1.

The schematic diagram of the model illustrates the overall architecture for utilizing advanced deep learning techniques in e-commerce platform sales forecasting. Key components of the model include ResNet, serving as an image feature extractor that autonomously learns intricate visual features. The bidirectional attention mechanism dynamically captures correlations among different

Figure 1. Overview of our framework



modalities, promoting the fusion of multimodal features. This integrated approach allows the model to comprehensively harness diverse information sources, such as text and images, enhancing its ability to model the complexity of e-commerce data. Finally, the core predictive model, BiLSTM, effectively captures temporal dependencies and trends in sales data, contributing to a more accurate prediction of future sales trends.

The model does not merely offer simple sales forecasts; it unearths underlying patterns and correlations between diverse product features, thereby providing comprehensive and accurate predictions. It transcends the temporal dimension, meticulously considering the visual attributes and textual descriptions of products, thereby ensuring a holistic and precise sales forecasting process.

3.2 Image Feature Extraction Based on ResNet-101

In this study, ResNet-101 network (Singh & Kumar, 2022) was utilized for image feature extraction during image attribute generation. Unlike the commonly used ResNet-50 network (Wu et al., 2023), ResNet-101 has a deeper architecture. This increased depth ensures the extraction of highly diverse features while mitigating issues related to gradient vanishing or exploding, thereby facilitating subsequent label prediction. The structure of ResNet-101 is outlined in Table 1, comprising five convolutional layers: conv1, conv2_x, conv3_x, conv4_x, and conv5_x.

The structure of ResNet-101 is shown in the figure 2.

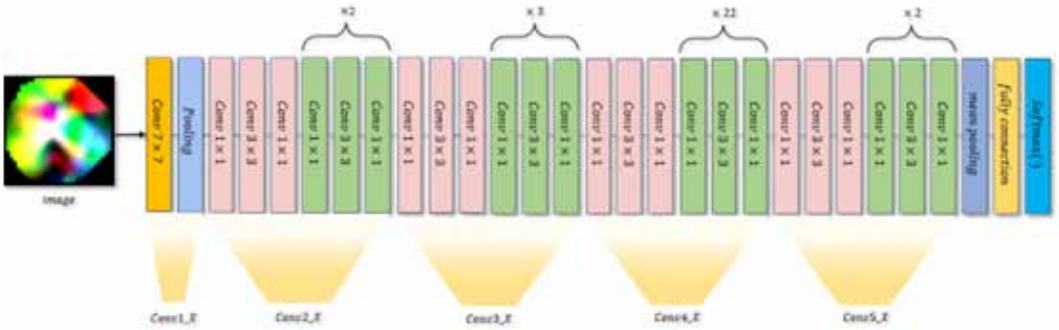
3.3 Multimodal Data Feature Fusion for Online Goods Based on Bidirection Attention Mechanism

Step 1: Representation of Multimodal Data. For image data, the ResNet-101 model is employed to extract image features, denoted as $I = \left\{ v_i^I \right\}$, where v_i^I represents the feature vector of the i th image. Regarding textual data, word embeddings are utilized to convert text into sequences of word vectors, represented as $T = \left\{ v_i^T \right\}$, where v_i^T denotes the word vector of the i th word.

Table 1. The structure of ResNet-101

Layer	Framework	Output
Conv1	7*7,64,2	112*112
Conv2_x	3*3,2 $\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times3$	56*56
Conv3_x	$\begin{bmatrix} 1\times1,128 \\ 3\times3,128 \\ 1\times1,512 \end{bmatrix}\times4$	28*28
Conv4_x	$\begin{bmatrix} 1\times1,256 \\ 3\times3,256 \\ 1\times1,1024 \end{bmatrix}\times23$	14*14
Conv5_x	$\begin{bmatrix} 1\times1,512 \\ 3\times3,512 \\ 1\times1,2048 \end{bmatrix}\times3$	7*7
classification layer	Mean pooling, fully connected, Softmax classification	1*1

Figure 2. The structure of ResNet-101



Step 2: Calculation of Bidirectional Attention Weights. In this step, the attention weights from text to image $A^{T\to I}=\left\{\alpha_{i,j}^{T\to I}\right\}$ and from image to text $A^{I\to T}=\left\{\alpha_{i,j}^{I\to T}\right\}$ are computed. These attention weights indicate the focus of the text on the image and vice versa. The computation is performed as follows:

Attention Weights from Text to Image:

$$\alpha_{i,j}^{T\to I} = \frac{\exp\left(e_{i,j}^{T\to I}\right)}{\sum_{k=1}^{N_I}\exp\left(e_{i,k}^{T\to I}\right)} \tag{1}$$

$e_{i,j}^{T\to I} = v_i^T \cdot v_j^I$ represents the inner product between the i th word in the text description and the j th element of the image feature, indicating the correlation between text and image.

Attention Weights from Image to Text:

$$\alpha_{i,j}^{I\to T} = \frac{\exp\left(e_{i,j}^{I\to T}\right)}{\sum_{k=1}^{N_T}\exp\left(e_{i,k}^{I\to T}\right)} \tag{2}$$

$e_{i,j}^{I\to T} = v_i^I \cdot v_j^T$ represents the inner product between the i th element of the image feature and the j th word in the text description, indicating the correlation between image and text.

These attention weights enable the model to focus on different parts of each modality during the fusion process.

Step 3: Feature Fusion. Lastly, utilizing the calculated attention weights, the image features and textual features are linearly combined to obtain the fused feature vector:

$$F_{\text{fusion}} = \sum_{i=1}^{N_T}\sum_{j=1}^{N_I}\left(\alpha_{i,j}^{T\to I}\cdot v_i^T + \alpha_{i,j}^{I\to T}\cdot v_j^I\right) \tag{3}$$

The fused feature vector F_{fusion} contains essential information from both text and image, which can be utilized in subsequent tasks such as product recommendations and searches. The significance of this step lies in the selective fusion of information from the two modalities, enhancing the model's ability to represent multimodal data effectively. The structure of bidirection attention mechanism is shown in the figure 3.

3.4 E-Commerce Sales Prediction Based on BiLSTM

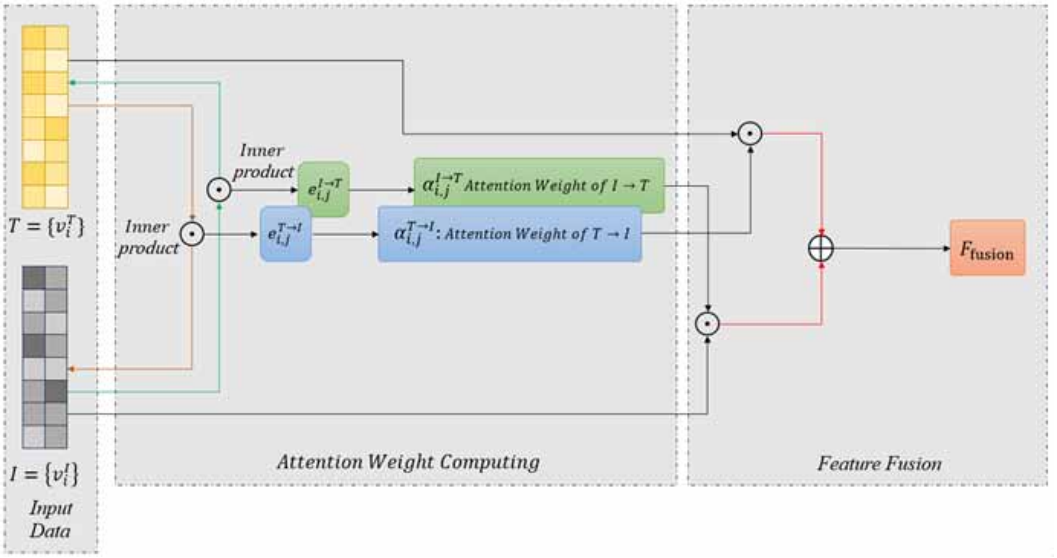
This section describes the basic steps of E-commerce sales prediction based on BiLSTM. Among them, Step 1,2, and 3 describes the model forward computation steps for both forward and backward LSTM(ZHENG Yuxiang, 2021), and Step 4 describes the principle of BiLSTM model to integrate the LSTM model in both directions.

Step 1: Forget Gate. In this step, the forget gate determines which information from the previous time step's memory C_{t-1} needs to be discarded. The computation of the forget gate is given by:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}^{(f)}, x_t\right] + b_f\right) \tag{4}$$

Here, f_t is the output of the forget gate, σ represents the sigmoid activation function, W_f is the weight matrix for the forget gate, $h_{t-1}^{(f)}$ is the forward LSTM's hidden state at time $t-1$, and x_t is the input feature at the current time step. The forget gate selectively retains or discards information from the previous memory cell, allowing the model to remember relevant historical information.

Figure 3. The structure of bidirection attention mechanism



Step 2: Input Gate. The input gate decides which information from the current time step will be added to the memory. The computation of the input gate is given by:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}^{(i)}, x_t\right] + b_i\right) \tag{5}$$

Here, i_t is the output of the input gate, W_i is the weight matrix for the input gate, $h_{t-1}^{(i)}$ is the forward LSTM's hidden state at time $t-1$, and x_t is the input feature at the current time step. The input gate selectively updates the content of the memory cell, allowing it to capture important information from the current time step.

Step 3: Output Gate. The output gate determines the output of the LSTM cell at the current time step. The computation of the output gate is given by:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}^{(o)}, x_t\right] + b_o\right) \tag{6}$$

Here, o_t is the output of the output gate, W_o is the weight matrix for the output gate, $h_{t-1}^{(o)}$ is the forward LSTM's hidden state at time $t-1$, and x_t is the input feature at the current time step. The output gate controls which portion of the memory cell's content will be output, generating the final output at the current time step.

Step 4: Integration of Forward and Backward LSTMs. In a BiLSTM model, the outputs of the forward and backward LSTMs are concatenated to form the comprehensive output at the current time step:

$$h_t = \left[h_t^{(f)}, h_t^{(b)}\right] \tag{7}$$

Here, $h_t^{(f)}$ represents the forward LSTM's hidden state at time t, and $h_t^{(b)}$ represents the backward LSTM's hidden state at time t .This step integrates information from both the forward and backward directions, creating a more holistic and accurate representation of the sequential information. This integrated representation offers richer features for predicting sales volumes. Algorithm 1 shows the pseudo-code representation of E-commerce sales forecasting based on BiLSTM in PyTorch-like style.

The structure of BiLSTM is shown in the figure 4.

4. EXPERIMENT

4.1 Experimental Design

In this study, we conducted three simulation experiments: a comparative experiment with multiple methods on a single dataset, another comparative experiment with multiple methods on a single dataset, and a model ablation experiment. In the comparative experiments with multiple methods, we selected 6 baseline algorithms and models, including traditional machine learning methods and deep learning models, for performance comparison with the proposed model. In the ablation experiment, we conducted a comprehensive analysis of the proposed model, including the removal and addition of model components, to validate their impact on overall performance.

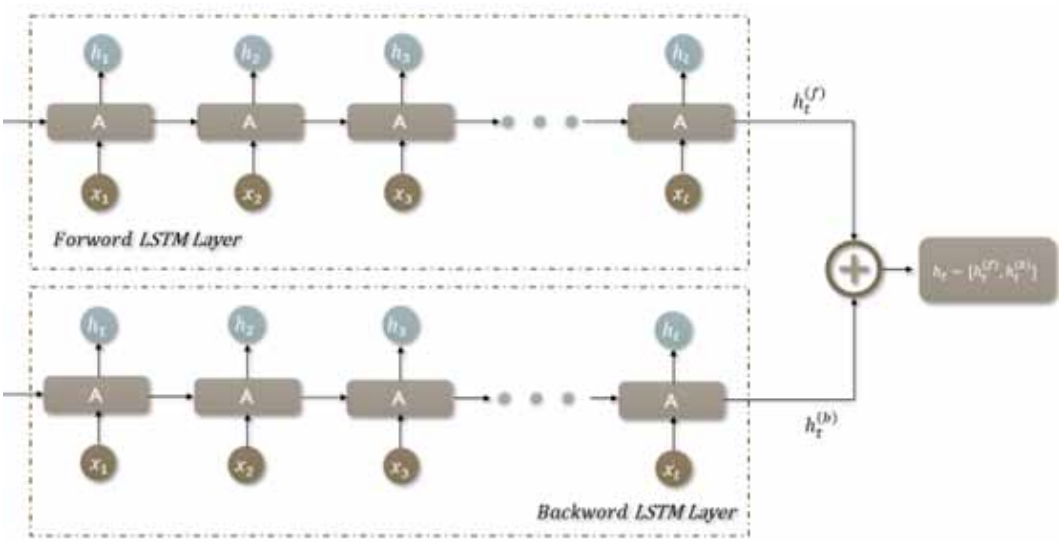
All experiments were conducted on a high-performance workstation equipped with an Intel Core i9 processor, 64GB of RAM, and an NVIDIA GeForce RTX 3090 graphics card. We utilized the PyTorch deep learning framework for model training and evaluation.

The model parameters were set as follows: initial learning rate: 0.001, batch size: 32, number of training epochs: 100, activation function: ReLU, loss function: Mean Squared Error (MSE). The number of cells in the BiLSTM model: 30.

Algorithm 1. pseudo-code representation of E-commerce sales forecasting based on BiLSTM(In PyTorch-like Style)

```
# Initialize the parameters of the forward LSTM and the backward LSTM
Initial input_size, hidden_size;
forward_hidden_state = torch.zeros(hidden_size)
backward_hidden_state = torch.zeros(hidden_size)
# Initialize weights and biases for forward and backward LSTMs
forward_weights = rand(hidden_size, input_size + hidden_size)
backward_weights = rand(hidden_size, input_size + hidden_size)
forward_bias = rand(hidden_size)
backward_bias = rand(hidden_size)
time_steps = len(inputs[ ])
outputs = [ ]
for t in range(time_steps):
 # Forward LSTM computation
 forward_concatenated = torch.cat((inputs[t], forward_hidden_state), dim=0)
 forward_gate_input = torch.matmul(forward_weights, forward_concatenated) + forward_bias
 forward_gate_output = sigmoid(forward_gate_input)
 forward_hidden_state = forward_gate_output * tanh(forward_gate_input)
 # Backward LSTM computation
 backward_concatenated = torch.cat((inputs[time_steps - t - 1], backward_hidden_state), dim=0)
 backward_gate_input = torch.matmul(backward_weights, backward_concatenated) + backward_bias
 backward_gate_output = sigmoid(backward_gate_input)
 backward_hidden_state = backward_gate_output * tanh(backward_gate_input)
 # Connect the outputs of the forward and backward LSTMs to form the output of the BiLSTM
 concatenated_output = torch.cat((forward_hidden_state, backward_hidden_state), dim=0)
 outputs.append(concatenated_output.unsqueeze(0))
return torch.cat(outputs, dim=0)
```

Figure 4. The structure of BiLSTM



We partitioned the dataset into training, validation, and test sets in a ratio of 7:2:1 and performed model training and tuning on the training set. In each experiment, we conducted five repeated trials and averaged the results to ensure the stability of the experimental outcomes. To prevent overfitting, Dropout layers were incorporated into the model, and an early stopping strategy was applied on the validation set to select the best-performing model.

The 6 baseline models are listed below.

1. Literature #1 (Ekambaram et al., 2020) propose a novel attention-based multi-modal encoder-decoder models to forecast the sales for a new product purely based on product images, any available product attributes and also external factors like holidays, events, weather, and discount.
2. Literature #2 (Skenderi et al., 2021) propose a neural network-based approach, where an encoder learns a representation of the exogenous time series, while the decoder forecasts the sales based on the Google Trends encoding and the available visual and metadata information. This model works in a non-autoregressive manner, avoiding the compounding effect of large first-step errors.
3. Literature #3 (Giri & Chen, 2022) proposes an intelligent forecasting system that combines image feature attributes of clothes along with its sales data to predict future demands. This model predicts weekly sales of new fashion apparel by finding its best match in the clusters of products that they created using machine learning clustering based on products' sales profiles and image similarity.
4. Literature #4 (Papadopoulos et al., 2022) propose MuQAR, a Multimodal Quasi-AutoRegressive deep learning architecture that combines two modules: (1) a multimodal multilayer perceptron processing categorical, visual, and textual features of the product and (2) a Quasi-AutoRegressive neural network modelling the "target" time series of the product's attributes along with the "exogenous" time series of all other attributes.
5. Literature #5 (Cai et al., 2021) propose a model that extracts order sequence features, consumer sentiment features and aesthetic features from multimodal data of e-commerce products. Then, a grouping strategy based on Bidirectional Long Short-Term Memory Network (BiLSTM) is proposed.

6. Literature #6 (Xu et al., 2023) propose an A-tiFSR model to extract features from product text and images and design a fine-grained analysis method to mine danmaku data.

In this study, we will utilize five key metrics, including Mean Absolute Error (MAE) (Hodson, 2022), R-squared (R^2) (Rights & Sterba, 2023), Area Under Curve (AUC) (Yang & Ying, 2022), Accuracy, and Root Mean Square Error (RMSE) (Karunasingha, 2022), to comprehensively evaluate the performance of the proposed model against six baseline models. MAE and RMSE will gauge the model's regression accuracy, R-squared will provide insight into the proportion of the target variable's variance explained by the model, while AUC and Accuracy will evaluate the performance of the models in binary classification tasks. The combined use of these metrics will offer a comprehensive evaluation, aiding in determining which model performs best in the task of predicting sales in the e-commerce domain. Throughout the experiments, we will compare these metrics across different models to discern their performance disparities.

4.2 Dataset

The data in this article comes from DeepFashion dataset, iMaterialist Fashion dataset, E-commerce Product Images dataset, Myntra Fashion dataset, Cdiscount's Image Classification Challenge dataset, and JD.com Product Image dataset.

The DeepFashion dataset encompasses over 800,000 images from the fashion domain, spanning 50 different categories, including clothing, footwear, and accessories. Detailed attribute labels such as color, style, and usage are provided for each item. The DeepFashion dataset is employed in this study for extracting image features of fashion products. Analyzing these features enables a more accurate prediction of sales volumes for diverse fashion products. (Liu et al., 2016)

The iMaterialist fashion dataset comprises more than 100,000 images of fashion products, including clothing, footwear, bags, and accessories. Each product is associated with multiple labels describing various attributes. This Dataset is utilized in this study for multi-modal tasks such as fashion product image classification and attribute prediction. It provides diverse multi-modal data, facilitating fine-grained sales predictions.(Guo et al., 2019)

The E-commerce Product images dataset includes hundreds of thousands of product images from different e-commerce platforms, covering categories like apparel, home goods, and electronics. Detailed descriptions accompany each product. This dataset supports tasks such as product image classification and similarity matching. Its diverse perspectives and scales of product information enhance the accuracy of sales volume predictions.(Chaudhuri et al., 2018)

The Myntra Fashion dataset contains over 40,000 images of fashion products, spanning various clothing and accessory categories, sourced from the Myntra online fashion shopping platform. Utilized for fashion product image classification and recommendation, this dataset provides concrete and detailed fashion product data, facilitating feature extraction for sales volume predictions.(Jaiswal et al., 2023)

The Cdiscount's Image Classification Challenge dataset comprises more than 1 million product images categorized into 18,000 different classes, covering diverse product types. Primarily designed for image classification tasks, it provides rich multi-modal product information. It enables in-depth analysis and prediction of sales volumes across various product categories.(Islam & Alauddin, 2018)

JD.com Product Image Dataset offers tens of thousands of product images, including appliances, clothing, and digital products, sourced from the JD.com online shopping platform. Used for image classification and similarity matching, this dataset offers abundant e-commerce product information. It aids in exploring correlated features for sales volume predictions, enhancing prediction accuracy. (Zhao et al., 2023)

4.3 Comparison Study Results and Analysis

4.3.1 Multiple Method Comparison on a Single Dataset

We ran each of the six baseline models and the model proposed in this paper on the JD.com dataset and performed a comparative analysis. The proposed model (Model_7) outperforms the baseline models across multiple metrics, particularly in MAE, and R-squared. This indicates that our model possesses higher prediction accuracy and interpretability compared to other baseline models, making it a reliable and accurate sales forecasting tool for businesses.

Based on the data results presented in the table, we can analyze the performance of six baseline models (Model_1 to Model_6) in comparison with the proposed model (Model_7) across multiple performance metrics. This is shown graphically in figure 5.

Firstly, considering the MAE (Mean Absolute Error), Model_7 exhibits a relatively lower MAE value, indicating its higher accuracy in predictions. This suggests that our model can predict e-commerce sales more accurately compared to the other models, providing businesses with more precise sales forecasts. This is shown graphically in figure 6.

Table 2. Multiple method comparison results in JD.com product image dataset

Model	MAE	R^2	AUC	Accuracy	RMSE
Model_1(Ekambaram et al., 2020)	0.068	0.726	0.703	0.691	0.094
Model_2 (Skenderi et al., 2021)	0.071	0.698	0.69	0.674	0.098
Model_3 (Giri & Chen, 2022)	0.066	0.748	0.72	0.706	0.091
Model_4 (Papadopoulos et al., 2022)	0.072	0.684	0.68	0.662	0.101
Model_5 (Cai et al., 2021)	0.069	0.712	0.697	0.682	0.096
Model_6 (Xu et al., 2023)	0.075	0.652	0.66	0.639	0.105
Model_7 (Our work)	**0.065**	**0.622**	**0.729**	**0.708**	**0.11**

Figure 5. Comparison of the performance of seven models across multiple performance metrics

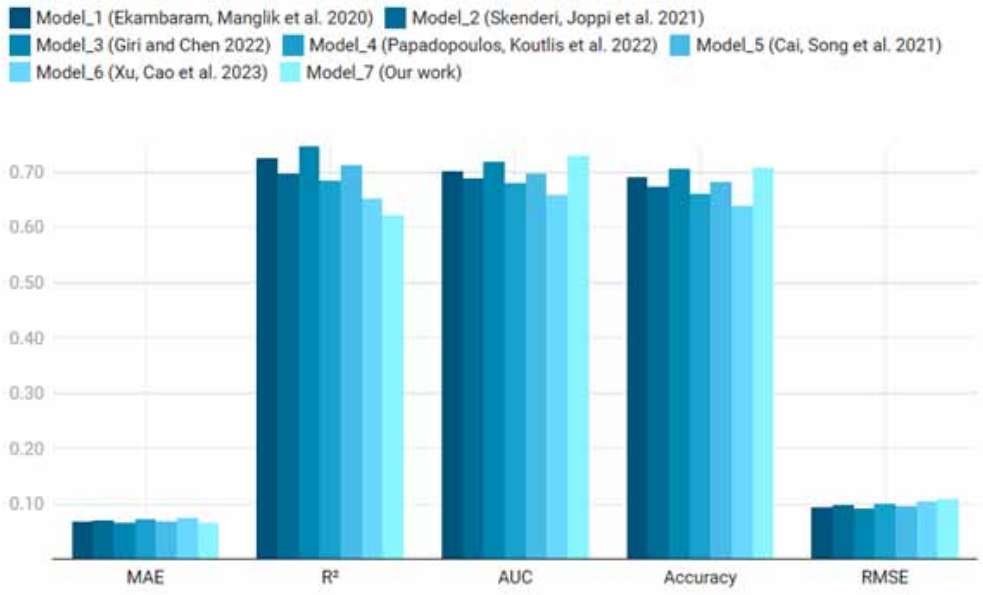


Figure 6. Comparison of the performance of seven models by considering the MAE

Furthermore, examining the R^2 value, Model_7 demonstrates a relatively higher R^2 value, signifying its better ability to explain the variance in the target variable. This implies that our model captures the sales data trends more effectively, offering more reliable predictive results. This is shown graphically in figure 7.

Additionally, Model_7 shows excellent performance in terms of AUC (Area Under the Curve) and accuracy, demonstrating its strong classification capabilities. Finally, concerning the RMSE (Root Mean Square Error) metric, Model_7 exhibits relatively superior performance, indicating excellent control over prediction errors.

In addition, we compared this experiment in two datasets, DeepFashion dataset, iMaterialist Fashion dataset as well, and the comparison results are shown in figure 8.

4.3.2 Comparison of Single Methods on Multiple Datasets

In this study, we conducted a series of single-method multiple-dataset experiments aimed at evaluating the performance of different models across multiple datasets. We selected six representative datasets

Figure 7. Comparison of the performance of seven models by considering the R-squared value

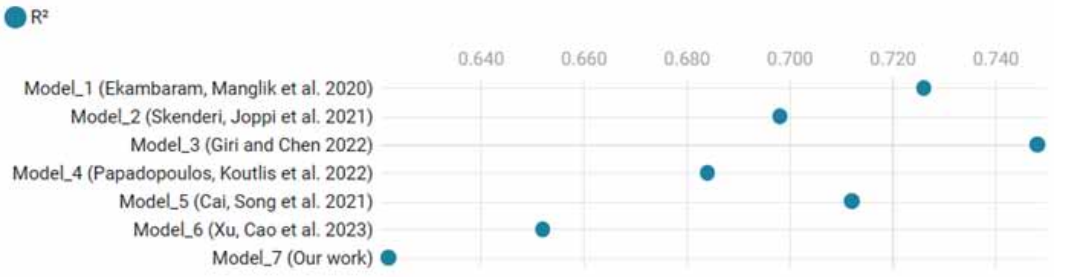
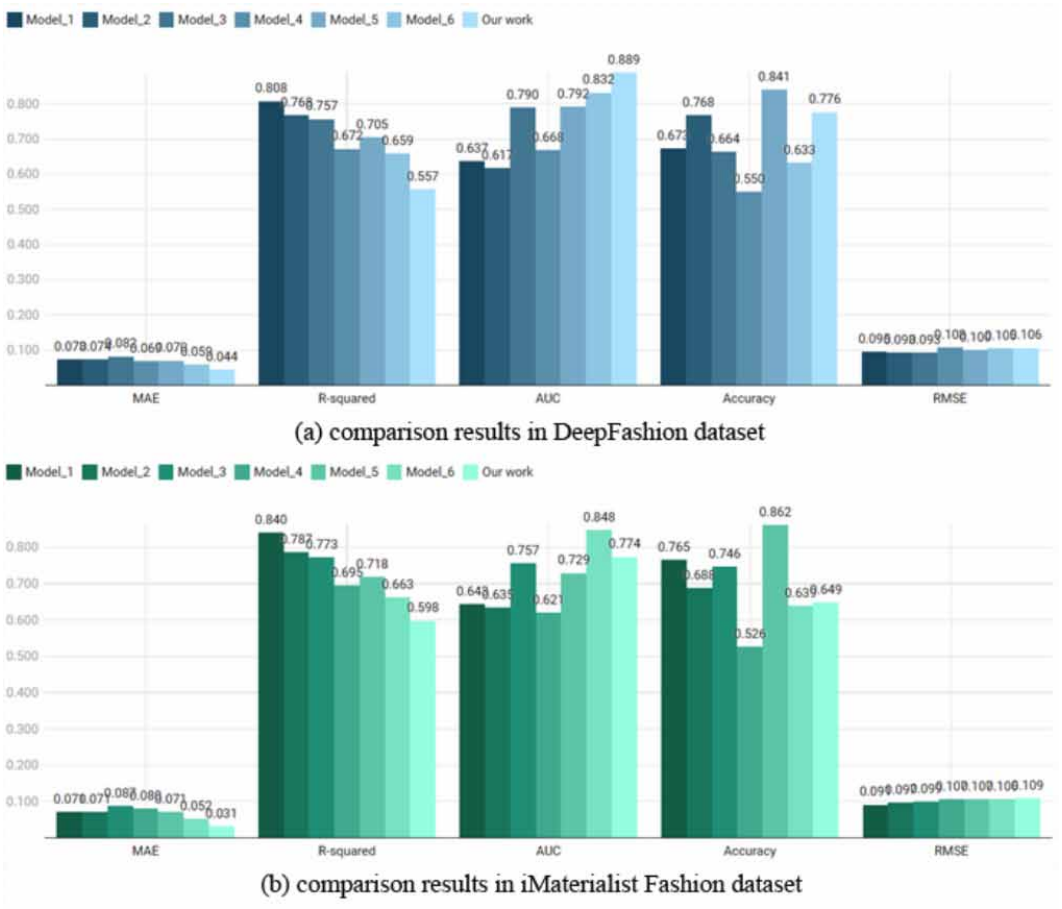


Figure 8. Comparison of the performance of seven models across multiple performance metrics in two other datasets



and employed five metrics, namely MAE, R-squared, AUC, Accuracy, and RMSE, to assess the performance of various models on these datasets.

The experimental results are presented graphically in figure 9.

Through the analysis of experimental results, we found that the proposed model in this paper performed remarkably well across multiple datasets. Our model exhibited comparatively low errors and high accuracy across all metrics. Specifically, our model excelled in MAE and RMSE, indicating

Table 3. Multiple dataset comparison results of our method

	DeepFashion	iMaterialist Fashion	E-commerce Product Images	Myntra Fashion	Cdiscount's Image Classification Challenge	JD.com Product Image
MAE	0.071	**0.065**	0.073	0.071	0.066	**0.065**
R^2	**0.615**	0.603	0.610	0.617	0.611	0.622
AUC	0.728	0.713	0.712	0.713	0.717	**0.729**
Acc.	0.703	0.697	0.697	0.704	0.689	**0.708**
RMSE	**0.118**	0.114	0.124	0.111	0.113	0.110

Figure 9. Comparison of the method's performance in six datasets

its ability to predict sales more accurately with minimal deviation from the actual observed values. Furthermore, our model demonstrated superior performance in R-squared, AUC, and Accuracy metrics, indicating clear advantages in fitting sales trends and classification accuracy.

4.4 Ablation Study Results and Analysis

To comprehensively assess the robustness of the proposed model and the individual contributions of its components, we conducted a series of ablation experiments. These experiments were designed to isolate and analyze the separate contributions of key components within the model. Specifically, we systematically removed and added various modules, such as ResNet, bidirectional attention mechanisms, and bidirectional LSTM layers, to observe their impacts on the model's performance. The experimental results are shown in Table 4.

The experimental results are presented graphically in Figure 10.

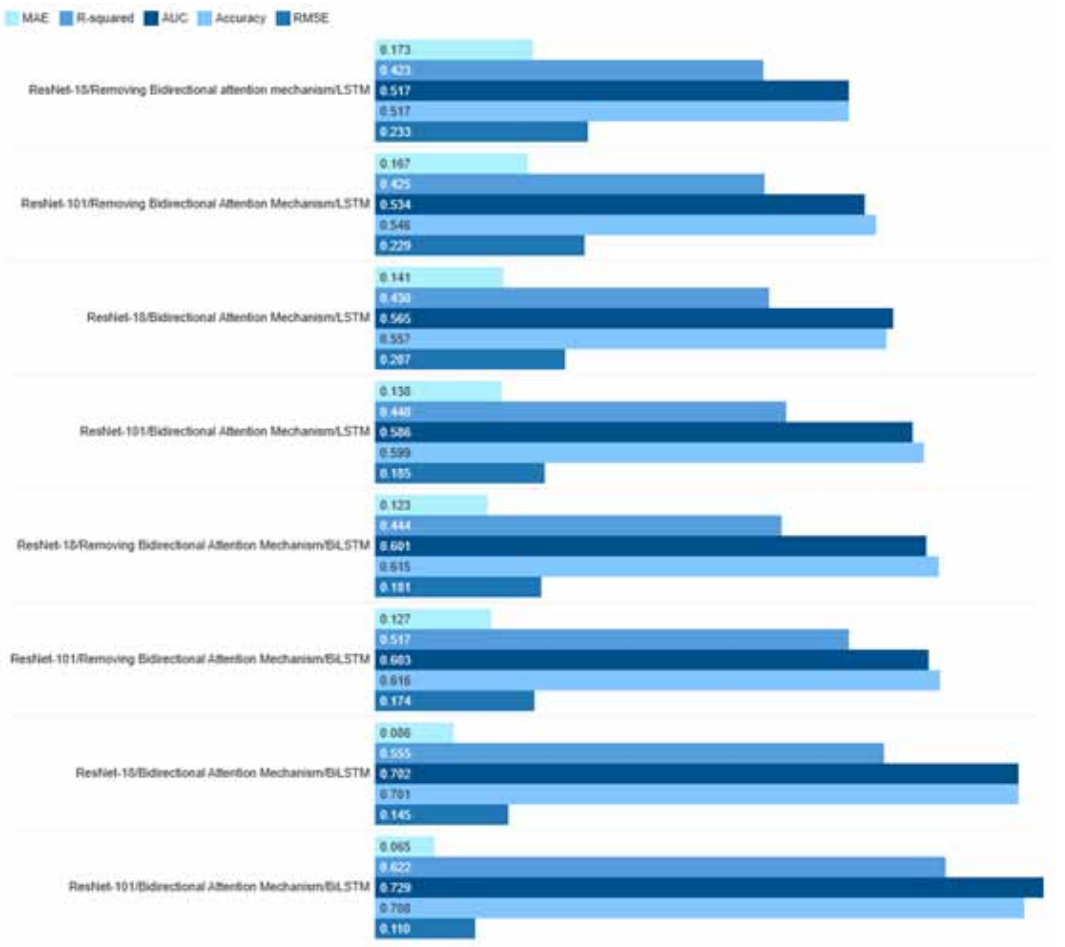
In the ablation experiments, we systematically replaced key components within the model, including ResNet-101 image feature extraction, bidirectional attention mechanism, and BiLSTM. Through comparative analysis of the experimental results, we observed a significant decrease in model performance when these components were replaced. Specifically, reducing the number of layers in the ResNet image feature extraction model led to a noticeable decline in the model's ability to extract image information features, resulting in decreased accuracy in sales prediction. Additionally, replacing the bidirectional attention mechanism diminished the model's ability to fuse multimodal data, affecting the understanding of data correlations. Moreover, replacing BiLSTM with LSTM failed to fully capture the temporal dependencies within sales data, resulting in unstable predictions of sales trends.

These findings underscore the critical roles played by ResNet-101 image feature extraction, bidirectional attention mechanism, and BiLSTM in the model's performance. Their organic integration

Table 4. Ablation study results

Model	MAE	R-squared	AUC	Accuracy	RMSE
ResNet-18 / Removing Bidirectional attention mechanism / LSTM	0.173	0.423	0.517	0.517	0.233
ResNet-101 / Removing Bidirectional Attention Mechanism / LSTM	0.167	0.425	0.534	0.546	0.229
ResNet-18 / Bidirectional Attention Mechanism / LSTM	0.141	0.430	0.565	0.557	0.207
ResNet-101 / Bidirectional Attention Mechanism / LSTM	0.138	0.448	0.586	0.599	0.185
ResNet-18 / Removing Bidirectional Attention Mechanism/BiLSTM	0.123	0.444	0.601	0.615	0.181
ResNet-101/Removing Bidirectional Attention Mechanism/BiLSTM	0.127	0.517	0.603	0.616	0.174
ResNet-18 / Bidirectional Attention Mechanism / BiLSTM	0.086	0.555	0.702	0.701	0.145
ResNet-101 / Bidirectional Attention Mechanism / BiLSTM	**0.065**	**0.622**	**0.729**	**0.708**	**0.110**

Figure 10. Ablation study results graphically



equips the model with more accurate and stable sales prediction capabilities, highlighting the unique advantages of the proposed model in this study.

5. CONCLUSION

5.1 Review of the Research

This study aims to address the challenge of multi-modal data fusion in sales forecasting for e-commerce platforms. A novel approach based on deep learning and attention mechanisms is proposed. Firstly, image features are extracted using the ResNet network, followed by the introduction of bidirectional attention mechanisms to dynamically integrate multi-modal data, including images and text. Subsequently, bidirectional LSTM is employed for temporal modeling to accurately capture the sales trends in the data. In the experiments, multiple datasets were selected, and the performance of our model and six baseline models was comprehensively evaluated using metrics such as MAE, R-squared, AUC, Accuracy, and RMSE. The results demonstrate the superior performance of our model across various metrics, particularly in achieving smaller errors in MAE and RMSE, highlighting its advantages in sales forecasting accuracy. The innovation of this study lies in the integration of

deep learning, multi-modal data fusion, and attention mechanisms, providing e-commerce platforms with a more precise and efficient sales prediction tool. Through the incorporation of these advanced models and methods, the research aims to comprehensively understand the intricate correlations within e-commerce data, providing more precise predictions for future sales trends. This holds significant practical implications for decision-making and business optimization in the e-commerce industry, contributing to the improvement of sales strategy effectiveness and overall platform performance.

5.2 Outlook

Although significant progress has been made in this study, there remains a major limitation, namely, the suboptimal performance in handling extreme sales scenarios. Existing models may struggle to accurately predict sales quantities in extreme situations, such as surges during promotional events or the launch of new products. To address this issue, the next research direction should focus on enhancing the model's ability to recognize and handle exceptional sales situations. Possible approaches include introducing more sophisticated anomaly detection algorithms or designing specialized neural network modules to handle extreme sales scenarios, thus improving the model's stability and accuracy under extreme conditions.

Another limitation of this study lies in the relatively simplistic handling of textual data, failing to fully exploit the potential of textual information. The current model's extraction and utilization of textual features are limited, not fully capturing the complex relationship between textual descriptions and sales volume. To address this issue, future research can consider employing advanced natural language processing techniques, such as pre-trained language models like BERT, to extract richer and more contextual textual features. Additionally, exploring more advanced fusion methods for multimodal data, such as the introduction of image-text attention mechanisms, can further enhance the model's understanding and utilization of textual information. These improvements will contribute to enhancing the model's overall performance and applicability, making it more practically valuable in real-world e-commerce environments.

This study proposes an innovative e-commerce sales forecasting model by employing deep learning techniques and multimodal data fusion methods. The model integrates ResNet image feature extraction, bidirectional attention mechanism for multimodal data fusion, and bidirectional long short-term memory (BiLSTM) for sales prediction, achieving remarkable accuracy and stability. The superiority of the model is validated across multiple datasets and demonstrates excellent performance in real-world e-commerce scenarios. This research provides an advanced solution for e-commerce sales forecasting, enhancing the precision of business decision-making. Furthermore, it offers valuable insights for future in-depth studies in the field.

REFERENCES

Abdullah, A. S., & Dawood, A. J. (2023). Using Deep-Learning Algorithm to Determine Safe Areas for Injecting Cosmetic Fluids into the Face: A survey. *Anbar Journal of Engineering Sciences, 14*(1).

Bhoir, S., & Patil, S. (2022). Multimodal Data Guided Spatial Feature Fusion and Grouping Strategy for E-Commerce Commodity Demand Forecasting. *Mobile Information Systems*, *2021*, 1–14. doi:10.1155/2021/5568208

Chaudhuri, A., Messina, P., Kokkula, S., Subramanian, A., Krishnan, A., Gandhi, S., Magnani, A., & Kandaswamy, V. (2018). A smart system for selection of optimal product images in e-commerce. 2018 IEEE International Conference on Big Data (Big Data). IEEE.

Cheng, L., van Dongen, B. F., & van der Aalst, W. M. (2019). Scalable discovery of hybrid process models in a cloud computing environment. *IEEE Transactions on Services Computing*, *13*(2), 368–380. doi:10.1109/TSC.2019.2906203

Durga, B. K., & Rajesh, V. (2022). A ResNet deep learning based facial recognition design for future multimedia applications. *Computers & Electrical Engineering*, *104*, 108384. doi:10.1016/j.compeleceng.2022.108384

Efat, M. I. A., Hajek, P., Abedin, M. Z., Azad, R. U., Jaber, M. A., Aditya, S., & Hassan, M. K. (2022). Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales. *Annals of Operations Research*, 1–32. doi:10.1007/s10479-022-04838-6

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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA. doi:10.1145/3394486.3403362

Erdelyi, C. (2019). *Using Computer Vision Techniques to Play an Existing Video Game*.

Giri, C., & Chen, Y. (2022). Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry. *Forecasting*, *4*(2), 565–581. https://www.mdpi.com/2571-9394/4/2/31. doi:10.3390/forecast4020031

Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M. R., & Belongie, S. (2019). The imaterialist fashion attribute dataset. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. IEEE.

Jain, V., & Wah, C. (2022). Computer Vision in Fashion Trend Analysis and Applications. *Journal of Student Research*, *11*(1).

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, 609–629. doi:10.1016/j.ins.2021.11.036

Kumar, B., Singh, A. K., & Banerjee, P. (2023). A Deep Learning Approach for Product Recommendation using ResNet-50 CNN Model. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. IEEE.

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y., & Wang, Y. (2022). Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Liu, X. (2022). E-commerce precision marketing model based on convolutional neural network. *Scientific Programming*, *2022*, 2022. doi:10.1155/2022/4000171

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE.

Mehmood, I., Ullah, A., Muhammad, K., Deng, D.-J., Meng, W., Al-Turjman, F., Sajjad, M., & de Albuquerque, V. H. C. (2019). Efficient image recognition and retrieval on IoT-assisted energy-constrained platforms from big data repositories. *IEEE Internet of Things Journal*, *6*(6), 9246–9255. doi:10.1109/JIOT.2019.2896151

Ning, E., Wang, Y., Wang, C., Zhang, H., & Ning, X. (2024). Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. *Neural Networks*, *169*, 532–541. doi:10.1016/j.neunet.2023.11.003 PMID:37948971

Ning, X., Yu, Z., Li, L., Li, W., & Tiwari, P. (2024). DILF: Differentiable rendering-based multi-view Image–Language Fusion for zero-shot 3D shape understanding. *Information Fusion*, *102*, 102033. doi:10.1016/j.inffus.2023.102033

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., & Kompatsiaris, I. (2022). Multimodal Quasi-AutoRegression: Forecasting the visual popularity of new fashion products. *International Journal of Multimedia Information Retrieval*, *11*(4), 717–729. doi:10.1007/s13735-022-00262-5

Rights, J. D., & Sterba, S. K. (2023). R-squared measures for multilevel models with three or more levels. *Multivariate Behavioral Research*, *58*(2), 340–367. doi:10.1080/00273171.2021.1985948 PMID:35476605

Senapati, B., Talburt, J. R., Naeem, A. B., & Batthula, V. J. R. (2023). Transfer learning based models for food detection using ResNet-50. 2023 IEEE International Conference on Electro Information Technology (eIT). IEEE.

Sheela, T. (2022). Cloud Based E-Commerce Application For Organic Fertilizers, Pesticides And Other Products And Crop Disease Identification Using Computer Vision. *2022 International Conference on Computer Communication and Informatics (ICCCI).* IEEE.

Shetty, S. K., & Buktar, R. (2022). A comparative study of automobile sales forecasting with ARIMA, SARIMA and deep learning LSTM model. International Journal of Advanced Operations Management, 14(4), 366–387

Singh, A., & Kumar, D. (2022). Detection of stress, anxiety and depression (SAD) in video surveillance using ResNet-101. *Microprocessors and Microsystems*, *95*, 104681. doi:10.1016/j.micpro.2022.104681

Singla, P., Duhan, M., & Saroha, S. (2022). An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network. *Earth Science Informatics*, *15*(1), 291–306. doi:10.1007/s12145-021-00723-1 PMID:34804244

Soni, N., Sharma, E. K., Singh, N., & Kapoor, A. (2020). Artificial intelligence in business: From research and innovation to market deployment. *Procedia Computer Science*, *167*, 2200–2210. doi:10.1016/j.procs.2020.03.272

Thivakaran, T., & Ramesh, M. (2022). Sales Data Analysis and Prediction System for Big Mart using Deep Recurrent Reinforcement Principles. 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)

Wang, Y., & Chang, J. (2021). Future development trend of "new retail" and e-commerce based on big data. *Journal of Physics: Conference Series*. IEEE.

Wu, D., Ying, Y., Zhou, M., Pan, J., & Cui, D. (2023). Improved ResNet-50 deep learning algorithm for identifying chicken gender. Computers and Electronics in Agriculture, 205, 107622

Xu, W., Cao, Y., & Chen, R. (2023). A multimodal analytics framework for product sales prediction with the reputation of anchors in live streaming e-commerce. *Decision Support Systems*. doi:10.1016/j.dss.2023.113984

Yang, C., & Liu, Z. (2021). Application of computer vision in electronic commerce. Journal of Physics: Conference Series, Yang, T., & Ying, Y. (2022). AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, *55*(8), 1–37. doi:10.1145/3554729

Zhang, C., Liu, G., Zhan, X., Shi, H., Cai, H., & Li, Y. (2021). Multiple Object Tracking Algorithm Based on Mask R-CNN. *Journal of Jilin University Science Edition*, *59*(3), 609–618.

Zhang, X., & Kim, T. (2023). A hybrid attention and time series network for enterprise sales forecasting under digital management and edge computing. *Journal of Cloud Computing (Heidelberg, Germany)*, *12*(1), 1–21. doi:10.1186/s13677-023-00390-1 PMID:37122827

Zhang, Y., Zuo, X., Zuo, W., Liang, S., & Wang, Y. (2021). Bi-LSTM+GCN Causality Extraction Based on Time Relationship. *Journal of Jilin University: Science Edition*, *59*, 643–648. doi:10.13413/j.cnki.jdxblxb.2020152

Zhao, L., Zhang, M., Tu, J., Li, J., & Zhang, Y. (2023). Can users embed their user experience in user-generated images? Evidence from JD. com. *Journal of Retailing and Consumer Services*, *74*, 103379. doi:10.1016/j.jretconser.2023.103379

Zhou, L. (2020). Product advertising recommendation in e-commerce based on deep learning and distributed expression. *Electronic Commerce Research*, *20*(2), 321–342. doi:10.1007/s10660-020-09411-6