MDF-Net: Multimodal Deep Fusion for Large-Scale Product Recognition

Yanling Pan[1], Ruizhi Zhou[1], Gang Zhao[2], Weijuan Zhang[1], Delong Chen[1], and Fan Liu[1(✉)]

[1] Hohai University, Nanjing, China
fanliu@hhu.edu.cn

[2] Jiangsu Institute of Water Resources and Hydropower Research, Nanjing, China

Abstract. Large-scale production recognition systems are crucial for building efficient E-commerce platforms. However, various traditional product recognition approaches are based on single-modal data input (e.g., image or text), which limits recognition performance. To tackle this issue, in this paper, we propose a Multimodal Deep Fusion Network (MDF-Net) for accurate large-scale product recognition. The MDF-Net has a two-stream late fusion architecture, with a CNN model and a bi-directional language model that respectively extract semantic latent features from multimodal inputs. Image and text features are fused via Hadamard product, then jointly generate results. Further, we investigated the integration of attention mechanism and residual connection to respectively improve the text and image representations. We conduct experiments on a large-scale multimodal E-commerce product dataset MEP-3M, which consists of three million image-text product data. MDF-Net achieves a 93.72% classification accuracy over 599 fine-grained classes. Empirical results demonstrated that the MDF-Net yields better performance than traditional approaches.

Keywords: Multimodal fusion · Product recognition · Attention mechanism · Deep learning

1 Introduction

With the rapid development and popularization of the e-commerce, the commodity management has become one of the main tasks of e-commerce platforms, deeply influencing their economic benefits. Only reasonably classifying commodities can improve the accuracy of shopping guides and the efficiency of searching for products on e-commerce platforms. Traditional Product Recognition methods [1–3] are mainly based on single-modal data. However, on existing e-commerce platforms, the information of commodities exists in multiple modalities, such as commodity texts, images, and videos. The traditional Product Recognition methods do not make full use of this information, which extremely limit its classification performance.

W. Deng et al. (Eds.): CCBR 2022, LNCS 13628, pp. 501–510, 2022.
https://doi.org/10.1007/978-3-031-20233-9_51

Multimodal Machine Learning [4] aims to integrate two or more modal information to make classification. It merges information from different modalities to obtain comprehensive features and improve classification accuracy. Ngiam et al. [5] first introduced deep learning to multimodal learning, using deep AutoEncoders to learn the shared representation features, and the experiment verified the effectiveness of the multimodal shared representation. Multimodal Machine Learning can make full use of each modal data and establish the association between different modal data.

In this paper, we propose MDF-Net: Multimodal Deep Fusion for Large-scale Product Recognition. We first introduce an Attention Mechanism in the text feature extraction model - Bi-LSTM to improve the ability of discriminating text features. We also use the deep neural network VGG-19 [6] to extract commodity image features. Then, the extracted text and image features are fused by Hadamard product to explore the internal relationship between text and image features. And a residual struction is added to enhance the image features. Subsequently, the enhanced image features are used as the initial weights of the original text features extraction model to enhance the text features. Ultimately, the enhanced text features and image features are enforced to the matrix Hadamard product to obtain the multimodal features of product. Compared with the single-modal Product Recognition method and the Multimodal Product Recognition model based on Concatenate Fusion, the proposed method has better classification effectiveness and accuracy. The main contribution of the paper can be summarized as below.

- We proposed MDF-Net, which is able to explore complementary information from multiple modalities. Experimental results show that it achieves better classification performance than traditional single-modal counterparts.
- We integrates attention mechanism to improve text representations, and used residual connection to improve image representations. Ablation experiments demonstrated the effectiveness of these modifications.

The rest of the paper is organized as follow. Section 2 reviews the related work. Section 3 elaborates on the implementation of the MDF-Net model. Section 4 presents the data collection, model evaluation, and experimental results. Section 5 finally concludes this paper.

2 Related Work

In this section, we briefly review the existing researches via Machine Learning in the field of product recognition. The approaches can be classified into two categories, Single-modal Machine Learning based Methods and Multi-modal Machine Learning based Methods, which are respectively shown below.

2.1 Single-modal Product Recognition

Traditional product classification methods are mainly based on single modal data. Common single modal data includes text data and image data of products.

The following shows a brief introduction to the Single-modal Machine Learning based Methods using text data or image data.

Text-Based Approaches. Using the titles or descriptions data to classify product is an important task in the e-commerce industry. Zhong et al. [1] proposed a Temporal Multiple-Convolutional Network (TMN), using the text titles of product for classification. They integrated Temporal Convolutional Network (TCN) model and Multiple-Convolutional Neural Network (MCNN) model to extract features from text, which achieved higher classification accuracy than that of the state-of-the-art models at that time. The Bert model [7] has further promoted the development of text classification. Zahera et al. [2] used the fine-tuned Bert model (ProBERT) to classify product into hierarchical taxonomy with product titles and descriptions. Their work achieved great prediction performance in MWPD2020 [8]. However, with the explosive growth of product images emerging in ECommerce platforms, this kind of approach shows poor performance for lacking enough visual features. The text-based product classification uses informative and intuitive data and is widely applied because of its simplicity and efficiency.

Image-Based Approaches. Product images are essential for consumers to select product. However, product images often contain noisy information, such as promotional text and product packaging information, which make image-based product classification more difficult. And how to fuse low-level features and high-level features from the product images better still remains a key task. Wazarkar et al. [9] used linear convolution to obtain core features of product images. They then combined them with the local features of product images to improve the classification effect. Considering that Attention Mechanism can capture local information, Yang et al. [3] constructed a two-level convolutional Attention network and a cluster-based grouped Attention network to fuse low-level visual features and high-level semantic features of product images. This model performs well in fine-grained product classification. Overall, compared with text-based product classification, image-based product classification often achieves poorer accuracy, since the information implied in images is more complex and difficult to extract. Nonetheless, the images sometimes involve the information that texts lacks. Hence, it is important to integrate text data and image data to extract complementary features.

2.2 Multi-modal Product Recognition

Multimodal learning aims to obtain comprehensive information from multimodal data, generally using text data and image data. However, it is challenging task to fuse multimodal data together to obtain higher accuracy. The fusing structure of the model may deeply influence the classification performance. Additionally, the traits of deep learning models can also affect the classification accuracy. Zhang et al. [10] believed that it was necessary to infer the complex relationship between

visual objects for visual question-and-answer. They used a bilinear attention module as the attention for problem guidance of visual objects to complement the fine-grained fusion of visual and textual features. Misikir et al. [11] adopted multiple neural networks, including CamemBERT, FlauBERT and SE-ResNeXt-50, to learn the features from textual data and visual data. With respect to the fusing techniques, they experimented on addition fusion, concatenation fusion and average fusion. However, these adopted fusion methods are too simple and could not solve the problem of losing low-level information when extracting visual features. Although these methods obtain better experimental results than Single-modal Machine Learning based Methods, they still achieved relatively moderate accuracy and there is still much room for improvement.

In summary, the Multimodal ML based Method can make full use of each modal data of the product, which leads to higher classification accuracy. However, there still exist two problems to solve, how to extract multimodal data better and how to fuse the extracted multimodal features to obtain higher product classification accuracy. In this paper, we propose two methods to solve these problems respectively.

3 Multimodal Deep Fusion Network

The structure of the Multimodal Deep Fusion Model for Large-scale Product Recognition (MDF-Net) is shown in Fig. 1. The model includes four parts: attention-based text feature extraction network, residual image feature extraction network, multimodal fusion, and multimodal classifier. Each part is described in detail below.

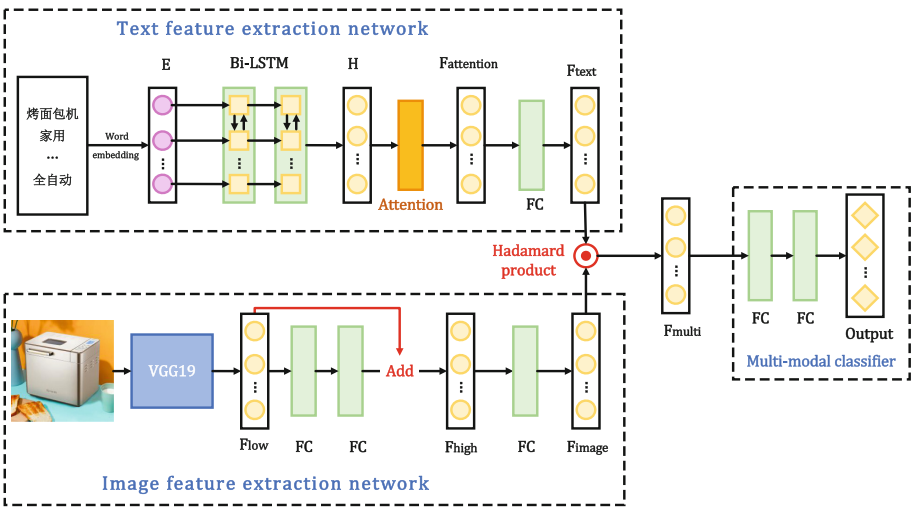

Fig. 1. The structure of the multimodal deep fusion product recognition model.

3.1 Attention-Based Bi-directional Language Model for Text Representation

We use Bi-LSTM as the backbone network for extracting features in the commodity title. Then the attention mechanism is introduced to capture semantic information in the text. The network structure is shown in Fig. 2. Our attention-based Bi-LSTM [12] model includes the input layer, embedding layer, Bi-LSTM layer, attention layer, and output layer.

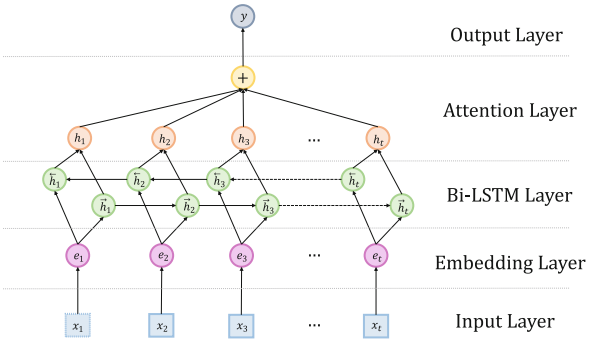

Fig. 2. The structure of Bi-LSTM model based on attention.

Firstly, word token sequences in the input layer denoted by $\{x_1, x_2, ..., x_t\}$ (t is the number of words in the text) are fed into an embedding layer that maps each word in the text to a low-dimensional space through the embedding matrix $W_{embedding}$, which is obtained by unsupervised pre-training. The output of the embedding layer is the word vector $E = \{e_1, e_2, ..., e_t\}$, where each element corresponds to a word. Then, Bi-LSTM layers perform feature extraction on the word embedding and derive $H = \{h_1, h_2, ..., h_t\}$, where h_i is represented as:

$$h_i = [\overrightarrow{h} \oplus \overleftarrow{h}]. \tag{1}$$

The $\overrightarrow{h}$ represents the historical context representation, while the $\overleftarrow{h}$ represents the future context representation, and $\oplus$ represents the concatenate operation. That is, the high-level feature H is composed of historical and future context information by concatenation. In the subsequent attention layer, we re-weight the features via attention mechanism as follows:

$$\alpha = \text{softmax}(w^T \tanh(H)) \tag{2}$$

$$F_{\text{attention}} = \tanh(H\alpha^T), \tag{3}$$

where $H \in \mathbb{R}^{d \times t}$ is the embedding extracted by the Bi-LSTM model; w represents a learnable linear projection, whose shape is $d \times 1$; the dimension of the attention map α is $1 \times t$, and $F_{\text{attention}}$ is the vector representation for classification. Finally, the output layer maps $F_{\text{attention}}$ to $F_{\text{text}} = \{f_1, f_2, ..., f_k\}$ through a fully connected layer, where k denotes the number of categories.

3.2 Residual Convolutional Neural Network for Image Representation

In this paper, we adopt VGG-19 [6] model to extract features from images. The deep network structure of VGG-19 can make the feature map broader and more suitable for large data sets. The applied VGG-19 has 19 layers, which can be divided into five groups of convolutional layers and three fully connected layers. VGG-19 also uses a 3×3 small convolution kernel, which can reduce the number of parameters and improve calculation efficiency. In spite of these advantages, there is a problem with this network. That is, the deep VGG-19 model is easy to lose low-level feature information when extracting image features in the transmission of information between the convolution layer and the full connection layer.

To deal with this problem and to further strengthen the complementarity of the two modal data of text and image, we introduce residual blocks to improve performance. Concatenate operation cannot complement the high-level and low-level features well. Nevertheless, the residual blocks can be used to directly transfer low-level features to high-level features so that they can reduce information loss and improve the identification of image features.

The network mainly includes the VGG-19 image feature extraction module and image feature reconstruction module based on the residual blocks. The image feature reconstruction module includes two layers of fully connected layer and addition of the skip connection.

Lastly, changing the dimension of F_{high} through a linear layer, We then obtain the final image features F_{image}.

3.3 Multimodal Fusion and Classification

In this paper, Hadamard product is applied to multimodal feature fusion. Hadamard product is a type of matrix operation between matrixes with same dimension. Suppose $A = (a_{ij})$ and $B = (b_{ij})$ are two matrixes in the same order, and the matrix $C = (c_{ij})$ is the Hadamard product of A and B, where $c_{ij} = a_{ij} \times b_{ij}$. For example, suppose $A, B \in R^{m \times n}$, then the $m \times n$ matrix shown in Eq. 4 is called the Hadamard product of matrices A and B, denoted by $A \odot B$.

$$A \odot B = \begin{bmatrix} a_{11} \cdot b_{11}, & a_{12} \cdot b_{12}, & \cdots & a_{1n} \cdot b_{1n} \\ a_{21} \cdot b_{21}, & a_{22} \cdot b_{22}, & \cdots & a_{2n} \cdot b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} \cdot b_{m1}, & a_{m2} \cdot b_{m2}, & \cdots & a_{mn} \cdot b_{mn} \end{bmatrix} \tag{4}$$

Compared with other traditional fusion methods, like concatenation and average fusion, the Hadamard product can mine the internal relationship between text and image. Beside, it has a smaller multimodal feature dimension than the concatenate fusion method. The parameter scale is also smaller when it is transferred to the fully connected layer.

The calculation formula of the multimodal fusion feature is shown in Eq. 5. In this case, the dimension of F_{multi} is unchanged.

$$F_{\text{multi}} = F_{\text{text}} \odot F_{\text{image}} \tag{5}$$

The multimodal classifier is composed of multiple fully connected layers and softmax activation functions. To evaluate the performance of the classification model, the cross-entropy loss function is used to optimize the model. In addition, the cross-entropy loss can also reduce the risk of gradient vanishing during stochastic gradient descent. Suppose the sample set is $\{x_1, x_2, ..., x_N\}$, the label set is $\{y_1, y_2, ..., y_K\}$, N is the total number of samples and K is the total number of labels. The cross-entropy loss function is defined as:

$$L = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{K} y_i \log\left(\frac{\exp(w_i^T \phi(x_i))}{\sum_{k=1}^{K} \exp(w_k^T \phi(x_i))}\right), \tag{6}$$

where x_i is the i-th sample and w is the final linear classification layer. The y_i represents the ground-truth class lable.

4 Experiment

4.1 Experimental Setting

The MDF-Net proposed in this paper is implemented using Python, TensorFlow, and Keras. All experiments were performed using Intel Core i5-9400F CPU with 2.90 GHz, TITAN GPU. During the experiment, the batch size of each iteration is 64 and the number of epochs is 200.

In our experiments, we use MEP-3M [13] dataset, which is large-scale, hierarchical categorized, multi-modal, fine-grained, and longtailed. This dataset includes 3 million product image-text pairs of 599 categories. Each product has both image and its description text. Some image samples and the text cloud of the titles are respectively given in Fig. 3(a) and (b). The training set and the test set are divided at a ratio of 8:2. The accuracy and the precision-recall (PR) curve were used to evaluate the classification performance of each model:

(a) Sampled images (b) Sampled text cloud

Fig. 3. Some image examples and the text cloud of the titles in the MEP-3M dataset.

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} TN_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i + \sum_{i=1}^{N} FP_i + \sum_{i=1}^{N} TN_i} \tag{7}$$

$$\text{Precision} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i} \tag{8}$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}, \tag{9}$$

where TP, FP, FN, and TN are the number of true positive, false positive, false negative, and true negative samples, respectively. The abscissa of the PR curve is Recall, and the ordinate is Precision.

4.2 Main Results

Table 1. Performance comparison between different product recognition methods

Classification type	Model	Accuracy
Text-only approaches	Attention-based Bi-LSTM	86.12%
Image-only approaches	VGG-19	74.82%
	Residual VGG-19	75.37%
Multimodal approaches	LMF [14]	89.22%
	TFN [15]	90.70%
	MDF-Net (concat)	90.69%
	MDF-Net (w/o-residual)	92.03%
	MDF-Net	**93.72%**

To verify the effectiveness of the proposed MDF-Net, we first compare the accuracy of the single-modal product recognition model and some multimodal product recognition models, such as LMF [14], TFN [15]. We further implement "MDF-Net (w/o-residual)" and "MDF-Net (concat)" to demonstrate the effectiveness of residual connection for image representation and Hadamard product for multimodal fusion.

The performance comparison between different product recognition methods are listed in Table 1. As shown in this table, the MDF-Net model proposed in this paper is more accurate than the single-mode methods and other multimodal models. The accuracy of MDF-Net reached 93.72%, achieving prominent experimental performance. Beside, with respect to the results of MDF-Net (concat), MDF-Net (w/o-residual), VGG-19 and Residual VGG-19, we can see that the residual module can make a distinct contribution to the product recognition accuracy. And with the comparison between the result of MDF-Net (w/o-residual) and MDF-Net, the Hadamard product shows great improvement of experimental performance with 1.69% increasement, which proves the effectiveness of the introduced Hadamard product.

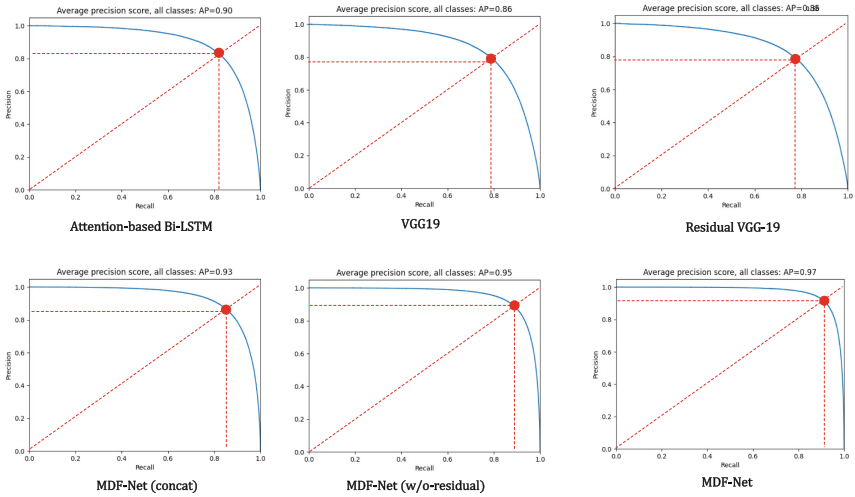

Fig. 4. PR curves. (Color figure online)

The PR curve is also used to evaluate the performance. It evaluates the model performance by comparing the value of the balance point (the red dot in the figure, where Precision = Recall) or the average precision (AP) value of different models. As shown in Fig. 4, the balance point and AP value of multimodal model is significantly larger than that of single-modal model, and both the direct and residual MDF-Net are larger than that of MDF-Net (concat) model. In addition, compared with MDF-Net (w/o-residual), the residual MDF-Net has higher AP value and further improves the performance of Product Recognition.

The accuracy and PR curves show that the MDF-Net proposed in this paper significantly improve the accuracy and performance of product recognition. Moreover, the residual one has better classification accuracy and performance.

5 Conclusion

This paper introduces a multimodal deep fusion method for large-scale product recognition (MDF-Net). The MDF-Net uses the attention mechanism to improve text feature discrimination and introduces residual blocks to enhance image features. In addition, our multimodal model uses the Hadamard Product to explore the internal connection between the text and image features, so as to obtain better multimodal features and improve the accuracy of Product Recognition. As a result, the MDF-Net model achieves 93.72% classification accuracy, significantly improving classification performance. And the comparisons of experimental results prove the effectiveness of the introduced residula blocks and Hadamard Product.

Acknowledgment. This work was partially funded by Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory(2021JCJQ-SYSJJ-LB06905), Water Science and Technology Project of Jiangsu Province under grant No. 2021072, 2021063.

References

1. Zhong, C., Jiang, L., Liang, Y., Sun, H., Ma, C.: Temporal multiple-convolutional network for commodity classification of online retail platform data. In: Proceedings of the 2020 12th International Conference on Machine Learning and Computing, pp. 236–241 (2020)
2. Zahera, H.M., Sherif, M.: ProBERT: product data classification with fine-tuning BERT model. In: MWPD@ ISWC (2020)
3. Yang, Y., Wang, X., Zhao, Q., Sui, T.: Two-level attentions and grouping attention convolutional network for fine-grained image classification. Appl. Sci. **9**(9), 1939 (2019)
4. Morency, L.P., Liang, P.P., Zadeh, A.: Tutorial on multimodal machine learning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, Seattle, United States, pp. 33–38. Association for Computational Linguistics, July 2022
5. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
6. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781 (2016)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Zhang, Z., Bizer, C., Peeters, R., Primpeli, A.: MWPD 2020: semantic web challenge on mining the web of html-embedded product data. In: MWPD@ ISWC (2020)
9. Wazarkar, S., Keshavamurthy, B.N.: Fashion image classification using matching points with linear convolution. Multimedia Tools Appl. **77**(19), 25941–25958 (2018). https://doi.org/10.1007/s11042-018-5829-4
10. Zhang, W., Yu, J., Hu, H., Hu, H., Qin, Z.: Multimodal feature fusion by relational reasoning and attention for visual question answering. Inf. Fusion **55**, 116–126 (2020)
11. Misikir Tashu, T., Fattouh, S., Kiss, P., Horvath, T.: Multimodal e-commerce product classification using hierarchical fusion. arXiv e-prints (2022) arXiv-2207
12. Li, L., Nie, Y., Han, W., Huang, J.: A multi-attention-based bidirectional long short-term memory network for relation extraction. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.-S.M. (eds.) ICONIP 2017. LNCS, vol. 10638, pp. 216–227. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70139-4_22
13. Chen, D., Liu, F., Du, X., Gao, R., Xu, F.: MEP-3M: a large-scale multi-modal e-commerce products dataset
14. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 (2018)
15. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017)