Summer 7-5-2020

Research on the Construction of Sales Forecasting Model of Fashion Products Based on Feature Representation of Multimodal and Deep Learning

Tingting Wang
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China, wangting199611@126.com

Li Li
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China; Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing, 210094, China, lily691111@njust.edu.cn

Wenming Huang
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China

Research on the Construction of Sales Forecasting Model of Fashion Products Based on Feature Representation of Multimodal and Deep Learning

Tingting Wang[1], *Li Li*[1,2], *Wenming Huang*[1*]
[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China
[2]Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing, 210094, China

Abstract: By improving the accuracy of sales forecasting, this paper provides support for fashion product sales enterprises to make better inventory management and operational decisions. The deep neural network is introduced into the construction of multimodal features, and the internal structure of different modes, such as historical sales features, picture features, and basic attribute features of products, are fully considered, and finally the sales forecasting model of fashion products based on multimodal feature fusion is constructed. In addition, combined with the actual data of the enterprise, the proposed model is compared with the exponential regression model and shallow neural network model. The paper finds that multimodal features and deep learning representation method has better performance than traditional methods (exponential regression and shallow neural network) in the task of predicting sales of fashion products. The results help enterprises use the deep learning method and the data of multiple modal to make accurate sales forecast.

Keywords: sales forecasting, deep learning, multimodal features, fashion products

1. INTRODUCTION

Since the value of fashion products will decline rapidly over time, it is of great significance for enterprises to improve the sales forecasting accuracy of fashion products for reducing the impact of demand fluctuation on inventory and reducing the probability of unsalable and out of stock.

Because the demand of fashion products fluctuates greatly, in the sales forecasting model， the more comprehensive the factors that reflect the sales trend contained, the higher the forecast accuracy. In the past, scholars used the structural static data such as product price and discount, or used the dynamic data of the sales changing with time in the historical sales time series to predict product sales, so there are many deficiencies in combining multiple data to forecast the sales of fashion products. In addition, in the task of sales forecasting under the e-commerce environment, few scholars research the role of product pictures, which have an important impact on the users' purchase decisions.

On the other hand, different modal data such as basic attributes, sales time series and pictures of the same product are a group of data with different structures. The traditional sales forecasting methods can't directly use these data, so it is very necessary to propose a feature extraction method for multimodal data to effectively improve the sales forecasting accuracy of fashion products. In recent years, deep learning model has been proved to be able to effectively learn the hidden information of different types of features, and have obvious advantages in establishing connections between different modes and feature fusion[1]. Therefore, when synthesizing multiple data to predict the sales of fashion products, we can consider designing the multimodal feature extraction and feature representation method based on deep learning.

Therefore, this paper focuses on the following three research issues: (1) How to use deep learning methods to represent multimodal features and achieve feature fusion. (2) How to build and evaluate the sales forecasting

* Corresponding author. Email: lily691111@njust.edu.cn (Li Li), wangting199611@126.com (Tingting Wang)

model of fashion products based on multimodal features. (3) How to verify the performance of the model in the task of sales forecasting of fashion products through empirical methods.

2. RELATED RESEARCH

2.1 Factors for sales forecasting of fashion products

The problem of sales forecasting has been studied for many years, and many researchers have proposed forecasting theoretical models. Through the summary, we can see that the main factors involved in sales forecasting include: time series features of historical sales [2-3], basic attribute features of products [4], etc.

In the early stage, scholars used methods based on time series and machine learning to predict sales. For example, Sim[2] proposed an autoregressive model, and Thissen et al.[3] used support vector machine to model and estimate time series data. Later, Ramanathan et al.[4] used the basic attribute features such as the discount of products that affect the user's purchase decisions to predict the sales of products, and obtained better results. In e-commerce, because users can't get the physical information of products, many product information needs to be displayed through pictures. Although pictures have never been used for sales forecasting, Kim et al.[5] confirmed that picture information can have a significant impact on the user's purchase decisions, and then affects product sales. In particular, deep learning provides an effective methodological basis for the extraction of picture features, making it possible to use pictures as an influencing factor for product sales forecasting.

In the field of fashion products, the existing sales forecasting researches mainly use the methods based on time series. For example, Choi et al.[6] solved the problem of less historical data of fashion products by combining the extreme learning machine (ELM) with the grey prediction model. At the same time, due to the short sales cycle of fashion products, the prediction accuracy of the time series methods is poor. Therefore, later scholars also explore more explanatory variables related to product sales to improve the forecast accuracy of the model. Sun et al.[7] explored the influence of product features such as color, size and price on the sales. Ni et al.[8] considered the influence of historical sales, seasons, holidays and discounts.

It can be seen that most of the existing researches predict the sales of fashion products based on the historical sales time series and product basic attribute data, but there is no research to prove the role of product pictures in the sales forecasting task of fashion products under the e-commerce environment. In addition, there is also a lack of research on the sales forecasting of fashion products by using historical sales time series, basic attributes and pictures of products simultaneously.

2.2 Multimodal feature representation based on deep learning

Data from different information channels, such as historical sales time series, basic attributes and pictures, describe different aspects of the same product, each different kind of data or data from each different observation perspective can be called a different mode. Neural networks with deep structure can effectively reveal the hidden internal structure among these data and extract high-level abstract features that are useful for classification or regression tasks. For example, in the research field of multimodal feature learning, through integrating the information of audio and video modes, Ngiam et al.[9] trained the depth confidence network and extracted the fused feature expression from the two modes, then obtained good results.

The methods of multimodal feature fusion based on deep learning have been widely used in many fields. In the task of human posture prediction, Chu et al.[10] designed a deep neural network, which combined the whole attention model and the part of the body attention model, and realized the fusion of various information from different sources. Feiran et al.[11] proposed a image and text sentiment analysis model, using a hybrid fusion framework of sentiment analysis to mine the recognition features and internal associations of visual and semantic content, and then make effective sentiment prediction. In the field of e-commerce, multimodal data can also better learn product features. For example, zahavy et al.[12] took the image and text description of e-commerce products as the research object, and carried out research on the product classification problems of multi-mode and multi category. Kannan et al.[13] used text descriptions and images of products to optimize product classification in commercial search.

In conclusion, although the existing feature representation and learning methods of deep learning have mature applications in multimodal data fusion, at present, under the specific task of fashion product sales forecasting, few people have discussed how to generate the most effective features representation based on multimodal data and deep learning feature representation methods, which is the focus and innovation of this study.

3. MODEL BUILDING

3.1 Feature construction

In view of the good performance of feature learning model with deep structure on multimodal high-dimensional unstructured data, we introduce deep neural network into multimodal feature extraction and multimodal feature representation model, and fully consider the internal structure of different modes, and then transform the original high-dimensional heterogeneous data into abstract semantic expression in the same feature space through multiple nonlinear transformation, to realize feature extraction of multi-modal shared fusion.

In the process of constructing the sales forecasting model of fashion products, we choose two kinds of characteristics: the historical sales features and the features that affect the user's purchasing behavior. The features in each category are divided in detail, as shown in Figure1. In product pictures, the color, texture, and contour of the product are the focus of the user’s attention, which can help the user to form an understanding of the product. In addition, the price and discount of the product will affect the user's perceived utility and perceived risk, and it will also have a significant impact on the user's purchasing decisions. For fashion products, the time to market will affect the value of the product, and then affect the user's purchase decisions[14]. So picture features and product basic attribute features together constitute the features that affect the user's purchase decisions in the model. In addition, we also selected the historical sales features commonly used in the study to add to the model.

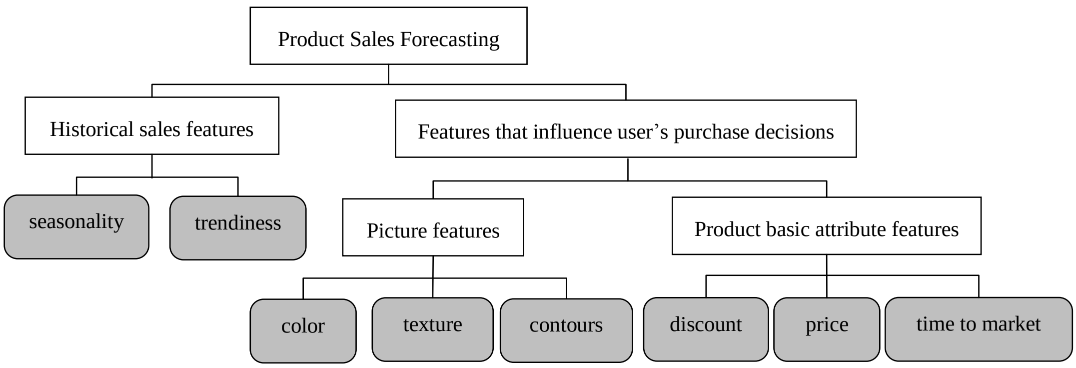


Figure 1. Data features of fashion products for sales forecasting

3.2 Feature representation

In recent years, deep learning has been widely used in feature extraction and high-dimensionality reduction. In the field of deep learning, representation refers to the input observation sample X of the model through the parameters of the model, in what form and in what way. Representation learning refers to learning the effective feature representation of observation sample X. In order to solve the problem of heterogeneous data structure of fashion product sales forecasting in multimodal data environment, we select suitable deep learning feature representation for different modal data describing fashion product sales.

(1) Deep learning representation of basic attribute features of fashion products

In this paper, aiming at the phenomenon that the relationship between sales and factors such as price in the sales forecasting is not simple linear, we use the feature representation based on the fully connected neural network (DNN), and connect all neuron nodes in adjacent layers such as the input layer, each hidden layer and the output layer through the fully connected neural network. Then we use the activation function (Relu) and cascade of neural network to express the non-linear features of structured data. The calculation method is shown in equation 1:

$$\mathrm{Relu}(x) = \max(0, x) \tag{1}$$

Let the output of the previous layer X be {x_1, x_2,..., x_n}, the connection weight between the previous layer and a node of the layer is {w_1, w_2,..., w_n}, the output of this node is o, and the calculation is shown in equation 2:

$$\mathrm{o} = \mathrm{Relu}\left(\sum_{i=1}^{n} (x_i * w_i)\right) \tag{2}$$

Assuming that this layer has *m* nodes, the matrix W composed of the connection weights between this layer and the previous layer is {W_1, W_2, ..., W_m}, where W_i is { $w_1^i, w_2^i, \cdots\cdots, w_n^i$ }, and the output of this layer is vector Output. The calculation method of Output is shown in equation 3:

$$\mathrm{Output} = \mathrm{Relu}(\mathrm{X} \cdot W^T) \tag{3}$$

It can be known that the fully connected neural network can construct the non-linear model based on structured data. Therefore, it can better learn the basic attribute features of products such as price, time to market.

(2) Deep learning representation of picture features of fashion products

In this paper, we use convolutional neural networks (CNN) to learn the picture features of fashion products. The core of CNN is to perform convolution operations on pictures to form new feature maps, extract high-dimensional features, so it can automatically extract the color, texture and contour feature of pictures. The convolution kernel consists of a third-order tensor. CNN uses the convolution kernel to traverse the input pictures, and is activated by the non-linear function Relu after element-level multiplication calculation, outputs an activation value, and finally gets a new feature map. Let the input of the convolutional neural network be X, which is used to represent the picture features of fashion products such as color and contour, and $\mathrm{X} \in \mathrm{R}_{m \times n}$, and the convolution kernel is $\mathrm{K} \in \mathrm{R}_{j \times k \times q}$. When the bias of the convolution kernel is b, the output of the convolutional neural network is O, and the calculation method is shown in equation 4:

$$\mathrm{O} = \mathrm{Relu}(\mathrm{conv}(\mathrm{X}, k) + b) \tag{4}$$

After the convolution calculation, the output of the convolution neural network is still a third-order tensor O, which can't be calculated directly. Then we expand the feature map into a vector O', eliminating the heterogeneity between different data modes, so as to realize the feature representation of picture data.

(3) Deep learning representation of historical sales features of fashion products

In view of the serialization feature of historical sales information, long short term memory network (LSTM) can learn and represent the feature of time series data better. Aiming at the seasonality and trendness feature in the historical sales data, LSTM solves the problem of sequence modeling by designing two state variables for storing short-term state and long-term state, saving the features of past time and output reference state variables of the current time. It is assumed that the historical sales information of fashion products at time t can be expressed by the long-term state C_t and the short-term state at that time as h_t, and the input x_t is the sales information at the current time. Then the LSTM uses the Sigmoid function and the forgetting gate operation with the symbol σ to determine which information in the current input and short-term state is forgotten and which information is updated into the long-term memory. The calculation method is shown in equation 5:

$$C_t^{'} = \sigma([h_{t-1}, x_t]) * \tanh([h_{t-1}, x_t]) \tag{5}$$

And then, the LSTM neural network updates the long-term state by formula X, and then combines the updated long-term state, the short-term state of the previous moment and the input of the current moment to calculate the short-term state of the current moment, that is, the output of the current moment. And the calculation method is shown in equation 6:

$$h_t = \sigma([h_{t-1}, x_t]) * \tanh(C_t) \tag{6}$$

Through the above calculation process, we can use LSTM neural network to better learn the time series features of fashion products that need to consider long-term dependence conditions, such as seasonality.

3.3 Feature fusion

In order to integrate data information of many different high-dimensional heterogeneous modes, and extract different formal feature vectors from each mode of the original data, we designed a new feature vector by combining the feature vectors obtained by different deep learning models. Finally, we got a fashion product sales forecasting model which considered different information modes, deep learning representation method, and feature fusion. The overall framework is shown in Figure2.

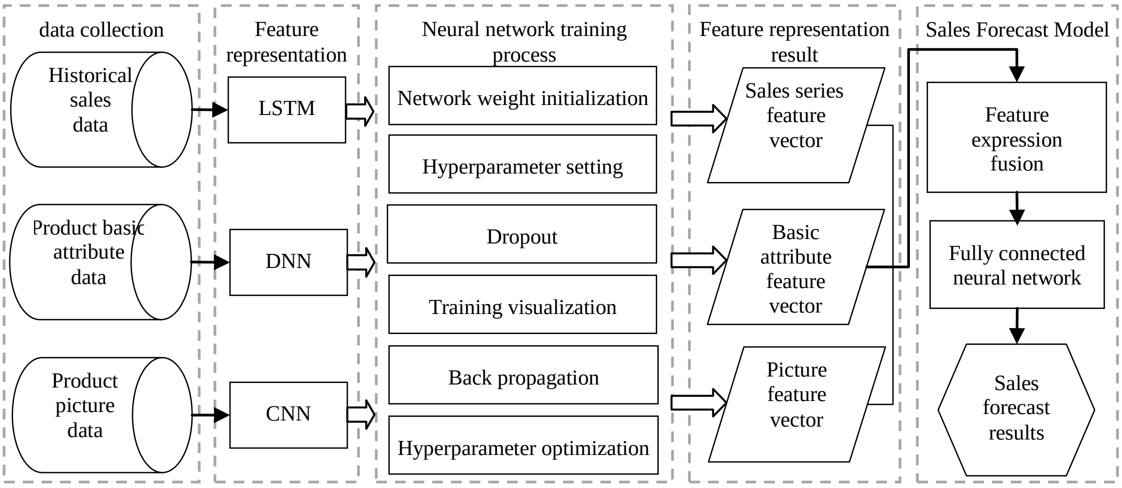


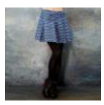
Figure 2. Framework of sales forecasting model of fashion products based on feature representation of multimodal and deep learning

4. EMPIRICAL RESEARCH AND RESULT ANALYSIS

4.1 Data

The data used in this experiment are 9,189,419 sales records from a clothing e-commerce company in Nanjing from September 2013 to December 2017, including 4,354 products. Among them, the data that can be used to reflect the time series feature of products are counted by week. Table 1 shows the original data with the product number 1427228, where *i* is the number of weeks in which the product is sold, indicating that the data provided is in the *i*-th week after the start of product sales. The product is sold for 32 weeks, x_i is the average of the original price of the product in the *i*-th week, y_i is the average of the discount of the product in the *i*-th week, z_i is the product sales in the *i*-th week.

Table 1. The original data for product number 1427228

product id	product original price (x_i, i=1, 2, …32)	time to market	product discount (y_i, i=1, 2, …32)	product image	product sales (z_i, i=1, 2, …32)
1427228	（135.32, 169.52, 162.86, 168.64, 162.1, 132.31, 148.32, 164.84, 160.11, 161.76, 161.75, 161.13, 168.29, 186.14, 119.67, 117.21, 99.97, 98.62, 98.36, 99.43, 153.3, 112.25, 143.42, 145.79, 137.11, 121.79, 139.15, 117.01, 103.29, 118.67, 107.57, 110.78）	Summer	（6.09, 6.74, 4.44, 9.01, 4.30, 12.41, 5, 6.31, 4.93, 7.44, 8.48, 7.97, 5.92, 2.99, 4, 14.64, 7.27, 1.64, 5, 2.44, 8.37, 8.29, 11.59, 12.43, 7.9, 4.34, 5.19, 5.69, 6.18, 8.24, 5.49, 2.83）		（84, 33, 37, 70, 194, 478, 150, 50, 42, 28, 38, 38, 31, 11, 3153, 38, 218, 56, 192, 19, 39, 1，59, 15, 152, 145, 29, 270, 96, 41, 64, 64）

4.2 Research process

(1) Data preprocessing. Because the sales and discount difference between different products is very large, we deal with the sales and discount logarithmically to avoid the influence of extreme values on the model. Therefore, we normalize all structured features. The normalization method is shown in equation 7:

$$\text{standarlization}(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{7}$$

Where, x represents an attribute in structured data, x_{min} represents the minimum value of the attribute, and x_{max} represents the maximum value of the attribute.

Since categorical data cannot be directly calculated, here one-hot encoding is used to encode the categorical data, and the categorical variables are converted into numeric vectors. For example, according to the time to market, we divide it into four seasons of listing. If the time to market is in summer, it is expressed as [0,1,0,0]. Because the resolution of the product pictures is inconsistent, and most product pictures have a resolution of 200×200, the bicubic linear interpolation algorithm is used to convert all the picture resolutions to 200×200.

(2) Using deep learning model to train. By adjusting the weights of neural networks such as CNN and LSTM, we can get the effective deep learning representation vectors of different modal features, avoiding problems that often occur in training such as overfitting. We use dropout for training during the training process. (3) Model parameter adjustment and optimization. In order to further improve the forecast accuracy, we adjust the learning rate and dropout retention rate to determine the optimal solution. In order to ensure the stable and effective training process, we visualized the loss function value and the output of each layer, and used the visual data chart to keep the training running normally. (4) Model effect comparison. In order to verify the efficiency of the multimodal learning model after the feature representation, we choose the classical exponential regression method and shallow neural network method to compare.

4.3 Analysis and results

(1) Model training and testing. We use random initialization to ensure the heterogeneity of neuron nodes. Suppose that each dimension of the input is a standard normal distribution, and the neural network weights are also initialized to a standard normal distribution, then the output of the current node is a normal distribution satisfying N (0, $(n_l-1)/2$), Where n_l-1 is the number of inputs for the current node. This will make the output of more neural network nodes in the inactive state of the activation function. For the modified linear unit, the local gradient in the inactive state is 0, which means that a large number of neuron nodes will not be trained. Therefore, we initialize the neural network weight to N (0, $2/(n_l-1)$), which can keep the output variance of each layer unchanged, and improve the training speed of the neural network.

In terms of setting hyperparameters, for basic attribute data and sales series data, we use the single hidden layer structure neural network with 8 nodes for feature representation respectively. For the neural network of picture data feature representation, due to computational resource constraints, we set the convolution kernel dimension to the common 5×5, the number of convolution kernels to 16, and the number of convolution layers to 4. The initial learning rate is set to 0.001 and the retention rate of dropout is set to 0.5. In order to further improve the accuracy of the model, we adjust the learning rate and dropout retention rate.

It can be seen from Figure3 that when the learning rate is less than 0.0005, the fitting speed of neural network is too slow to fit correctly in 100000 times of training, resulting in the rapid rise of the loss function value of train set and test set. Therefore, this paper determines the learning rate of the model to be 0.0005. In addition, we take 0.5 as the center, take five equally spaced points for testing, and search for the best retention rate. It can be seen from Figure4 that when the retention is 0.5 and 0.6, the test set loss function value performs better, while the train set loss function value is lower and the training speed is faster when the retention is 0.6. So this paper uses 0.6 as the retention rate of model training. The following experiments are carried out under

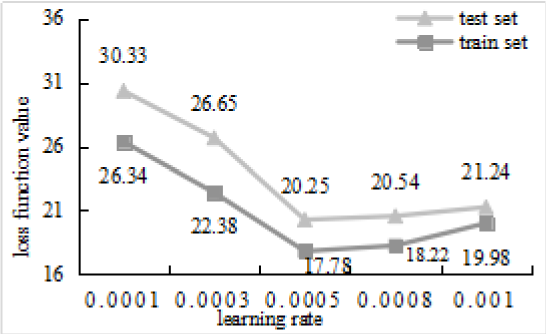


Figure 3. **Trend of learning rate - loss function value**

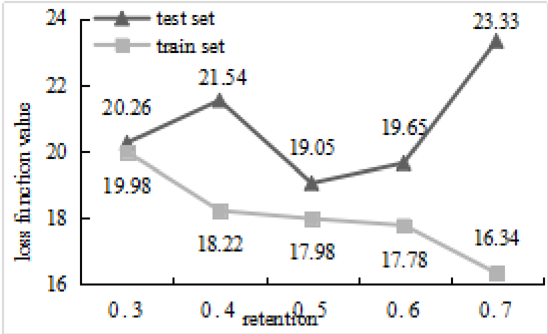


Figure 4. **Trend of retention - loss function value**

the above optimal parameters.

(2) Comparative test. In order to test the performance of our multimodal deep learning model in the fashion product sales forecasting task, we selected the exponential regression and shallow neural network commonly used in actual production for comparison. Then we select the company's summer dress from the sales date of 20 weeks of sales data for testing. The comparison between the predicted sales of each model and the real sales is shown in Figure5.

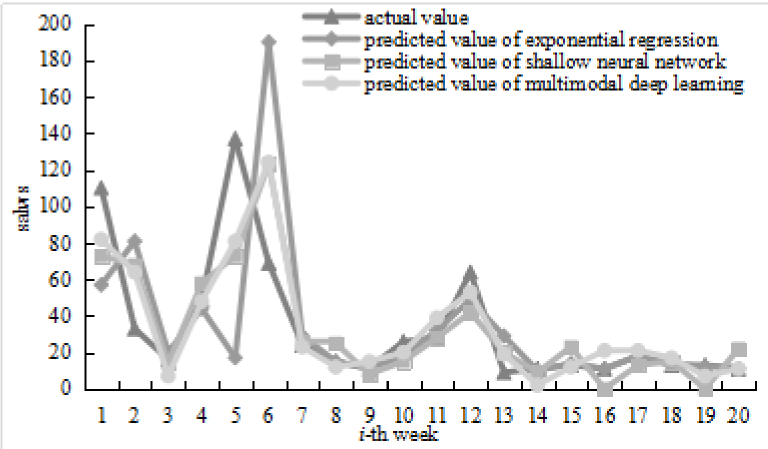


Figure 5. **Comparison between predicted sales and real sales of models constructed by different methods**

Table 2 shows the mean absolute deviation (MAD), mean absolute percentage error (MAPE) and root mean square error (RMSE) of three different models. It can be seen that under the three error indicators of MAD, MAPE and RMSE, the prediction results of the model proposed in this paper are better than the exponential regression model and the shallow neural network model. It can be said that the sales forecasting model of fashion products proposed in this paper is more accurate than the traditional models.

Table 2. **Error comparison of prediction models constructed by different methods**

	Multimodal Deep Learning Model	Exponential regression model	Shallow neural network model
MAD	12.95	20.35	15.70
MAPE	41.17%	41.78%	51.75%
RMSE	20.78	41.79	23.53

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we combine multimodal data and deep learning feature representation methods to build a relevant prediction model for the task of sales forecasting of fashion products, and prove that different modal features and deep learning representation methods have better results than the traditional methods in this prediction task. The research results show that when the basic attribute features and historical sales features of products are represented respectively by the single hidden layer neural network with eight nodes, the picture features are represented by a neural network with convolution kernel dimension set to 5×5 and convolution kernel number set to 16, and the learning rate and dropout retention rate are 0.0005 and 0.6 respectively, the accuracy of the sales forecasting model proposed in this paper is high, and the effect of preventing overfitting is the best. The results of the comparative test show that the model proposed in this paper has better prediction effect than the exponential regression and shallow neural network model.

This paper innovatively uses picture data, historical sales data and basic attribute data for feature representation and feature fusion, and applies them to the sales forecasting of fashion products, improving the forecast accuracy. Therefore, enterprises can use the model to predict product sales, in order to prepare the right products at the right time, better control seasonal operations costs, and improve their profitability of selling. And this paper provides a reference for enterprises to use deep learning methods in actual production operations. However, there are still some shortcomings. On the one hand, the scale of convolution neural network in this paper is small, which may lead to the inadequate use of picture data; on the other hand, the prediction accuracy of the model proposed in this paper in the later stage of product sales still needs to be improved. The follow-up research will further expand the network scale and improve the prediction accuracy of the model.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 71771122 and Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant SJCX19_0042.

REFERENCES

[1] Baltrušaitis T, Ahuja C, Morency L P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443.

[2] Sims C A. (1980). Macroeconomics and Reality. Econometrica, 48(1): 1-48.

[3] Thissen U, Brakel R V, Weijer A P D, et al. (2003). Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems, 69(1-2): 35-49.

[4] Ramanathan U, Muyldermans L. (2010). Identifying demand factors for promotional planning and forecasting: A case of a soft drink company in the UK. International Journal of Production Economics, 128(2): 538-545.

[5] Kim M, Lenonn S. (2008). The effects of visual and verbal information on attitudes and purchase intentions in Internet shopping. Psychology and Marketing, 25(2): 146-178.

[6] Choi T M, Hui C L, Liu N, et al. (2014). Fast fashion sales forecasting with limited data and time. Decision Support Systems, 59: 84-92.

[7] Sun Z L, Choi T M, Au K F, et al. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1): 411-419.

[8] Ni Y, Fan F. (2011). A two-stage dynamic sales forecasting model for the fashion retail. Expert Systems with Applications, 38(3): 1529–1536.

[9] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011.

[10] Chu X, Yang W, Ouyang W, et al. (2017). Multi-context attention for human pose estimation. https://arxiv.org/abs/1702.07432.

[11] Feiran H, Zhang X, Zhao Z, et al. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167: 26-37.

[12] Zahavy T, Magnani A, Krishnan A, et al. (2016). Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce. https://arxiv.org/abs/1611.09534.

[13] Kannan A, Talukdar P P, Rasiwasia N, et al. Improving Product Classification Using Images. In: the 2011 IEEE 11th International Conference on Data Mining, Vancouver, BC, Canada, 2011.

[14] Mao Zhaofang, Liu Wanwan, Li Hui. (2016). Joint strategy of advance-selling and buy-back for seasonal perishable products. Journal of Management Sciences in China, 19(2): 74-84. (in Chinese)