



Learning compatibility knowledge for outfit recommendation with complementary clothing matching

Ruomei Wang, Jianfeng Wang, Zhuo Su^{*}

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Keywords:

Outfit composition
Clothing matching
Fashion compatibility

ABSTRACT

With the rapid development of mobile networks and e-commerce, clothing recommendation has achieved considerable success in recent years. Fashion outfit matching has become an essential component to users while shopping, which helps users to select and present items to individuals in a personalized fashion recommendation. Apparently, it is an arduous task to guide complementary clothing matching due to the complexity and subjectivity of fashion items. Some existing solutions have been presented in recent years, which are tending to discover a series of visual cues to establish the matching relations. However, it would be mismatched easily due to these methods being hard to represent all the potential semantic information from the appearance of clothes. To thoroughly make use of the visual characteristics of clothing products and the related description information, we propose a complementary clothing matching method with some compatibility knowledge, named it CCMCK shortly. For visual compatibility, we adopt the graph neural network to model the visual relationship between items. To generate an outfit that satisfies the requirement of fashion compatibility, we propose a matching way under the compatibility constraint and seek to recommend compatible items based on multi-modal compatibility. Finally, we performed a qualitative investigation on the fill-in-the-blank and fashion outfit compatibility tasks to evaluate the proposed method.

1. Introduction

With the rapid development of Internet technology, more and more user-oriented services and social multimedia network platforms have been formed in various fields, and personalized recommendation has become an active field to explore [1]. Outfit recommendation plays an increasingly important role in the online retail market. The purpose of outfit recommendation is to promote people's interest and participation in online shopping by recommending fashionable outfits that they may be interested in [2]. Outfit composition is a difficult problem to tackle due to the complex interplay of human creativity, style expertise, and self-expression involved in the process of transforming a collection of seemingly disjoint items into a cohesive concept [3]. To learn how to compose outfits means compatibility that is the items of possibly different type that can go together in an outfit. The research of clothing compatibility is a challenge task, there are different complex factors be considered (garment style, clothing texture, color, environment, even personal preference).

The artificial intelligence technology and availability of large-scale rich-annotated fashion dataset gained astonishing success in clothing item retrieval, fashion image classification and so on [4]. CNN was used to learn visual clothing style by making feature transformation according to mapping the image of items to the style space [5]. Recently,

graph neural networks have been applied to clothing compatibility [6, 7]. Clothing items and their pairwise compatibility are constructed as a graph, where vertices are the fashion items and edges connect pairs of items that are compatible, and compatibility between garments is obtained by learning and predicting the edges [8]. In addition, the scheme of personalized outfit recommendation with attribute-wise interpretability [9] and the method of learning tuple compatibility [10] are presented to model clothing compatibility. The research of clothing compatibility is more challenging since it needs to infer the compatibility relationships among fashion items that go beyond merely learning visual similarities [11]. To attain a satisfactory clothing compatibility capability, overall or partial visual information, compatible about style and material, textual representation for the fashion item and so on some multi tuple factors need to be computationally interpreted in terms of their latent semantics.

In response to the above requirements, we proposed a model about complementary clothing matching based on compatibility knowledge, dubbed CCMCK. This model focus on predicting the fashion compatibility from the visual-semantic embedding based on the consistency between the visual representation and textual representation for the same fashion item. The research of item-based clothing fashion compatibility is successful to match outfit from the style, texture and detail

^{*} Corresponding author.

E-mail address: suzhuo3@mail.sysu.edu.cn (Z. Su).

accessories [12], it can recommend compatible items corresponding to the queried item, but it is lack to consider the factors about partial visual information and textual information in clothing fashion compatibility. Fig. 1 shows the samples about compatibility recommendation. From Fig. 1(a) we can see the outfits compatibility is better to consider the factors about style, texture and detail accessories. And in Fig. 1(b), the effect of partial visual information in clothing fashion compatibility recommendation will be considered.

Moreover, the diversity of the outfits depends largely on the multi-modality compatibility, multi-modality information fusion technology has become an important technology in the field of intelligent image processing [13]. The image multi-modality information [14] consists of the visual image and textual descriptions. Since there is a semantic gap between the visual modality and textual modality, our goal is to learn the visual-semantic compatibility. We mainly concentrate on learning multi-modality compatibility with their visual-semantic embedding for the outfits. A completely CCMCK framework represents the high-level features of the outfit as the style of the clothing which comprises three function modules, a visual compatibility learning with graph neural network to extract visual information and make use of the word dictionary model to extract textual information, a compatibility model based on multi-modal embedding to learn compatibility and a complementary fashion outfit matching to learn the matching scheme based on compatibility knowledge. Our approach is evaluated on Polyvore¹ which is a popular social commerce website for fashion. The main contributions are as follows:

- We build a CCMCK model to explore how to guide complementary clothing matching based on multi-modality embedding to make more effective fashion recommendations. We make the most of visual information based on graph neural networks and textual information based on BERT model to learn visual-semantic embedding in latent space.
- We design a visual-aware model to learn the visual compatibility between items and research the visual relationships. Different from previous work, we mainly discuss multi-modality compatibility in complementary clothing matching and realize a complementary clothing matching algorithm to predict the compatibility scores of outfits, and replace incompatible items in outfits to generate compatible outfits according to the visualization.
- The experimental results demonstrate that the multi-modality compatibility obtains the superior effect in complementary clothing matching.

2. Related work

2.1. Fashion compatibility learning

Fashion compatibility has been widely deployed to model the clothing matching in shopping platforms. There are extensive works committed to learn fashion compatibility effectively. In order to enhance the performance of compatibility, various successful attempts towards fashion compatibility have been made [15], the early methods [5,16] are based on a great deal of data in the “co-occurrence” from the user’s data. Veit et al. [5] used the Siamese network to capture the semantic information of visual style. Lin et al. [17] applied an end-to-end network to calculate item–item compatibility from couples of the images to seek similarity features. Besides, Cucurull et al. [18] adopted a graph neural network to predict visual compatibility by combining the visual information from the context of items view. However, the context-based visual compatibility needs to consider the relationship of visual compatibility as well as to consider the attribute-level compatibility. Most of the methods have been proposed to learn compatibility in the embedded space [16,19,20]. Han et al. [16] put forward to learn compatibility embedding and then capture the relationships of compatibility. Yang et al. [19] exploited an attribute-wise method to learn

fashion compatibility from the attribute-level view and achieved good results. Zhan et al. [20] made appearance-preserve clothing synthesis by learning the visual features to generate superior quality clothing. Furthermore, in the research of apparel style, the consistency of the context is applied to learn the fashion compatibility [21–23].

2.2. Complementary clothing matching

Clothing matching is more increasing attention in recent years, there are many studies focusing on fashion compatibility modeling recently [24,25]. Yang et al. [19] proposed a ExFCM model to generate the item-level compatibility evaluation and made explanations from the attribute-level view. Fashion compatibility is a subjective sense of human for relationships between fashion items, which is essential for fashion clothing matching. Sun et al. [26] combined semantic and visual embeddings to learn fashion compatibility, which adopt triplet ranking loss with compatible weights to measure fine-grained relationships between fashion items. Recently, the methods of estimating pairwise compatibility between a pair of items are presented [21,24,27,28]. He et al. [29] developed a personalized ranking with visual features based on implicit feedback data to retrieve visually similar apparel products from the database, the concept of visual style was captured. Chen et al. [21] extended this approach by identifying differences between styles (for example, fashion and gothic styles). Karpathy et al. [27] used automatic line-wrapping methods and manual annotations to identify the various style. Liu et al. [24] adopted multiple neural network to extract visual features and bag of words model to extract text features to learn compatibility in the latent space. In addition, there are also some studies on the concept of matching [30], such as the establishment of fashion semantic space (FSS) based on Kobayashi’s aesthetic theory. Li et al. [31] proposed semi-supervised compatibility learning across categories for clothing matching to learn the compatibility between fashion items across categories, which map items into a latent style space where minimize the distances between distributions. Gao et al. [32] proposed an novel model with siamese network and autoencoder based on both labeled data and unlabeled data to make clothing matching.

From the literature review we can see that there are some significant research results about the clothing compatibility have been reported to infer the compatibility relationships among fashion items and learns visual similarities [11]. Researchers built models of fashion compatibility of clothing from different perspectives. Because of the complex of clothing in structure, material, style and so on, the more considerations about clothing detail feature, textual representation feature will be more improvement clothing fashion compatibility adaptability under multimodal data. The model proposed in this paper is just to solve this problem. Different from previous work, we mainly discuss multi-modality compatibility in complementary clothing matching and realize a complementary clothing matching algorithm to predict the compatibility scores of outfits, and replace incompatible items in outfits to generate compatible outfits according to the visualization.

3. The proposed method

In this paper, we establish a complementary clothing matching method from multiple modalities based on the conventional graph network. The framework consists of three significant stages, including a visual compatibility learning process, a multi-modal compatibility computational model, and a complementary clothing matching scheme. First, we adopt a convolutional neural network to extract visual information and use a word dictionary model to extract textual information simultaneously. Then, the visual and textual information are jointly embedded into the latent space to learn the potential compatibility relationship. At last, the model constructs some matching schemes based on the compatibility knowledge, and it further recommends a series of suitable compatible items to compose the final outfit. The entire pipeline of the framework is illustrated in Fig. 2, and the technical detail will be described in the following.



Fig. 1. Samples of the clothing compatibility recommendation. (a) Item-based clothing fashion compatibility recommendation. (b) The clothing fashion compatibility recommendation with considering partial visual information. For example, the Mickey Mouse topic is a significant semantic cue that should be considered in the recommended results.

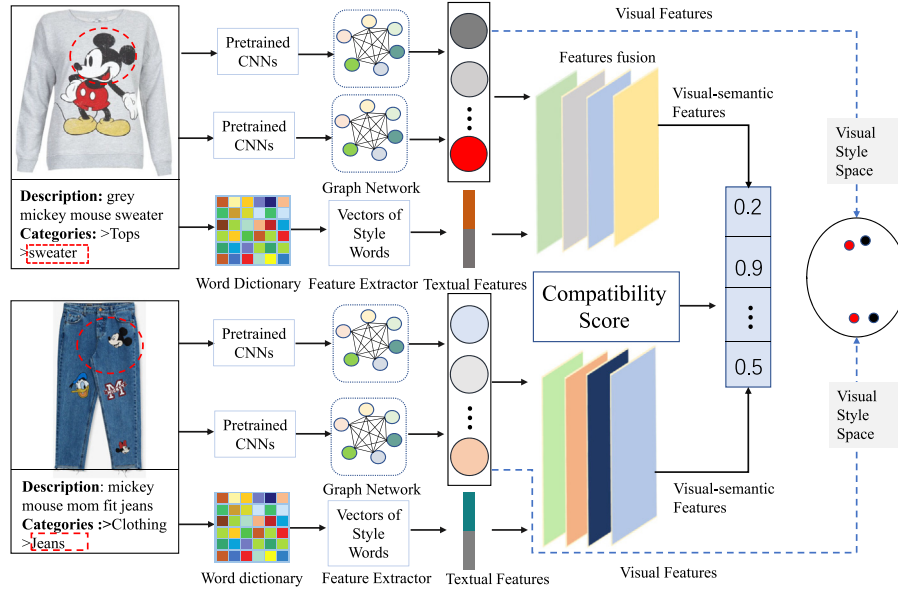


Fig. 2. The pipeline of the proposed framework. It is significant that a complementary clothing matching scheme is established based on multiple modalities compatibility information.

3.1. Problem formulation

Fashion clothing could be applied to express personalization and raise the effect of self-expression. Taking into account some clothes elements, such as fabric, style, pattern, and some attributes of outfits, we try to make up a meaningful clothes composition that is regarded as a fashionable collection of dresses, shoes, skirts, and other accessories of garments. Formally, a set of outfits is formulated as $S = \{s_1, s_2, \dots, s_i, \dots, s_n\}$, which consists of n items and kinds of multi-modal information from image, textual descriptions, and so on. All descriptions and images of outfits will be embedded into a style space, which aims to compute the compatibility of the outfit. The compatible score could be regarded as an indicator to evaluate the reasonable of the outfit composition.

3.2. Visual compatibility learning with graph neural network

Since the compatibility is affected by various factors, such as color, shape, and pattern, which makes it difficult to calculate the compatibility between each item of clothes. Therefore, it is significant to establish

a latent compatibility space for the relationship analysis of items. A possible way is that uses a kind of deep neural network to extract some practical features from the corresponding images of clothes, which could be considered to learn the visual embedding further and share style coherently. This visual embedding could be regarded as a compatibility space. Formally, it constructs a graph $G = (V, E)$ for fashionable outfits, where E is the set of edges and V is the set of nodes. Each item of the outfit could be regarded as a node. To enhance the visual representation between item pairs, a graph attention network is adopted in our solution, which is formulated as:

$$g_i = \tanh(W_i^r z_i). \quad (1)$$

Here, the visual feature z_i is derived from the convolutional neural network for the i th item in the outfits. The other details of the corresponding node would be taken as the input to enrich its entire features, which is simulated by propagating information from node to node. In other to distinguish the diversity for the items of outfits, we integrate an attention mechanism to calculate the different effects between nodes and neighbors. The graph attention network is imported to make a distinction between essential neighbor nodes, and the visual feature of

the node s_i is obtained by:

$$f_i = \text{ReLU} \left(\frac{1}{h} \sum_{h=1}^h \sum_{s_c \in N_i} a_{ic}^h g_c u_g^h \right), \quad (2)$$

where h stands for the size of the group. The weight matrix w_g^h is applied to calculate the scores between the adjacent nodes. The attention network plays an important role in discovering the essential nodes from their neighbors. In our solution, the following normalized equation is designed to calculate the attentive coefficients of the network:

$$a_{ic}^h = \frac{\exp(\sigma(\mathbf{a}^T [g_i \parallel g_c]))}{\sum_{s_c \in N(i)} \exp(\sigma(\mathbf{a}^T [g_i \parallel g_c]))}. \quad (3)$$

where \mathbf{a}^T denotes the normalized adjacency matrix. The nonlinear activation function $\sigma(\cdot)$ is adopted Leaky ReLU implementation. Besides, all the output features are transformed by the fully connected neural networks. That is,

$$\mathbf{z}_i = \sigma(\mathbf{w}_i \mathbf{f}_i + \mathbf{b}_i), \quad (4)$$

where the weight matrix $\mathbf{w}_i \in \mathbb{R}^f$, and bias vector $\mathbf{b}_i \in \mathbb{R}^f$. The predicted visual compatibility q_{ij} between a pair of items i and j could be evaluated by the following measurement,

$$q_{ij} = \varphi \left(\left| \mathbf{z}_i - \mathbf{z}_j \right| \mathbf{w}^T + \mathbf{b} \right). \quad (5)$$

Empirically, the weight matrix \mathbf{w}^T and the bias vector \mathbf{b} should be both learnable parameters, and $\varphi(\cdot)$ is an Sigmoid function. However, in some practices, we found that a few hot sellers or daily necessities would appear more frequently than other outfit items. Therefore, in order to deal with the data imbalance phenomenon, we further integrate a focus loss factor into cross-entropy loss function, which is defined as:

$$L_v = \sum_{i=1}^n y_{ij} (1 - q_{ij})^\gamma \log(q_{ij}), \quad (6)$$

where y_{ij} is the ground truth, and γ is a hyper-parameter.

3.3. Compatibility model based on multi-modal embedding

Except for the image appearance of clothes, an outfit also has abundant textual information provided by the manufacturer and customers to describe its fashionable styles. In order to bridge the gap between visual and textual information, we adopted the BERT model to extract features of words, which could capture the inherent consistency essentially. Consequently, the extracted semantic features \mathbf{t}_i and the visual features \mathbf{z}_i are uniformly embedded into the visual-semantic embedding space as,

$$\mathbf{x}_i = \sigma(w_{if} [\mathbf{t}_i; \mathbf{z}_i] + \mathbf{b}_{if}), \quad (7)$$

where w_{if} is a weight matrix, and \mathbf{b}_{if} is a bias vector. The training set contains compatible pairs $(\mathbf{x}_i, \mathbf{x}_j)$ and incompatible pairs $(\mathbf{x}_i, \mathbf{x}_k)$. The compatible pair is made up of the matched items, and the incompatible pairs of the item are randomly selected to the unsuitable outfits to form the negative sets. The relationship could be reflected by:

$$\mathbf{x}_{ij} = \sigma(w_{pi} \mathbf{x}_i + w_{pj} \mathbf{x}_j + \mathbf{b}_p), \quad (8)$$

where w_{pi} and w_{pj} are the weight matrices, and \mathbf{b}_p is a bias vector. The compatibility probability p_{ij} between items s_i and s_j are evaluated as,

$$p_{ij} = \varphi(w_{ij} \mathbf{x}_{ij} + \mathbf{b}_{ij}). \quad (9)$$

Similar to Eq. (7)–(8), w_{ij} is a weight matrix, and \mathbf{b}_{ij} is a bias vector as well. Besides, φ is an Sigmoid function as above. Let $d(\cdot)$ be an Euclidean distance function between a pair of items, we have an L_2 -norm measurement defined in the following:

$$d(\mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_d) = \|\mathbf{x}_i \odot \mathbf{w}_d - \mathbf{x}_j \odot \mathbf{w}_d\|_2^2. \quad (10)$$

Here, the weight \mathbf{w}_d is used to constrain the pair $\{\mathbf{x}_i, \mathbf{x}_j\}$, and the operator \odot denotes the component-wise multiplication. The relationship between the visual-semantic embedding in the multi-modal space is formulated as a triplet loss L_{vse} , and it could be optimized by:

$$L_{vse} = \max(0, m - d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_i, \mathbf{x}_k)). \quad (11)$$

It is easy to conclude that the distance between compatible items is relatively closer than that of incompatible items by a certain margin m . Owing to the coherent relationship between images and contextual information at the same item, a consistency loss L_t is set up in logarithmic terms as follows:

$$L_t = -\ln(\varphi(\mathbf{z}_i^T \mathbf{t}_i)) - \ln(\varphi(\mathbf{z}_j^T \mathbf{t}_j)) - \ln(\varphi(\mathbf{z}_k^T \mathbf{t}_k)). \quad (12)$$

3.4. Complementary fashion outfit matching

For the sake of making the best matching between the candidates, we employ a Bayesian Personalized Ranking (BPR) framework to raise the method's performance further. Generally, the BPR optimization function is defined in logarithmic terms as well, that is:

$$L_{bpr} = \sum_{(i,j,k) \in S} -\ln(\varphi(\mathbf{x}_{ij} - \mathbf{x}_{ik})). \quad (13)$$

Especially, j belongs to the compatible samples that denote the set of compatible item, and k is the set of incompatible items from the training samples. In summary, the total loss function $L(\cdot)$ of the solution is formulated as follows:

$$L(\lambda, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \theta) = \lambda_1 L_v + \lambda_2 L_t + \lambda_3 L_{vse} + \lambda_4 L_{bpr} + \frac{\lambda}{2} \|\theta\|^2, \quad (14)$$

where λ and λ_1 – λ_4 are the scalar parameters. The parameters of the model θ taken as the regularizer term is designed to avoid the over-fitting problem.

4. Dataset and features

4.1. Dataset

We conduct some experiments on Polyvore dataset in which data is collected from the fashion website polyvore.com. Different from general e-commerce websites, this website is released by fashion experts to publish their fashion outfits. The fashion items in the outfits contain abundant information, such as image, category, title, price, popularity, and so on. The apparel comprises three categories: upper body apparel (blouse, coat, etc.), lower body apparel (jeans, shorts, etc.), and full-body apparel (dress, jumpsuit, etc.). Two outfit samples from dataset are shown in Fig. 3.

The Polyvore dataset has 21,889 sets of outfits. The average number of items of the outfits is 8. We divide 70% into the training set, 10% into the validation set, and 20% into the test set. We discard the outfits that contain less than 3 items, or the items of outfit lack the textual information. So the dataset used in the experiment includes 17,316 for training samples, 3076 for testing samples.

4.2. Feature extraction

The efficientnet-b5 is used to extract visual features from the images of items, and the output features of its last convolutional layers are passed to the graph network. The graph of an outfit is constructed by the matching clothing which is selected by the fashion experts. In this paper, we applied the graph attention network to enhance the visual features of items and calculate the visual compatibility between items.

Considering the textual information, we followed the bert model to use the wordpiece tokenizer to split each text into language tokens [33]. The textual description has different semantics in the context, we adopted the bert model to initialize the parameters of architecture, which are used to encode the text information. Since the text



Fig. 3. An outfit is consisted of some items. Every item is described by image and text information about item's name and category.

datasets contain a lot of noise, we removed some words that appear less than five times. After filtering the data, there were 2,757 words left in the vocabulary. In order to keep with the consistent length for the textual information, the contextual information is truncated that the length of the text is greater than the fixed length, and the short text requires filling in the gaps.

5. Experiments

5.1. Experiment setting and training details

The outfit contains the image which was extracted by the convolutional neural network and aesthetic words that was extracted by the bert model according the word dictionary. The node of the graph is 512-dimensional feature vector. We adopt the Adam optimizer to optimize parameters. In addition, the propagation steps, hidden size and learning rate are set 3, 12, and 0.001, respectively. Due to the GPU memory limitation, the batch size is set as 16. Taking the complementary fashion item matching as an example, we use metric AUC to evaluate the model as Eq. (15):

$$AUC = \frac{1}{|S|} \sum_{(x_{i,j}, x_{i,k}) \in S} \delta(x_{i,j} > x_{i,k}), \quad (15)$$

where $\delta(\cdot)$ is an activate function.

5.2. Comparison with state-of-the-art methods

We employ two tasks to checkout the performance of CCMCK, one is the fashion compatibility task and the other is fill-in-the-blank to generate the compatible outfits [12]. At the same time, we assess the compatibility of outfits and fill the most compatible items to constitute the outfits. We evaluate the performance of the model by the FIFB accuracy and compatibility AUC metric, respectively. To estimate the proposed method, we make a comprehensive comparison with the advanced methods (LMT [34], SiameseCNNs [3], Bi-LSTM [16], GCN [18], NGNN [6] and MCAN [10]). The brief descriptions of the six methods are as follows,

- LMT: It is a method for the first to map images to the style space, and Mahalanobis distance is used to judge whether the items be going well together.
- SiameseCNNs¹: A mechanism of sharing weights for the feature vector is adopted in the network to estimate the distance between items in the style space. The model was the primary method to convert the image of items to the style space.

Table 1

The performance comparison of CCMCK with different methods.

Method	FIFB Acc. (%)	Compat. AUC (%)
LMT [34]	50.91	67.82
SiameseCNNs [3]	48.09	70.87
Bi-LSTM [16]	67.01	84.27
GCN [18]	76.41	91.40
NGNN [6]	78.89	97.22
MCAN [10]	86.50	96.00
CCMCK	87.75	97.67

- Bi-LSTM²: The whole fashion suit was regarded as a sequence and modeled in the form of a suit of this sequence. The bidirectional lstm memory network is used to learn fashion compatibility to recommend the compatible items.
- GCN³: A graph neural network is adopted to model the fashion graph to make outfits matching based on their context and the pairwise items.
- NGNN⁴: A fashion graph model with multi-modal information is adopted and achieved the compatibility score from the visual information view and textual information view.
- MCAN: In this model, fashion clothing is divided into tuples according to the multi-category to learn the visual-semantic compatibility among multiple tuples, finally the model recommended compatible clothing according to the user's category preference.

Table 1 shows the performance comparison of CCMCK with the advanced baselines in terms of AUC and FIFB metrics. In Table 1, our method obtains notable improvements that the outfit compatibility AUC is improved from 0.9600 to 0.9767. From this table, we infer the following conclusion, 1) Our method is superior to the advanced methods, which shows the effectiveness of CCMCK to generate a complementary clothing matching. 2) The performance of original baseline LMT is not good, the probable cause is the lack of the item's description, the textual description represents the high-level semantic features of fashion items. 3) NGNN and GCN significantly surpass all the item-based methods (i.e., LMT, SiameseCNNs, and BiLSTM), the methods based on graph network have the advantage on account of the context of outfits. 4) MCAN outperforms the NGNN and GCN, it was probably that MCAN can effectively summarize the key features of clothing by categories, such as patterns and styles. 5) CCMCK obtains better

¹ <https://github.com/mvasil/fashion-compatibility>.

² <https://github.com/xthan/polyvore>.

³ <https://github.com/CRIPAC-DIG/NGNN>.

⁴ <https://github.com/gcucurull/visual-compatibility>.

	Com.Score.	Tops	Down	Shoes	Watches	Bags	Accessories
1	0.9054						
2	0.8746						
3	0.8705						
4	0.8328						

Fig. 4. Compatible scores of the outfits with visual information.

Table 2

Ablation study for in our framework including the visual embedding module (CCMCK-V), text embedding module (CCMCK-C) and visual semantic embedding (CCMCK).

Method	Compat. AUC (%)	FIFB Acc. (%)
NGNN	96.58	78.89
GCN	91.40	76.41
CCMCK-V	89.52	85.04
CCMCK-C	83.16	83.28
CCMCK	97.67	87.75

performance than MCAN, it is a reason that CCMCK makes full use of graph neural networks to learn the visual feature and then combines the textual information to form a unified representation for matching complementary outfits.

5.3. Ablation study on component comparison

In order to check the effectiveness of each component in CCMCK, we perform an ablation study on different components of CCMCK. The experimental results are shown in Table 2, in which CCMCK-V is a visual feature module, CCMCK-C is a textual feature module, and CCMCK is a multi-modal embedding module.

We conclude that different graph model has different effects on the experimental results shown in Table 2. CCMCK has exhibited excellent evaluation performance by making the best of image information and textual information. It is observed from Table 2 that the performance of the graph neural network (i.e., GCN) is inferior to the performance of NGNN, and CCMCK-C shows the excellent performance in the outfit compatibility, the reason is that the textual information acts as a supplementary role to understand the semantic information of the context. CCMCK is superior compatibility for modules with textual information since CCMCK depends on text information as well as visual information. Additionally, the contextual data usually conveys some high-level semantic cues like the item brand.

5.4. Case study on complementary fashion outfit matching

For the purpose of matching complementary fashion outfits, we calculate the compatibility of the items in the outfit from a visual perspective. Fig. 4 shows the compatibility scores for the outfits with visual information. There are six parts in total, namely, tops, downs, shoes, watches, bags, and accessories.

We present the qualitative results of the outfits with visual information in Fig. 4. We observe that our method is capable of recommending similar products fairly well. In Fig. 4, the top outfit generated according to Mickey Mouse is a Mickey Mouse style outfit with the same topic and different patterns. The items with Mickey Mouse patterns are compatible and assembled into an outfit. In the second outfit, different from the Mickey Mouse style of the first outfit, the outfit regards the visual information of bear as the topic style, the items are suitable with bears style. In the third outfit, the items have a cat topic and make up the outfit with a consistent style. Although the shoes are the absence of cat pattern, however, the textual information is displayed on the shoes. Different from the others, the items recommended by CCMCK for the last outfit have a watermelon sign which belongs to the cool style.

5.5. Compatibility modeling of overcoat for given outfits

To assess the practical applicability of our model, we evaluate it on the compatibility of fashion outfits. Fig. 5, displays some qualitative instances illustrating the compatibility results for the outfits with multi-modal information, each line represents different types of items, and each category is compatible with each other. Our CCMCK is exhibited to perform six different recommendation outfits as shown in Fig. 5.

From Fig. 5 we can see that our method generates the multi-modal compatible outfits. For case A, CCMCK recommends a yellow overcoat to fit the outfits, the items are the faux fur style and compatible with a score of 0.9775. Case B shows that the fabric style coat is more suitable to the outfit and the outfit has a compatible score of 0.9657. A cardigan overcoat is recommended in case C by CCMCK to the outfit which has the score of 0.9896. And the overcoat is more suitable to go together with the outfit with coarse categories. For the case D, the outfit with leather style has a relatively high compatibility score of 0.9944. It is compatible with the pattern as well as color, which indicates that CCMCK is great a reliance upon multi-modal information. Due to the diversity of the outfit in case E, the compatible score is a valley. Nevertheless, in case F, CCMCK recommends the buff overcoat to the outfit and matches the Balenciaga embroidered organza shirt. This demonstrates that CCMCK is capable of generating satisfactory recommendations by considering the category requirements. All in all, the top four cases show that the compatible outfits keep the homogeneous style in multi-modal information. The last three cases display that the compatible outfits are leather-like apparel.

To further study the effect of compatibility knowledge in complementary clothing matching, we show the curve varying with the





	Com. Score.	Compatible Category					
		Overcoats	Tops	Downs	Shoes	Bags	Accessories
A	0.9775						
B	0.9657						
C	0.9896						
D	0.9944						
E	0.9572						
F	0.9831						

Fig. 5. The compatibility for the outfits that consist of visual information and textual information.

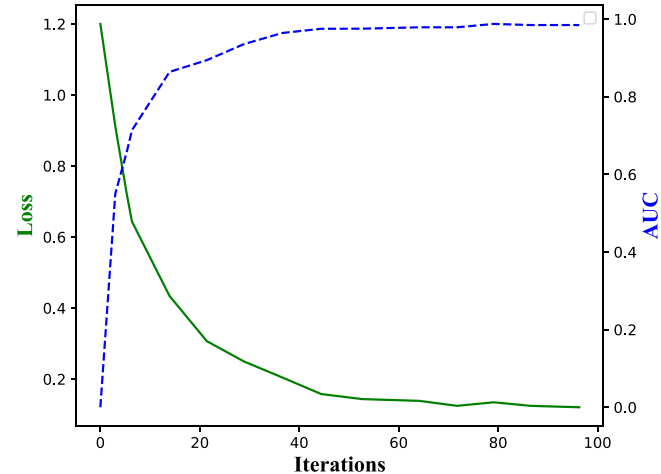


Fig. 6. Illustration of the training loss and AUC convergence of the CCMCK model.

number of iterations in Fig. 6, the blue line and the green line represents the AUC value and the loss value, respectively. It can be observed that CCMCK achieves great performance with respect to iterations on the AUC metrics. The loss value gradually decreases, while the value of ACU gradually increases, and finally, both of them get a stable value.

In order to fairness, CCMCK is compared with the baseline in the complementary clothing matching task, where the accuracy is regarded as the evaluation metric. As shown in Fig. 7, CCMCK obtains the best performance rather than all baselines. As the epoch increases, the accuracy of the model becomes more better and finally tends to be stable. CCMCK requires more epochs to achieve stable performance due to a large amount of data. On the other hand, MCAN has a better performance than GCN and NGNN for its clothing repartition via the multi-category tuples. GCN and NGNN have an outstanding superiority over Bi-LSTM, the reason is that graph neural network models the relationship between items according to the context. Bi-LSTM requires lesser epochs to achieve stable performance. It is noted that Bi-LSTM

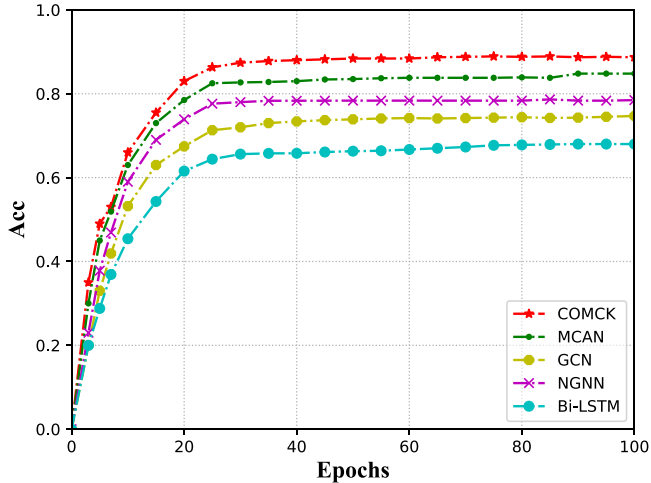


Fig. 7. Performance of different models with regard to accuracy in different epochs.

treats the relationships of items as a sequence, the next prediction depends on the previous items.

To make profound insights, we conduct some researches on clothing matching and replacement, where the items are randomly selected from the outfits and make a compatibility prediction as shown in Fig. 8. For the incompatible outfits, and we replace the garments from the outfits and predict the compatibility of outfits in the candidate items. The benefit of outfit diagnosis is that we can revise the outfit based on the diagnosis result. CCMCK attempts to discover the consistency between the textual information and the visual of the same fashion item, and the revised outfits show excellent compatibility with a consistent style.

Fig. 9 is the performance visualization of our model in the tasks of outfit matching. For each compatible outfit, we visualize instances of our method for the compatibility task. We list the compatibility of two sets of outfits and the revised outfits corresponding to the outfits, as well as make fashion compatibility analysis, and visualize the last four layers of the network. In the visualization graph, each graph represents



Fig. 8. The compatible scores of the revised outfits.

the compatibility score between items. We make outfit research for the top outfit, the women's embossed skater skirts are compatible with the brian printed cotton t-shirt, along with the deepening of the network, they are not together in the outfit, the outfit has high compatibility by replacing the printed cotton t-shirt with the white short sleeves. For the bottom outfit, the short sleeve top and the vintage metallic flared trouser match well, but the trouser is incompatible with other items. When we replace the vintage metallic flared trousers, the whole shirt has high compatibility that the outfit is compatible with cat-eye sunglasses.

6. Conclusions

In this paper, we propose a complementary clothing matching model based on multi-modal information. In the task of clothing compatibility matching, our model achieves superior performance to the other methods on Accuracy and AUC metrics. Ablation study on component comparison shows that the components in the model have different functions to predict clothing compatibility, the contextual data can convey some high-level semantic cues. For the purpose of matching complementary fashion outfits, we can calculate the compatibility of the items in the outfit from a partial visual information. The experiment results demonstrate that our method is capable of generating satisfactory recommendations by considering the category requirements. For the view of multi-modal, our model recommends suitable clothing from the different categories, which further verifies the effectiveness of multi-modal. The CCMCK effectively achieves a favorable performance on complementary clothing matching. In the future, we will further classify and refine the local features to achieve comprehensive compatibility prediction of clothing in the multi-scale.

CCRediT authorship contribution statement

Ruomei Wang: Conceptualization, Resources, Writing – review & editing. **Jianfeng Wang:** Methodology, Software, Writing – original draft. **Zhuo Su:** Methodology, Validation, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This research is supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011953) and the National Natural Science Foundation of China (No. 61872394).



Fig. 9. The visualization results of compatibility where showed the variations in compatibility between different outfits.

References

- [1] M. Jian, J. Guo, C. Zhang, T. Jia, L. Wu, X. Yang, L. Huo, Semantic manifold modularization-based ranking for image recommendation, *Pattern Recognit.* June (2021) 1–39.
- [2] Y. Lin, P. Ren, Z. Chen, Z. Ren, J.M.M. de Rijke, Explainable outfit recommendation with joint outfit matching and comment generation, *IEEE Trans. Knowl. Data Eng.* June (2019) 1–16.
- [3] M.I. Vasileva, B.A. Plummer, K. Dusad, S. Rajpal, R. Kumar, D.A. Forsyth, Learning type-aware embeddings for fashion compatibility, in: *ECCV*, Vol. 11220, Springer, 2018, pp. 405–421.
- [4] Y. Zhang, P. Zhang, C. Yuan, Z. Wang, Texture and shape biased two-stream networks for clothing classification and attribute recognition, *CVPR* (2020) 13535–13544.
- [5] A. Veit, B. Kovacs, S. Bell, J.J. McAuley, K. Bala, S.J. Belongie, Learning visual clothing style with heterogeneous dyadic co-occurrences, in: *ICCV*, IEEE Computer Society, 2015, pp. 4642–4650.
- [6] Z. Cui, Z. Li, S. Wu, X. Zhang, L. Wang, Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks, in: *WWW*, ACM, 2019, pp. 307–317.
- [7] Y. Lin, M. Moosaei, H. Yang, Outfitnet: Fashion outfit recommendation with attention-based multiple instance learning, in: *WWW*, ACM, 2020, pp. 77–87.
- [8] G. Cucurull, P. Taslakian, D. Vazquez, Context-aware visual compatibility prediction, *CVPR* (2019) 12617–12626.
- [9] D. Sagar, J. Garg, P. Kansal, S. Bhalla, R.R. Shah, Y. Yu, PAI-BPR: personalized outfit recommendation scheme with attribute-wise interpretability, in: *BigMM*, IEEE, 2020, pp. 221–230.
- [10] X. Yang, D. Xie, X. Wang, J. Yuan, W. Ding, P. Yan, Learning tuple compatibility for conditional outfit recommendation, in: *MM*, ACM, 2020, pp. 2636–2644.
- [11] X. Yang, Y. Ma, L. Liao, M. Wang, T.-S. Chua, TransNFCM: Translation-based neural fashion compatibility modeling, *AAAI-19* (2019) 403–410.
- [12] J. Wang, X. Cheng, R. Wang, S. Liu, Learning outfit compatibility with graph attention network and visual-semantic embedding, in: *ICME*, IEEE, 2021, pp. 1–6.
- [13] Y. Gao, N. Chang, K. Shang, Multi-layer and multi-order fine-grained feature learning for artwork attribute recognition, *Comput. Commun.* 173 (2021) 214–219.
- [14] K. Mao, G. Srivastava, R.M. Parizi, M.S. Khan, Multi-source fusion for weak target images in the industrial internet of things, *Comput. Commun.* 173 (2021) 150–159.
- [15] D. Verma, K. Gulati, R.R. Shah, Addressing the cold-start problem in outfit recommendation using visual preference modelling, in: *BigMM*, IEEE, 2020, pp. 251–256.
- [16] X. Han, Z. Wu, Y. Jiang, L.S. Davis, Learning fashion compatibility with bidirectional LSTMs, in: *MM*, ACM, 2017, pp. 1078–1086.
- [17] M. Lin, R. Ji, S. Chen, X. Sun, C. Lin, Similarity-preserving linkage hashing for online image retrieval, *IEEE Trans. Image Process.* 29 (2020) 5289–5300.
- [18] G. Cucurull, P. Taslakian, D. Vázquez, Context-aware visual compatibility prediction, in: *CVPR*, IEEE, 2019, pp. 12617–12626.
- [19] X. Yang, X. Song, F. Feng, H. Wen, L. Duan, L. Nie, Attribute-wise explainable fashion compatibility modeling, *ACM Trans. Multim. Comput. Commun. Appl.* 17 (1) (2021) 36:1–36:21.
- [20] H. Zhan, C. Yi, B. Shi, J. Lin, L. Duan, A.C. Kot, Pose-normalized and appearance-preserved street-to-shop clothing image generation and feature learning, *IEEE Trans. Multim.* 23 (2021) 133–144.
- [21] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, B. Zhao, POG: personalized outfit generation for fashion recommendation at alibaba ifashion, in: *SIGKDD*, ACM, 2019, pp. 2662–2670.
- [22] M. Dong, X. Zeng, L. Koehl, J. Zhang, An interactive knowledge-based recommender system for fashion product design in the big data environment, *Inform. Sci.* 540 (2020) 469–488.
- [23] W. Kang, E. Kim, J. Leskovec, C. Rosenberg, J.J. McAuley, Complete the look: Scene-based complementary product recommendation, in: *CVPR*, IEEE, 2019, pp. 10532–10541.
- [24] J. Liu, X. Song, Z. Chen, J. Ma, Neural fashion experts: I know how to make the complementary clothing matching, *Neurocomputing* 359 (2019) 249–263.
- [25] J. Liu, X. Song, L. Nie, T. Gan, J. Ma, An end-to-end attention-based neural model for complementary clothing matching, *ACM Trans. Multim. Comput. Commun. Appl.* 15 (4) (2020) 114:1–114:16.
- [26] G. Sun, J. He, X. Wu, B. Zhao, Q. Peng, Learning fashion compatibility across categories with deep multimodal neural networks, *Neurocomputing* 395 (2020) 237–246.
- [27] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4) (2017) 664–676.
- [28] X. Li, X. Wang, X. He, L. Chen, J. Xiao, T. Chua, SIGIR, ACM, 2020, pp. 159–168.
- [29] R. He, J.J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback, in: *AAAI*, AAAI Press, 2016, pp. 144–150.
- [30] X. Fu, T. Ouyang, Z. Yang, S. Liu, A product ranking method combining the features-opinion pairs mining and interval-valued pythagorean fuzzy sets, *Appl. Soft Comput.* 97 (Part B) (2020) 106803.
- [31] Z. Li, Z. Cui, S. Wu, X. Zhang, L. Wang, Semi-supervised compatibility learning across categories for clothing matching, in: *ICME*, IEEE, 2019, pp. 484–489.
- [32] G. Gao, L. Liu, L. Wang, Y. Zhang, Fashion clothes matching scheme based on siamese network and AutoEncoder, *Multim. Syst.* 25 (6) (2019) 593–602.
- [33] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, J. Fu, Seeing out of the box: End-to-end pre-training for vision-language representation learning, *CoRR* (2021) arXiv:2104.03135.
- [34] J.J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: *SIGIR*, ACM, 2015, pp. 43–52.