Prediction of Outfit Compatibility Based on Weighted Multimodal Fusion

Yan Li[1][0000-0001-6773-6299] *, Xiaobao Fu[1][0000-0001-5503-3668], You Du[1][0000-0001-6460-1540]

[1]Department of Information Engineering, Lankao Vocational College of San Nong, Zhongzhou Street, Henan, China

liyanjuly@163.com*, flyxiang99@163.com, lksnjwc@163.com

Abstract. Owing to the influence of too many complex factors between items, it is difficult to outfit compatibility. Although the current outfit compatibility technology has good results, however, previous works which focus on the compatibility of modeling based on the image mode or the text mode fail to make full use of the complex relations between text and image information interaction. To solve the clothing matching prediction of a single mode limit the variety of modes different information interaction, a method of clothing matching based on multi-modal fusion with weights is proposed. Firstly, this method extracts clothing images as visual features, and at the same time extracts text information as textual features. Secondly, weighted fusion of the extracted visual features and text features. Finally, the fused features are used to input into the graph neural network model to learn the outfit compatibility. The fused features capture the most important clothing features into the clothing representation, thereby effectively improving the accuracy of outfit compatibility prediction.

Keywords: Multi-modal fusion; Outfit compatibility; Graph neural network; Visual feature; Textual features.

1 Introduction

With the rapid development of online fashion industry, new visual problems also arise (Dong 1998) [6]. In recent years, the problem of clothing matching has attracted extensive attention in the field of computer vision [10] [25]. "Predicting the matching degree of clothing" refers to determining whether a set of fashion clothes can match well. However, outfit compatibility is a complex task, which depends on subjective concepts of style, context and trend - all of which may vary from different persons and evolve over time. Therefore, this problem goes beyond the traditional visual similarity problem. It needs to model and infer the collocation relationship between different clothing categories, as well as the relationship between color, material, pattern, texture, shape and other clothing factors. This paper aims to propose an effective clothing matching prediction method to help people make appropriate clothing matching.

H. Wu et al. (Eds.): CSIEDE 2022, ACSR 103, pp. 3–14, 2022.
https://doi.org/10.2991/978-94-6463-108-1_2

In recent years, clothing matching technology based on deep learning has attracted more and more attention, because deep neural network can extract important clothing features from images, and the characteristics of clothing products play an important role in clothing design, which is particularly important for clothing matching prediction. Previous clothing matching research [4] [18] [23] [27] mainly focused on the image information of clothing. Although the clothing image contains the most important information of clothing matching, it can not capture all the features in the image for clothing matching. Relevant clothing descriptions usually highlight some important information and supplement the imperceptible clothing features in the image (such as those related to fit, fabric, brand, style or occasion) to help predict and model clothing collocation.

This paper focuses on the two most common modes in clothing matching prediction: clothing image information and text information in clothing description. Previous clothing matching studies (Song 2018) [29] mainly used text information to improve image embedding and improve clothing matching degree. For example, Han (Han 2017) [30] et al. Used two-way long-term and short-term memory (LSTM) to model the garment image sequence, and took the multimodal data as the input. In the multimodal semantic space, the garment image was embedded in a position close to its corresponding description. Vasileva [28] et al. Proposed a multimodal semantic space, in which the corresponding image and description are closely embedded together, and then the embedded image is mapped to the matching space specific to the clothing type. By closely mapping the corresponding image and description in the multimodal semantic space, these methods do not capture the complementarity and related associations between image and text. Therefore, this paper proposes a weighted multi-modal fusion method to solve the limitation of information exchange between image mode and text mode in clothing matching. Weighted modal fusion highlights some parts of images or words in clothing matching, which correspond to the important clothing features of clothing matching, so as to effectively improve the prediction accuracy of clothing matching. The main contributions of this paper are:

- This paper proposes a multi-modal fusion clothing matching prediction method, including image mode and text mode of fashion products, so as to make better use of different modes and different information interaction between modes to obtain rich feature information.
- In this paper, weighted modal fusion is proposed to highlight some important clothing features of images or words in clothing matching, so as to effectively improve the prediction accuracy of clothing matching.

2 Related work

The field of packaging is a very important application of computer vision [11]. Research in this field mainly focuses on garment image retrieval [12] [19] garment image attribute learning [2] [13] [20] garment fashion trend prediction [1] [9] and garment matching degree modelling [28]. For example, Liu et al. proposed a potential support vector machine (SVM) [7] model for clothing recommendation in different scenes. Iwata et al. proposed a theme model, recommending "top" to "bottom" [16]. Hu et al. [14] studied personalized clothing recommendation using data sets collected from fashion websites. McAuley et al. [23] Based on Amazon's co purchase data set, mapped clothing in potential space and introduced similarity measurement to calculate clothing matching degree. He et al. [15] introduced a scalable matrix decomposition method, which adds the visual information of commodity pictures to the learning of partial order pairs of commodity preferences for clothing recommendation.

Previous fashion analysis mainly focused on visual features, but failed to consider text information. In order to make up for the above defects, Han and others regard a set of clothing as an ordered sequence, and model the clothing sequence through two-way long-term and short-term memory (LSTM), so as to learn the compatibility relationship between clothing. However, the interaction between text information and visual information is limited, there is no good capture, and there are various relationships between image and text. Li et al. [21] proposed a multi-channel deep learning framework to classify whether a given garment is popular. This method does not highlight some information of images or words corresponding to the important items features of outfits matching. Different from the above research, this paper extracts the weighted multi-modal fusion method to find the complex compatibility between item modes.

3 Model

This paper uses the method based on the improved graph neural network [5] proposed by Cui et al to construct the clothing matching prediction model. The outfit data are denoted as a set $V=\{v_1,v_2,v_3\cdots\}$ in the training, in which is an outfit. We use a set $V=\{v_1,v_2,v_3\cdots\}$ to represents a collection of clothes, which v_i represents a piece of clothes. Each piece of clothes has an image and a text description. Firstly, the visual features and text features of the clothing are extracted, and then the visual features and text features are fused. The fused features $f_{\mathrm{i}}^{\mathrm{vt}}$ are input into the graph neural network model, and finally the matching degree of the clothing is calculated.

3.1 Problem Formulation

According to the different action results between different types of clothes in clothing, a directed graph model is constructed. In the graph, each node represents a clothing category, the fusion characteristics of each clothing are represented by, each edge represents the information transmission between two clothes, and the adjacency matrix is used to describe how the nodes in the graph interact. First, calculate the weight from node to node according to the training set (1):

$$w_{(n_i,n_j)} = \frac{\frac{T_{c_i,c_j}}{T_{c_j}}}{\frac{\sum_k T_{c_j,c_k}}{T_{c_k}}} \tag{1}$$

Where T_{c_j,c_k} is the co-occurrence frequency of the categories c_j and c_k, and T_{c_j} is the occurrence frequency of category c_j;

Then the graph neural network model is used to calculate the information transmission between the nodes in the graph, the state information of the final node is obtained through multiple information exchanges between the nodes, and the matching degree of the suit of clothing is calculated by the attention mechanism. As shown in Figure 1:

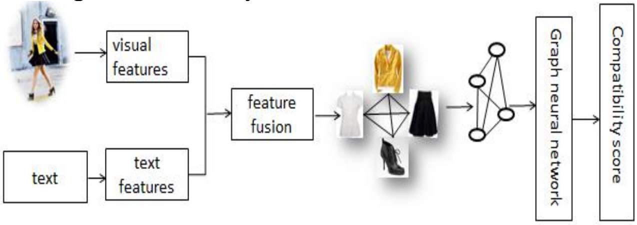


Fig. 1. Outfit scoring model

3.2 Datasets

The data used in this paper is based on the Polyvore (Han 2017) data set proposed by Han et al. The data set collects 21889 sets of clothing from Polyvore fashion website. Each clothing contains rich multimodal information, such as clothing pictures, text description, popularity score and price. According to the graph segmentation algorithm proposed by Han et al. (Han 2017, Li 2017), the data set is divided into 17316 sets as the training set, 1497 as the verification set and 3076 for the test set.

3.3 Feature Extraction

This section extracts visual features from clothing images and text features from clothing text description. The extracted visual features and text features are used to construct the matching degree prediction model of clothing.

3.3.1 Visual Feature.

In the process of clothing matching, image is the key factor of clothing performance. No matter how expressive the text description is, low-level visual features are difficult to express in words. This paper uses the deep convolution neural network googlenet perception V3 model [26] to extract the visual features of the image; The clothing image samples are input into the neural network model, and the linear layer outputs 2048 dimensional vectors as the visual features of the clothing.

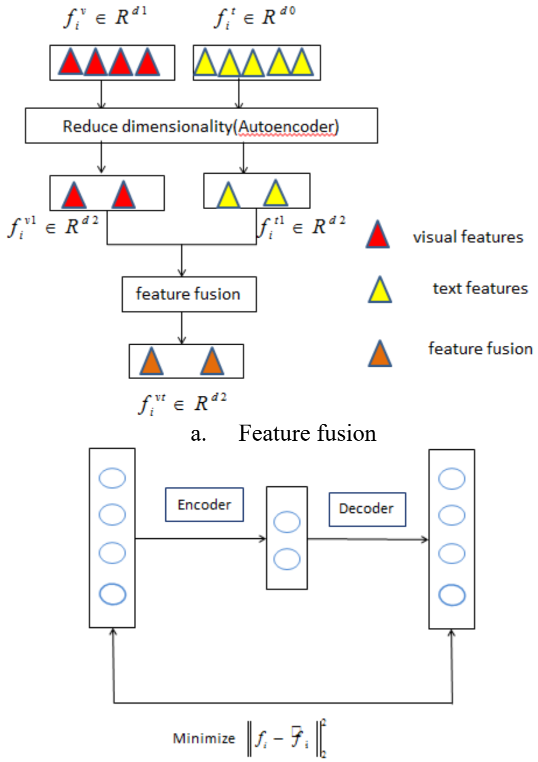
3.3.2 Textual Feature.

Due to the short text description, this paper uses bag of words Scheme [17] to extract text features. Firstly, build a vocabulary according to the words in the text description: filter the words that appear less than 5 times and the useless words with a length of less than 3 characters, and finally obtain a vocabulary of 2757 words. Therefore, this paper represents the text feature of each garment as a 2757 dimensional Boolean vector.

3.4 Feature fusion

Using visual features or text to represent a dress, the interaction between visual features and text features is very limited. On the contrary, this paper hopes to highlight some parts of images or words in feature expression, which correspond to the important clothing features of constructing a set of fashion clothing combination. This paper proposes to integrate the visual features and text features of clothing into a multimodal clothing representation method, so as to enhance the better interaction between visual features and text features, and make the model more accurately predict the collocation degree of a set of clothing.

The feature fusion process used in this paper is shown in Figure 2. Image visual features and text features are fused together. The specific operations are as follows:



a. Feature fusion

b. Reduce the dimension (Autoencoder)

Fig. 2. Feature preprocessing.

1. Firstly, autoencode is used to reduce the dimension of visual features and text features to obtain a unified dimension d2. Autoencode consists of two parts: encoder and decoder. The encoder injects the input features f_i into the potential representation space f_{i}^{1} through the function F, and the decoder computes the potential representation space f_{i}^{1} through the function to reconstruct the representation f_i, so as to minimize the error ζ between the input features and the reconstructed features. As shown in formula (2) and formula (3):

$$\begin{cases} F: f_i^1 = (\mathrm{w}_i \cdot \mathrm{f}_i + \mathrm{b}_i) \\ g: f_i^1 \rightarrow \hat{f}_i \end{cases} \tag{2}$$

$$\zeta = \frac{1}{2}\left\| f_i - \widehat{f}_i \right\|_2^2 \tag{3}$$

The weight w_i and deviation b_i of the decoder are respectively, and the model parameters are trained through the coding layer and decoding layer.

2. The processed visual features and text features are weighted and fused. ψ is a trade-off parameter (0 to 1) to balance the weight of the visual feature channel and text feature channel. The formula is as follows:

$$f^{vt} = \psi f^{v1} + (1 - \psi) f^{t1} \tag{4}$$

3. At the same time, add or Hadamard point multiplication fusion is performed to compare the two.

3.5 Mode via Graph Neural Networks for Outfit Compatibility

The basic idea of graph neural network is to aggregate the information of each node and surrounding nodes in the graph, so that the new representation of each node contains not only information about itself, but also information about its neighbor nodes. In order to better capture the close and complex relationship between multiple clothes, this paper proposes to use graph neural network to construct clothing matching prediction model.

The garment matching model based on graph neural network includes three steps. The first step is to learn the initial state of nodes based on the constructed graph model; The second step is to model the node interaction and update the node state. The third step is to use the attention mechanism to calculate the clothing matching degree. The specific operations are as follows:

Step 1: for each set of clothes containing different types of clothes, this paper uses the linear mapping matrix w_h^i to map the characteristics of each clothes to the potential space, and the obtained potential space representation h_i^0 is used to initialize the node state, as shown in formula (5):

$$h_i^0 = tanh\left(w_h^i f_i^{\mathrm{vt}}\right) \tag{5}$$

Step 2: the node transmits its status information to other nodes and receives the information of other nodes, which is called node information dissemination. In the propagation of step t, the sum of information received by the node from its neighbor nodes is:

$$\begin{cases} a_i^t = \sum_{n_j \to n_i \in E} A[n_j, n_i] W_p^{n_j \to n_i} h_j^{t-1} + b_p \\ A[n_i, n_j] = \begin{cases} w(n_i, n_j), n_i \to n_j \in E, \\ 0, others. \end{cases} \end{cases} \tag{6}$$

Where $W_p^{n_j \to n_i}$ and b_p are the weights and deviations of the edge shared linear transformation.

After receiving the neighbor node information a_i^t, the status of n_i will be updated:

$$\begin{cases} z_i^t = \sigma(W_z a_i^t + U_z h_i^{t-1} + b_z), \\ r_i^t = \sigma(W_r a_i^t + U_r h_i^{t-1} + b_r), \\ \tilde{h}_i^t = tanh(W_h a_i^t + U_h(r_i^t \odot h_i^{t-1}) + b_h), \\ h_i^t = \tilde{h}_i^t \odot z_i^t + h_i^{t-1} \odot (1 - z_i^t), \end{cases} \tag{7}$$

Where $w_z, w_r, w_h, b_z, b_r, b_h$ is the weight and deviation of update gated recursive unit (GRU) [22], z_i^t and r_i^t represent update gate vector and reset gate vector respectively.

Step 3: we can get the final state of the node through secondary information dissemination. This paper aims to design an attention mechanism to simulate the influence of different types of clothes on the matching degree of the whole set of clothes. Since nodes receive status information from other nodes, their node representations can perceive global information. Similar to Li et al. [24], the attention mechanism is used to calculate the graph-level output to obtain the clothing matching degree x_s, as shown in formula (8):

$$x_s = \sum_i^{|s|} \sigma(\theta(h_i^t)) \bullet \alpha(\delta(h_i^t)) \tag{8}$$

Where $\theta(\bullet)$ gives weight to a single piece of clothes (different clothes have different effects on a set of clothes), and $\delta(\bullet)$ is scores the compatibility between clothes (evaluates the matching effect between a certain clothes and other clothes), representing the information transmission state of the node in step;

3.6 Objective function

This paper is similar to Cui [5], and the objective function is:

$$\zeta_1 = \zeta_{bpr} + \frac{\lambda}{2}||\Theta||^2 \tag{9}$$

Where θ is the hyperparameter of the optimization graph neural network model and λ is the hyperparameter of norm L_2. According to Bayesian framework [31], the training set is defined as: $D_s = \left\{ \left(s, s^- \right) | s = \left\{ v_1, \mathrm{v}_2 ... v_{p-1}, v_p, v_{p+1} ... \right\}, s^- = \left\{ v_1, \mathrm{v}_2 ... v_{p-1}, \tilde{v}_p, v_{p+1} ... \right\}, s \in S \right\}$; For each set of clothing s , random selection $\tilde{v}_p \in V$, s^- is incompatible. Its objective function is:

$$\zeta_{bpr} = \sum_{(s,s^-) \in P_{\text{train}}} - ln\, \sigma\, (x_s - x_{s^-}) \tag{10}$$

No dot should be included after the sub subsection title number.

4 Experimental results and analysis

4.1 Model evaluation

In this paper, AUC (area under curve) is introduced as the standard of evaluating the model. The larger the value of AUC, the better the model represents. The expression is:

$$AUC = \frac{1}{|p_{test}|} \sum_{(s,s^-)} \delta(x_s > x_{s^-}) \tag{11}$$

s is the positive sample in the data set, s^- is the negative sample in the data set, p_{test} is the sample pair composed of positive samples and negative samples, $\delta(x)$ is the indicator function. If $x_s > x_{s^-}$, $\delta(x) = 1$,otherwise $\delta(x) = 0$.

4.2 Analysis of experimental results

(1) Compare different fusion dimensions.

The visual features and text features are weighted and fused, and the clothing features are fused into the dimension vector. The performance under different dimensions is shown in Table 1:

Table 1. The assessment results of model based on different dimensions (d2).

d2	AUC
64	73.50%
128	76.71%
256	79.93%
512	76.28%

In this experiment, this section attempts to find out the impact of fused features on performance in dimensions. When the learning rate, training batch and super parameters are the same, the fusion feature dimension is changed from 64 to 512. According to the results in the table, when the fusion feature dimension is 256, the model performance reaches the best value. If the feature dimension is too large, it is easy to cause information redundancy, and if the dimension is too small, it is easy to cause important information. Even if the feature dimension is too large and too small, it will have an adverse impact on the model performance.

(2) Compare different learning rates.

Similarly, in order to find the best learning rate to control the weight adjustment speed in the network, this paper carries out the control variable experiment, and controls the learning rate variable to explore the influence of the learning rate on the performance of the model while keeping other parameters unchanged. It can be seen from table 2 that when the learning rate is 0.001, the effect is significantly better than that when the learning rate is 0.01 and 0.0001. Therefore, this paper determines that the learning rate of the model is 0.001.

Table 2. The assessment results of model based on different learning rate.

learning rate	AUC
0.01	77.02%
0.001	79.93%
0.0001	76.56%

(3) Compare different fusion methods.

In order to verify the effectiveness of the feature weighted fusion method, the comparative experiments of different algorithms are carried out in the same data set and experimental environment. AUC is used to evaluate these algorithms. Specifically, this paper uses the siamesenet algorithm proposed by vasileva et al., Hadamard point multiplication, add operation and the weighted fusion proposed in this paper to predict the matching degree of clothing.

Table 3. The assessment results of different algorithm.

algorithm models	AUC
SiameseNet [28]	77.20%
Hadamard	76.12%
Add	77.32%
This paper	79.93%

Based on the statistical results in Table 3, it can be seen that the performance of weighted feature fusion model is better than that of other fusion methods. Based on the method proposed by vasileva et al., the concepts of image similarity, text similarity and graphic similarity are captured, but the interaction between visual features and text features is very limited. Compared with Hadamard point multiplication and add operation, weighted feature fusion can better highlight some important features of images or words for clothing combination, so it can better predict clothing matching.

5 Conclusions

In this paper, a two channel weighted modal fusion is designed based on image mode and text mode to predict clothing matching degree. Firstly, the visual features of clothing image are extracted, and the text features of clothing text description are extracted; Then the extracted visual features and text features are weighted and fused.

Finally, the fused features are input into the graph neural network model for clothing matching prediction. The results show that highlighting some characteristics in clothing is of significance to clothing combination. This paper improves the prediction accuracy by weighted fusion of features. In the future, we will study the novel fusion mechanism of visual features and text features, so as to better integrate the interaction of fine-grained visual attributes and text attributes and improve the prediction accuracy of clothing matching.

References

1. Al-Halah, Z., R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion [C]//in Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 388-397.
2. Chen, Q., et al. Deep domain adaptation for describing people based on fine-grained clothing attributes [C]//in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 5315-5324.
3. Cheng, Z.-Q., et al. Video2shop: Exact matching clothes in videos to online shopping images[C]//in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 4048-4056.
4. Cucurull, G., P. Taslakian, and D. Vazquez. Context-aware visual compatibility prediction [C]//in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. p. 12617-12626.
5. Cui, Z., et al. Dressing as a Whole: Outfit Compatibility Learning Based on Node-wise Graph Neural Networks[C]//in The World Wide Web Conference. 2019. ACM. p. 307-317.
6. Dong, Q., S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes[C] //in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017. IEEE. p. 520-529.
7. Felzenszwalb, P., D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model[C]//in 2008 IEEE conference on computer vision and pattern recognition. 2008. IEEE. p. 1-8.
8. Gao, G., et al., Fashion clothes matching scheme based on Siamese Network and AutoEncoder [J]. Multimedia Systems, 2019. 25(6): p. 593-602.
9. Gu, X., et al. Understanding fashion trends from street photos via neighbor-constrained embedding learning[C]//in Proceedings of the 25th ACM international conference on Multimedia. 2017. p. 190-198.
10. Hsiao, W.-L. and K. Grauman. Creating capsule wardrobes from fashion images [C]//in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 7161-7170.
11. Hidayati, S.C., et al. What are the fashion trends in New York? [C]//in Proceedings of the 22nd ACM international conference on Multimedia. 2014. p. 197-200.
12. Hadi Kiapour, M., et al. Where to buy it: Matching street clothing photos in online shops [C]//in Proceedings of the IEEE international conference on computer vision. 2015. p. 3343-3351.
13. Huang, J., et al. Cross-domain image retrieval with a dual attribute-aware ranking network [C]//in Proceedings of the IEEE international conference on computer vision. 2015. p. 1062-1070.

14. Hu, Y., X. Yi, and L.S. Davis. Collaborative fashion recommendation: A functional tensor factorization approach[C]//in Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 129-138.
15. He, R. and J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback [J]. arXiv preprint arXiv:1510.01784, 2015.
16. Iwata, T., S. Wanatabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines[C]//in IJCAI. 2011. Citeseer. p. 2262.
17. Ji, R., et al., Mining city landmarks from blogs by graph modeling[C]// Multimedia, 2009. p.105-114
18. Liu Yu-jie, et al. FMatchNet algorithm for fast clothing matching[J]. Chinese Journal of Image and Graphics.2019. 24(06): p. 979-986. (in Chinese)
19. Liu, S., et al. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set [C]//in 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. IEEE. p. 3330-3337.
20. Liu, S., et al. Hi, magic closet, tell me what to wear! [C]//in Proceedings of the 20th ACM international conference on Multimedia. 2012. ACM. p. 619-628.
21. Li, Y., et al., Mining fashion outfit composition using an end-to-end deep learning approach on set data[J]. IEEE Transactions on Multimedia, 2017. 19(8): p. 1946-1955.
22. Li, Y., et al., Gated graph sequence neural networks [J]. arXiv preprint arXiv:1511.05493, 2015.
23. McAuley, J., et al. Image-based recommendations on styles and substitutes [C]//in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015. ACM. p. 43-52.
24. Rendle, S., et al., BPR: Bayesian personalized ranking from implicit feedback [J]. arXiv preprint arXiv:1205.2618, 2012.
25. Song, X., et al. Neurostylist: Neural compatibility modeling for clothing matching [C]//in Proceedings of the 25th ACM international conference on Multimedia. 2017. ACM. p. 753-761.
26. Szegedy, C., et al. Rethinking the inception architecture for computer vision [C]//in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 2818-2826.
27. Veit, A., et al. Learning visual clothing style with heterogeneous dyadic co-occurrences[C]//in Proceedings of the IEEE International Conference on Computer Vision. 2015. p. 4642-4650.
28. Vasileva, M.I., et al. Learning type-aware embeddings for fashion compatibility [C]//in Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 390-405.
29. Song, X., et al. Neural compatibility modeling with attentive knowledge distillation[C]//in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018. p. 5-14.
30. Han, X., et al. Learning fashion compatibility with bidirectional lstms[C]//in Proceedings of the 25th ACM international conference on Multimedia. 2017. ACM. p. 1078-1086.
31. Rendle, S., et al., *BPR: Bayesian personalized ranking from implicit feedback.* arXiv preprint arXiv:1205.2618, 2012.