An End-to-End Attention-Based Neural Model for Complementary Clothing Matching

JINHUAN LIU, XUEMENG SONG, LIQIANG NIE, TIAN GAN, and JUN MA,
Shandong University

In modern society, people tend to prefer fashionable and decent outfits that can meet more than basic physiological needs. In fact, a proper outfit usually relies on good matching among complementary fashion items (e.g., the top, bottom, and shoes) that compose it, which thus propels us to investigate the automatic complementary clothing matching scheme. However, this is non-trivial due to the following challenges. First, the main challenge lies in how to accurately model the compatibility between complementary fashion items (e.g., the top and bottom) that come from the heterogeneous spaces with multi-modalities (e.g., the visual modality and textual modality). Second, since different features (e.g., the color, style, and pattern) of fashion items may contribute differently to compatibility modeling, how to encode the confidence of different pairwise features presents a tough challenge. Third, how to jointly learn the latent representation of multi-modal data and the compatibility between complementary fashion items contributes to the last challenge. Toward this end, in this work, we present an end-to-end attention-based neural framework for the compatibility modeling, where we introduce a feature-level attention model to adaptively learn the confidence for different pairwise features. Extensive experiments on a public available real-world dataset show the superiority of our model over state-of-the-art methods.

CCS Concepts: • **Information systems → Retrieval tasks and goals**; *World Wide Web*;

Additional Key Words and Phrases: End-to-end, feature-level attention, complementary clothing matching

ACM Reference format:
Jinhuan Liu, Xuemeng Song, Liqiang Nie, Tian Gan, and Jun Ma. 2019. An End-to-End Attention-Based Neural Model for Complementary Clothing Matching. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 4, Article 114 (December 2019), 16 pages.
https://doi.org/10.1145/3368071

1 INTRODUCTION

According to Plunkett Research, the global retail market for clothing and footwear has reached $1.8 trillion in 2018.[1] The great economic value indicates the increasingly important role of

[1] https://www.plunkettresearch.com/industries/apparel-shoes-textiles-market-research/.

This work was supported by the National Basic Research Program of China (973 Program), no. 2015CB352502; the National Natural Science Foundation of China, nos. 61772310, 61702300, 61702302, and 61672322; Young Scholars Program of Shandong University; and the Project of Thousand Youth Talents 2016.
Authors' address: J. Liu, X. Song (corresponding author), L. Nie, T. Gan, and J. Ma, School of Computer Science and Technology, Shandong University, Qingdao, 266237, China; emails: {liujinhuan.sdu, sxmustc, nieliqiang}@gmail.com, {gantian, majun}@sdu.edu.cn.

1551-6857/2019/12-ART114 $15.00
https://doi.org/10.1145/3368071

(a) Top1 and bottom1 (b) Top2 and bottom2

Fig. 1. Illustration of the contribution of different aspects for different fashion items.

clothing in people's daily lives. In fact, more and more people pursue fashionable and decent outfits, which directly leads to a headache for many individuals who need to seek advice from their professional friends or shopping guides about clothing matching. Therefore, devising an automatic clothing matching scheme to alleviate people from such inconvenience merits our attention.

In this work, we aim to investigate the problem of clothing matching. Without loss of generality, we take the bottom recommendation for a given top as an example.[2] Essentially, the clothing matching problem can be cast as the task of compatibility modeling between complementary fashion items. However, the compatibility modeling is non-trivial due to the following challenges. First, the main challenge lies in how to model the compatibility between fashion items of different categories (e.g., the top and bottom). In addition, the multiple modalities (i.e., the visual modality and textual modality) can describe the same fashion item from different perspectives. Consequently, how to enhance the compatibility measurement by exploiting the visual and textual metadata of fashion items contributes to a tough challenge. Second, the different features of the clothing may contribute differently to the compatibility modeling given the different fashion items. For example, as can be seen from Figure 1, the harmonious match between *top1* and *bottom1* is largely due to the consistent pattern, whereas the match between *top2* and *bottom2* should be attributed to the the compatible color. Therefore, how to adaptively assign the confidence of different features for different fashion items poses a crucial challenge. Last, how to jointly learn the latent representation of multi-modal data and the compatibility between different fashion items contributes to the last challenge.

To address these challenges, we devise a content-based multi-modal clothing matching framework, as shown in Figure 2. Given a top and a bottom along with their visual and textual metadata, we aim to learn the effective latent multi-modal representations of the fashion items, which would enable us to accurately capture the various aspects (e.g., patterns and categories) and measure the compatibility between fashion items. Moreover, different from previous work, we further take into consideration the confidence modeling of different features (i.e., the high-level features that extracted from the neural networks) in terms of clothing matching. In particular, we propose an end-to-end multi-modal deep neural network, which is capable of jointly learning the feature encoding of the multiple modalities and modeling the compatibility preference among fashion items of the complementary categories. Moreover, to distinguish the contributions of different pairwise features in compatibility modeling, we propose a feature-level attention model to adaptively assign the confidence of different features for different fashion items.

[2]The proposed method can also be applied to other scenarios, such as recommending the shoes for a given bottom.

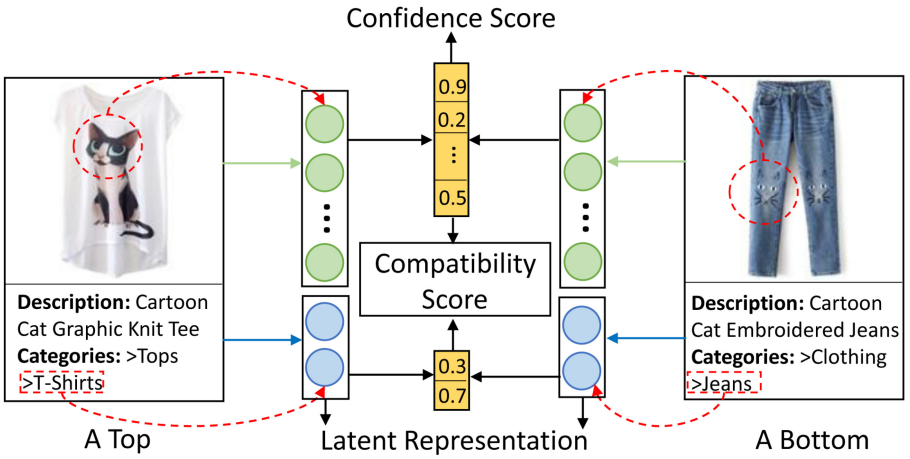

Fig. 2. Illustration of the main research task. Given the images, textual descriptions and categories of a top and a bottom, our model can capture various aspects of fashion items and learn the confidence score based on the contribution of various aspects, thus achieving compatible clothing matching.

The main contributions of this article can be summarized as follows:

- To the best of our knowledge, we are the first to introduce the feature-level attentive mechanism to model the contributions of different pairwise features of fashion items in the context of clothing matching.
- We propose an end-to-end multi-modal deep neural framework to model the compatibility between fashion items based on multi-modal data (i.e., visual and textual modalities).
- Experimental results on a real-world dataset demonstrate the effectiveness of our proposed framework compared with other state-of-the-art models. We have released our code and parameters[3] to facilitate other researchers to repeat our experiments and verify their proposed approaches.

The rest of the article is organized as follows. In Section 2, we review the related work on fashion analysis techniques and the attentive mechanism. Section 3 gives a detailed description of our proposed end-to-end attention-based neural model for complementary clothing matching and the optimization algorithm. Experimental results and discussions are presented in Section 4. Finally, we conclude the article and discuss future work in Section 5.

2 RELATED WORK

2.1 Fashion Analysis Techniques

In recent years, great efforts have been dedicated to the fashion domain, which mainly focuses on clothing style prediction [31, 35], clothing retrieval [14, 15, 30], clothing recommendation [11, 28], outfit composition [22, 35], and clothing matching [39, 42]. For instance, Li et al. [18] proposed a session-based fashion recommendation system to model the user's sequential behavior and capture the user's main purpose in the current session. This method only relies on user clicks or purchase activities without taking the content information of fashion items into consideration. In addition, Iwata et al. [12] utilized the photographs from a fashion magazine to tackle the problem of the top (bottom) recommendation for a given bottom (top). Similarly, Jagadeesh et al. [13]

[3] https://github.com/coderepositary/EAN.

proposed a large-scale visual recommendation system with street fashion images. Notably, the images in both datasets have complex and noisy backgrounds, which may hurt the performance of clothing recommendation. Toward this end, He and McAuley [9] investigated personalized recommendation with clean images crawled from Amazon.com and Tradesy.com. However, the authors only considered the visual information of fashion items. In fact, the textual metadata can also deliver important characteristics of fashion items. To bridge this gap, Ma et al. [31] proposed a fashion-oriented multi-modal deep learning–based model to classify clothing styles. Meanwhile, Song et al. [42, 43] studied the problem of compatibility modeling between fashion items with the multi-modal data of fashion items, where a latent space is learned to measure compatibility. Despite the success of this work, the authors separated the feature extraction and compatibility modeling, which may devastate the performance to a certain extent. Moreover, existing work overlooks the confidence of different features in compatibility modeling, which is a major concern of our work.

2.2 Attentive Mechanism

The attention mechanism aims to select more critical information from a multitude of information by focusing on the relevant parts of the input data [34]. Recently, attention-based methods have been widely applied in machine learning tasks, such as sentence classification [26], multimedia recommendation [2], information retrieval [49], and image caption generation [3]. For example, Bahdanau et al. [1] first introduced the attention mechanism to the encoder-decoder neural network in the context of machine translation, which is able to pay different attention to different words in the input sequence. Then, inspired by their work, Xu et al. [50] presented an attention-based model with the long short-term memory (LSTM) network to automatically generate captions for images. In addition, Chen et al. [2] proposed a novel attention mechanism within the collaborative filtering (CF) framework to model item- and component-level implicit feedback in multimedia recommendation. Although the attention mechanism has achieved tremendous success in various contexts, its potential in the fashion domain has not been well exploited. For example, Li et al. [18] introduced a novel hybrid encoder for session-based item recommendation, where the attention mechanism is used to model the user's sequential clicking behaviors and thus infer the user's purchase intention. Different from their work that focuses on the users' clicking behaviors, we tend to apply the attention mechanism to model the contents of fashion items in the context of clothing matching. In addition, as far as we know, most existing works treat the different features of items equally, which overlooks the different contributions of different features in compatibility modeling. To bridge this gap, we aim to employ the attention mechanism to distinguish the confidence of different pairwise features and thus accurately match the clothing.

3 PROBLEM FORMULATION

In this section, we detail the proposed end-to-end attention-based neural model (EAN), which is able to model compatibility between fashion items with multiple modalities while distinguishing the confidence of different pairwise features of different fashion items.

3.1 Notations

Let $\mathcal{T} = \{t_1, t_2, \ldots, t_{N_t}\}$ and $\mathcal{B} = \{b_1, b_2, \ldots, b_{N_b}\}$ stand for the set of tops and bottoms, where N_t and N_b stand for the total number of tops and bottoms, respectively. Each fashion item (e.g., a top) involves a 227×227 pixels image with three channels and certain textual description. Let $\mathcal{P} = \{(t_{i_1}, b_{j_1}), (t_{i_2}, b_{j_2}), \ldots, (t_{i_M}, b_{j_M})\}$ stand for the set of positive top-bottom pairs, where M is the number of top-bottom pairs. We use $\tilde{m}_{ij}$ to represent the compatibility between top t_i and bottom b_j. To accurately measure the $\tilde{m}_{ij}$, we propose a feature-level attention model to learn a

confidence $s_l(i, j)$ for the l-th pair of features of top t_i and bottom b_j. Furthermore, we exploit the visual information $v_i(v_j)$ and textual information $c_i(c_j)$ to enhance the compatibility score $\tilde{m}_{ij}$. In the following, we detail the proposed EAN model, which is the main contribution of this article.

3.2 Visual Feature Encoding

Images undoubtedly are the most informative signals of fashion items, characterizing their most intuitive features, such as the color and shape. To encode the visual features of fashion items, we employ the deep convolutional neural networks (CNNs) (e.g., AlexNet [17, 24], VGGNet [10, 40], GoogleNet [45, 46], and ResNet [8, 44]), which has proven to be effective in various computer vision tasks [19–21, 23, 36, 47, 48]. In particular, we encoded the fashion images with the AlexNet model [17], which is composed of five convolutional layers and three fully connected layers. Specifically, the AlexNet model has been pre-trained on 1.2 million images in the ImageNet LSVRC-2010 contest. We feed the fashion images into AlexNet and employ the output of the "FC7" layer as the visual feature encoding. Let $\mathbf{v}_{t_i} \in \mathbb{R}^{d_v}$ and $\mathbf{v}_{b_j} \in \mathbb{R}^{d_v}$ denote the visual encoding of top t_i and bottom b_j, respectively, where $d_v = 4,096$ is the dimension of the image encoding.

3.3 Textual Feature Encoding

The textual data of fashion items in the form of the categories and descriptions also delivers valuable information such as the material and category. To effectively encode the textual information, we first encode each word with a 300-dimensional vector using the word2vector vectors pre-trained on 100 billion words from Google News. Accordingly, each word of the fashion items in the pre-trained words can be represented as a 300-dimensional vector, whereas the words not present are randomly initialized. Moreover, we concatenate the word vectors of each fashion item and feed the word matrix to a CNN model [16], which consists of a convolutional layer and a max-pooling layer. We adopt k kernels of different sizes to encode the textual features of fashion items. In practice, we set $k = 4$ and the sizes of the kernels as 2, 3, 4, and 5. Formally, we use $\mathbf{c}_{t_i} \in \mathbb{R}^{d_c}$ and $\mathbf{c}_{b_j} \in \mathbb{R}^{d_c}$ to represent the textual feature encoding for top t_i and bottom b_j, respectively. $d_c = (k * 100)$ is the dimension of the textual feature encoding.

3.4 EAN for Compatibility Modeling

Apparently, it is not advisable to directly measure the compatibility between fashion items of different categories with the aforementioned feature encodings from heterogeneous spaces. In this work, we assume that there exists a latent space that can unify the heterogeneous fashion items (i.e., the tops and bottoms) with multiple modalities (i.e., the visual and textual modalities). In particular, we employ the multi-layer perceptron (MLP) [8] to model the semantic relation between fashion items of different categories. We take the mapping from the visual encoding of top t_i to the latent representation as an example. Suppose that the MLP consists of K hidden layers; the hidden representation can be defined as

$$\begin{aligned} \mathbf{h}_{t_i}^1 &= \sigma\left(\mathbf{W}_{vt}^1 \mathbf{v}_{t_i} + \mathbf{b}_t^1\right), \\ \mathbf{h}_{t_i}^k &= \sigma\left(\mathbf{W}_{vt}^k \mathbf{h}_{t_i}^{k-1} + \mathbf{b}_t^k\right),\ k = 2, \ldots, K, \end{aligned} \tag{1}$$

where $\mathbf{h}_{t_i}^k$ is the hidden representation and $\mathbf{W}_{vt}^k, \mathbf{b}_t^k, k = 1, 2, \ldots, K$ are the parameters of the k-th layer, where the subscripts v and t denote the "visual encoding" and the "top," respectively. In this article, $\sigma(.)$ stands for the sigmoid activation function. We define the latent visual representation $\hat{\mathbf{v}}_{t_i} = \{v_{t_i}^1, v_{t_i}^2, \ldots, v_{t_i}^d\}^T = \mathbf{h}_{t_i}^K \in \mathbb{R}^d$, where d is the dimension of the latent representation. In a similar manner, we can get $\hat{\mathbf{c}}_{t_i} \in \mathbb{R}^d$, $\hat{\mathbf{v}}_{b_j} \in \mathbb{R}^d$, and $\hat{\mathbf{c}}_{b_j} \in \mathbb{R}^d$ as the latent textual representation of top t_i, visual, and textual representations of bottom b_j.

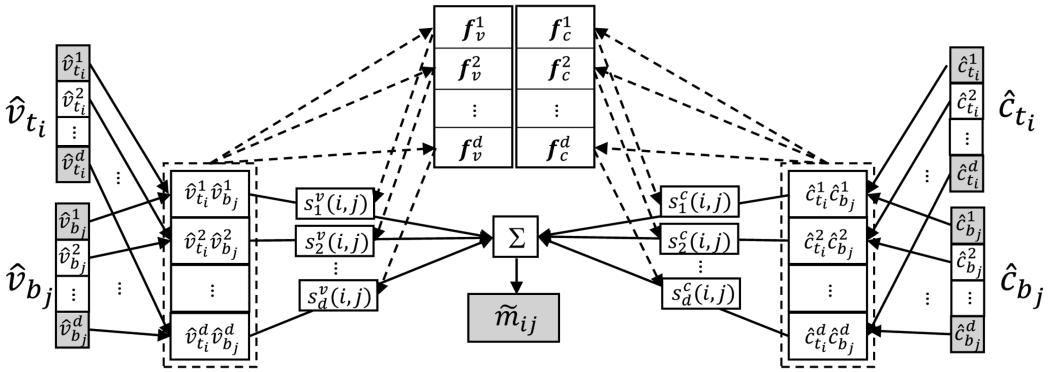

Fig. 3. Architecture of the feature-level attention model. The input is visual and textual feature encoding of tops and bottoms. This model can learn confidence for each pairwise feature of different categories of fashion items.

Traditionally, the compatibility between fashion items of different categories can be calculated by summing over all interactions between each pair of features with uniform confidence. Namely, for top t_i and bottom b_j, we have

$$m_{ij} = (\hat{\mathbf{v}}_{t_i})^T \hat{\mathbf{v}}_{b_j} + (\hat{\mathbf{c}}_{t_i})^T \hat{\mathbf{c}}_{b_j} = \sum_{l=1}^{d} \left(\hat{v}_{t_i}^l \hat{v}_{b_j}^l + \hat{c}_{t_i}^l \hat{c}_{b_j}^l \right), \tag{2}$$

where $\hat{v}_{t_i}^l$, $\hat{v}_{b_j}^l$, $\hat{c}_{t_i}^l$, and $\hat{c}_{b_j}^l$ represent the l-th features of the corresponding latent representations.

Nevertheless, different pairwise features of the latent representations may focus on characterizing different aspects (e.g., color, style, and pattern) of the fashions items and hence have different effects toward compatibility modeling. Moreover, even the same features can contribute differently for different fashion items. For example, the match between a "River Island White Print Halter Neck Crop Top" and a "River Island White Print Tassel Shorts" is mainly due to the compatible pattern features (i.e., River Island white print), whereas the match between a "Flora Denim Jacket" and a "Frayed Denim Pencil Skirt" can be attributed to material features. Accordingly, in this work, we assume that not all pairwise features contribute equally for compatibility modeling, and the confidence of each pairwise feature should be adaptively learned for different samples. In particular, we adopt the attention mechanism, which has proven to be effective in various computer vision tasks, such as personalized recommendation [38], image captioning [4, 27], and human action recognition [33, 41], and propose the feature-level attention model to learn the confidence of different features for different fashion items.

Figure 3 illustrates the architecture of the proposed feature-level attention model, which is able to adaptively learn the confidence of different pairwise features. We first define the confidence $q_l^v(t_i, b_j)$ for the l-th pairwise visual feature of the top-bottom pair (t_i, b_j) as follows:

$$q_l^v(t_i, b_j) = \mathbf{w}^T \phi\left(\mathbf{W}_{att}^v \left(\hat{\mathbf{v}}_{t_i} \odot \hat{\mathbf{v}}_{b_j} \odot \mathbf{f}_v^l \right) \right) + \mathbf{b}_a) + c, \tag{3}$$

where $\odot$ denotes the elementwise product of two vectors. $\mathbf{f}_v^l \in \mathbb{R}^d$ stands for the one-hot vector representation of the l-th pairwise feature, where the l-th element is 1 and the others are 0. $\mathbf{W}_{att}^v \in \mathbb{R}^{t \times d}$, $\mathbf{w} \in \mathbb{R}^t$, $\mathbf{b}_a \in \mathbb{R}^d$, and $c \in \mathbb{R}$ are the parameters of the attention network. t represents the number of the hidden units of the attention network. $\phi(.)$ stands for the activation function, where

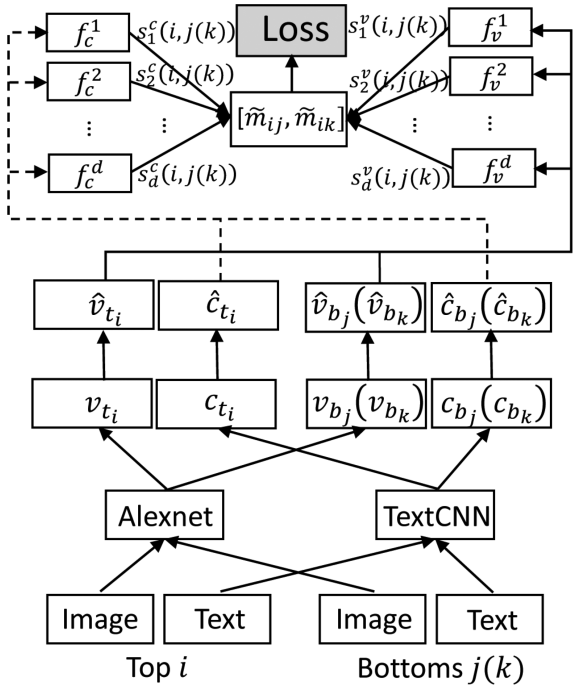

Fig. 4. Workflow of our proposed end-to-end attention-based neural model. We first encode the visual and textual information of tops and bottoms with the AlexNet and TextCNN model, respectively. Then, we employ the MLP to map the visual and textual feature encoding to the latent representation. After the feature mapping, we present a feature-level attention model to learn the confidence for different pairwise features. Ultimately, the compatibility score and loss are obtained for complementary clothing matching.

we adopt the ReLU function. Then, we normalize the l-th visual feature confidence as follows:

$$s_l^v(t_i, b_j) = \frac{exp(q_l^v(t_i, b_j))}{\sum_{l=1}^{d} exp(q_l^v(t_i, b_j))}. \tag{4}$$

In a similar fashion, we can derive the confidence $s_l^c(t_i, b_j)$ for the l-th feature of the textual encoding. We can thus re-define the compatibility $\tilde{m}_{ij}$ between top t_i and bottom b_j as follows:

$$\tilde{m}_{ij} = \sum_{l=1}^{d} \left(s_l^v(t_i, b_j) \hat{v}_{ti}^l \hat{v}_{bj}^l + s_l^c(t_i, b_j) \hat{c}_{ti}^l \hat{c}_{bj}^l \right). \tag{5}$$

By now, we have shown how to model compatibility between fashion items of different categories and assign confidence for the different pairwise features. We can thus proceed to introduce the proposed end-to-end attention-based neural model for clothing matching, which is shown in Figure 4.

Inspired by Song et al. [43], to thoroughly encode the compatibility preference between the tops and bottoms, we employ the Bayesian personalized ranking (BPR) [7, 37] model. In particular, we argue that the top-bottom pairs appearing in the same outfits (i.e., composed together by fashion experts) are more compatible than the unobserved ones. Accordingly, we build the training set:

$$\mathcal{M} := \{(i, j, k) | (t_i, b_j) \in \mathcal{P}, b_k \in \mathcal{B} \wedge (t_i, b_k) \notin \mathcal{P}\}, \tag{6}$$

ALGORITHM 1: EAN for Compatibility Modeling

Input: $\mathcal{M}, \mathbf{f}_v^l, \mathbf{f}_c^l$.
Output: Parameters Θ_{vc} and Θ_{att}.
Initialize parameters Θ_{vc} and Θ_{att};
repeat
  Draw (i, j, k) from $\mathcal{M}$;
  for *each* l **do**
    Compute $s_l^v(t_i, b_j)$, $s_l^c(t_i, b_j)$, $s_l^v(t_i, b_k)$, $s_l^c(t_i, b_k)$ according to Equations (3) and (4);
  end
  Construct the EAN according to Equations (5) and (7);
  Using a back-propagation strategy, update parameters Θ_{vc} and Θ_{att} according to Equation (11);
until *Converge*;
Return Θ_{vc} and Θ_{att}.

where the triplet (i, j, k) indicates that bottom b_j goes better with the given top t_i than bottom b_k. According to Rendle et al. [37], we define the loss function as follows:

$$\mathcal{L}_{BPR} = \sum_{(i,j,k)\in\mathcal{M}} -ln(\sigma(\tilde{m}_{ij} - \tilde{m}_{ik})) + \frac{\mu}{2}\left(\|\Theta_{vc}\|_F^2 + \|\Theta_{att}\|_F^2\right), \tag{7}$$

where Θ_{vc} and Θ_{att} refer to the parameters of the end-to-end multi-modal deep neural network and the feature-level attention network, respectively. μ is the non-negative hyper-parameter to control the strength of the regularization item. It is worth noting that $\tilde{m}_{ik}$ can be derived in a similar manner as $\tilde{m}_{ij}$.

3.5 Optimization

The general optimization procedure of our proposed EAN model is shown in Algorithm 1. To optimize the loss in Equation (7), we adopt the back-propagation strategy, where the main procedure lies in the calculation of the partial derivative of $\mathcal{L}_{BPR}$ with respect to each parameter in Θ_{att} and Θ_{vc}. Here, we only show the derivation of $\partial\mathcal{L}_{BPR}/\partial\mathbf{W}_{vt}^k$ and $\partial\mathcal{L}_{BPR}/\partial\mathbf{W}_{att}^v$ as the examples, whereas the other parameters can be obtained in a similar manner.

First, it is easy to obtain $\partial\mathcal{L}_{BPR}/\partial\mathbf{W}_{vt}^K$ as follows:

$$\frac{\partial\mathcal{L}_{BPR}}{\partial\mathbf{W}_{vt}^K} = -\sigma(\tilde{m}_{ik} - \tilde{m}_{ij})\sum_{l=1}^{d} q_l^v(t_i, b_j)\hat{v}_{bj}^l \frac{\partial\hat{v}_{t_i}^l}{\partial\mathbf{W}_{vt}^K} + \frac{\mu}{2}, \tag{8}$$

where $\partial\hat{v}_{t_i}^l/\partial\mathbf{W}_{vt}^K$ can be derived from $\hat{\mathbf{v}}_{t_i} = \sigma(\mathbf{W}_{vt}^K\mathbf{h}_{t_i}^{K-1} + \mathbf{b}_t^K)$. Then, we can iteratively obtain $\partial\mathcal{L}_{BPR}/\partial\mathbf{W}_{vt}^k, k = K-1, \ldots, 1$ using the chain rule.

Second, we can calculate $\partial\mathcal{L}_{BPR}/\partial\mathbf{W}_{att}^v$ as follows:

$$\frac{\partial\mathcal{L}_{BPR}}{\partial\mathbf{W}_{att}^v} = -\sigma(\tilde{m}_{ik} - \tilde{m}_{ij})\left(\sum_{l=1}^{d}\hat{v}_{t_i}^l\hat{v}_{b_j}^l\frac{\partial s_l^v(t_i, b_j)}{\partial\mathbf{W}_{att}^v} + \sum_{l=1}^{d}\hat{v}_{t_i}^l\hat{v}_{b_k}^l\frac{\partial s_l^v(t_i, b_k)}{\partial\mathbf{W}_{att}^v}\right) + \frac{\mu}{2}, \tag{9}$$

where $\partial s_l^v(t_i, b_j)/\partial\mathbf{W}_{att}^v$ and $\partial s_l^v(t_i, b_k)/\partial\mathbf{W}_{att}^v$ can be calculated according to Equation (5). Here, we take $\partial s_l^v(t_i, b_j)/\partial\mathbf{W}_{att}^v$ as an example. For simplicity, we define $U = exp(q_l^v(t_i, b_j)), V = \sum_{l=1}^{d} exp(-q_l^v(t_i, b_j))$. Then, we can access $\partial s_l^v(t_i, b_j)/\partial\mathbf{W}_{att}^v$ as follows:

$$\frac{\partial s_l^v(t_i, b_j)}{\partial\mathbf{W}_{att}^v} = \frac{\frac{\partial U}{\partial\mathbf{W}_{att}^v}V - \frac{\partial V}{\partial\mathbf{W}_{att}^v}U}{V^2}, \tag{10}$$

where

$$\begin{cases} \dfrac{\partial U}{\partial \mathbf{W}_{att}^{v}} = exp(q_l^v(t_i, b_j)) \dfrac{\partial q_l^v(t_i, b_j)}{\partial \mathbf{W}_{att}^{v}}, \\ \dfrac{\partial V}{\partial \mathbf{W}_{att}^{v}} = -\sum_{l=1}^{d} exp(-q_l^v(t_i, b_j)) \dfrac{\partial q_l^v(t_i, b_j)}{\partial \mathbf{W}_{att}^{v}}. \end{cases} \tag{11}$$

Finally, the parameters of our proposed model can be updated as follows:

$$\begin{cases} \mathbf{W}_{vt}^{k} \leftarrow \mathbf{W}_{vt}^{k} - \eta \left(\dfrac{\partial \mathcal{L}_{BPR}}{\partial \mathbf{W}_{vt}^{k}} \right), \\ \mathbf{W}_{att}^{v} \leftarrow \mathbf{W}_{att}^{v} - \eta \left(\dfrac{\partial \mathcal{L}_{BPR}}{\partial \mathbf{W}_{att}^{v}} \right), \end{cases} \tag{12}$$

where η is the learning rate.

4 EXPERIMENTS

To comprehensively verify the effectiveness of our proposed EAN model, we conduct extensive experiments to answer the following questions:

- **RQ1:** Is EAN superior to state-of-the-art methods?
- **RQ2:** How does the proposed feature-level attention model affect the performance?
- **RQ3:** What is the effect of the number of negative examples on the model performance?

4.1 Dataset

All experiments are conducted on the publicly accessible dataset FashionVC [43], which consists of 20,726 outfits with 14,870 tops and 13,662 bottoms. Each fashion item involves visual and contextual multi-modal data, including the visual image, the categories, and the fashion item description. Because the textual data is generated by users and may be noisy, we sanitized the noisy textual data using the following steps:

(1) We removed the words appearing fewer than 10 times.
(2) We filtered out the stop words and obtained a vocabulary of size 12,706.
(3) To unify the length of text, we used the zero padding method and finally retained 43 words for each fashion item.

4.2 Experimental Setting

We adopt 10-fold cross-validation to evaluate the proposed model, where we randomly split the dataset of the positive top-bottom pairs into 10 parts, choose 8 parts for the training, 1 part for the validation, and 1 part for the testing in turn. The validation set is used for tuning the hyper-parameters, and all experimental comparisons in this article are performed on the test set. According to Equation (6), for each positive top-bottom pair (t_i, b_j), we generate the triplets by randomly sampling $S = 3$ negative bottoms' b_ks, where bottom b_k has not been paired with top t_i in the dataset. To calculate m_{ij} and m_{ik} equally, the positive bottom and the negative bottom share the same weights, when passing the end-to-end multi-modal deep neural network. In terms of the evaluation method, we adopt the area under the ROC curve (AUC), which has been widely used in various recommendation systems [25, 29].

To optimize the objective function, we adopt stochastic gradient decent (SGD) [5]. In addition, we adopt the grid search strategy to determine the optimal values of hyper-parameters, where we tested the batch size of $[50, 100, 150, 300]$, the learning rate (i.e., η) of $[0.01, 0.05, 0.1, 0.5]$, the

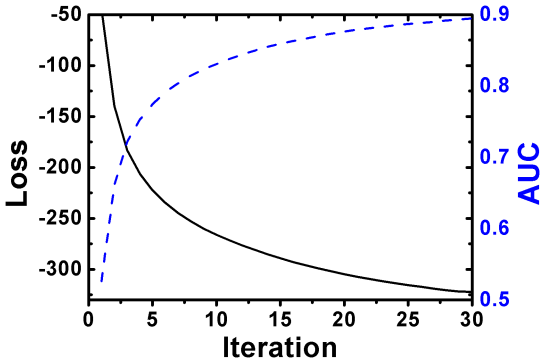

Fig. 5. Training loss and AUC of each epoch comparison of our proposed EAN model.

number of hidden units and attention layer units of [64, 128, 256, 512, 1, 024], and the regularization parameter (i.e., μ) of [0.001, 0.01, 0.1]. In addition, the proposed model was fine tuned based on the training set and the validation set for 30 iterations.

We first verify the convergence of our proposed model. We show the cures of the training loss in Equation (7) (black solid line) and the training AUC (blue dashed line) in Figure 5. We can see that the two values of different models first change sharply within a few iterations and then tend to be stable, which well validates the convergence of our model.

4.3 On Model Comparison (RQ1)

To validate the effectiveness of EAN, we first compare our model with the following baselines in the context of the compatibility modeling between tops and bottoms:

- **POP:** We measure the compatibility between top t_i and bottom b_j by the "popularity" of bottom b_j. The "popularity" refers to the number of bottoms matched with the top in the training set.
- **Bi-LSTM:** The bidirectional LSTM (Bi-LSTM) method [6] is designed to model outfit compatibility by exploring the sequential relationships among fashion items in an outfit. In our context, we adapt this method to deal with outfits consisting of only two items (i.e., the top and bottom) and replace the visual and textual encoders with our AlexNet and TextCNN modules, respectively.
- **IBR:** This complementary item recommendation method [32] models the relationships between objects with a latent style space, which is learned purely based on visual information. For the sake of fairness, we adapt IBR to an end-to-end model that would automatically learn the representation of fashion items with the CNNs instead of pre-extract the visual features with the Caffe network.
- **IBR-VC:** We extend IBR to enable it to measure the distance between fashion items with both visual and textual information. Similar to the preceding adaptation, we can derive IBR-VC from IBR with the visual feature encoding and textual feature encoding used in this work.
- **EXBPR-DAE:** We choose the content-based clothing matching scheme BPR-DAE proposed in Song et al. [43], which focuses on jointly modeling the coherent relation between different modalities of fashion items and the implicit preferences among them. Similarly, we transform BPR-DAE to EXBPR-DAE following an end-to-end fashion. The main difference

Table 1. Performance Comparison of Different Models

Approaches	AUC	p Value
POP	0.4332	$1.82E-13$
Bi-LSTM	0.5596	$2.51E-11$
IBR	0.5884	$1.20E-11$
IBR-VC	0.6809	$7.23E-09$
EXBPR-DAE	0.7680	$3.76E-03$
EAN	**0.7859**	–

between EXBPR-DAE and our EAN lies in that we consider the confidence of different pairwise features.

To evaluate the effectiveness of our proposed EAN, we adopt the paired t-test to perform significance analysis. Table 1 shows the performance comparison among different approaches. From this table, we have the following observations:

(1) EAN outperforms all state-of-the-art methods with a value of $p < 0.01$, which demonstrates significant improvement of our model over the baselines in the context of compatibility modeling between fashion items of different categories.
(2) EAN shows superiority over the end-to-end neural-based method EXBPR-DAE, which also exploits the multi-modal data of fashion items. This confirms to the assumption that different latent pairwise features of fashion items have different confidence in compatibility modeling.
(3) Bi-LSTM performs worse than the other content-based baselines, which may be attributed to the fact that the compatibility between the top and bottom pairs can be well modeled by the pairwise relationship rather than the sequential one.
(4) EAN, EXBPR-DAE, and IBR-VC all perform better than the image-based recommendation method IBR, which indicates that the textual modality also plays an important role in characterizing the fashion items and that compatibility modeling can benefit from exploiting the multiple modalities of fashion items.
(5) It is no surprise that POP achieves the worst performance, as POP overlooks the content information (i.e., visual information and textual information) that may describe the important characteristics of fashion items, such as the colors, patterns, and categories.

4.4 On the Feature-Level Attention Model (RQ2)

To evaluate the contribution of the proposed feature-level attention model, we compared EAN with its derivation EAN-E, where we disable the feature-level attention component by uniformly assigning the confidence to different pairwise features. To comprehensively verify the effectiveness of the proposed attention mechanism, we compare EAN-E and EAN with different modality combinations: the pure visual modality (V), the pure textual modality (T), and the multiple modalities (T+V).

Table 2 shows the comparison between EAN-E and EAN with different modality combinations. From this table, we can observe the following:

(1) EAN significantly outperforms EAN-E with all of the different modality combinations, which well validates the advantages of the proposed feature-level attention model in

Table 2. Performance Comparison on EAN-E and EAN with Different Modality Combinations

Approaches	T	V	T+V
EAN-E	0.7009	0.7252	**0.7461**
EAN	0.7277	0.7336	**0.7859**
p Value	$7.78E-06$	$2.19E-03$	$7.42E-07$

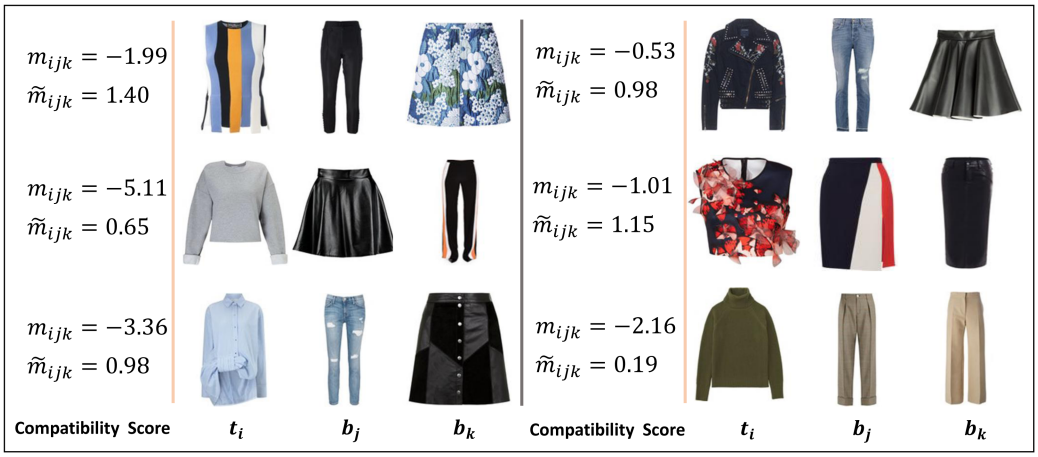

Fig. 6. Illustration of the impact of our proposed feature-level attention model. We annotate the compatibility scores learned by EAN and EAN-E as m_{ijk} and $\tilde{m}_{ijk}$, respectively. The $m_{ijk} = m_{ij} - m_{ik}$ and $\tilde{m}_{ijk}$ can be calculated in a similar manner.

compatibility modeling. This also confirms the initial assumption that the confidence of different pairwise features should be adaptively assigned rather than be equally treated.

(2) Both EAN and EAN-E with multiple modalities show superiority over that with any single modality, which demonstrates that both textual and visual modalities contribute to the compatibility modeling again.

(3) EAN-E and EAN with visual modality perform better than with textual modality. This may be due to the complex factors that affecting complementary clothing matching can be reflected in visual information (e.g., the color and pattern) more than textual information (e.g., the material and category).

To fully understand the impact of the proposed feature-level attention model, we illustrate the comparison between EAN and EAN-E with several examples in Figure 6. As can be seen, top t_i seems to be compatible with negative bottom b_k (i.e., $m_{ijk} < 0$) when the feature-level attention model is not considered. However, if we further take the different contributions of different pairwise features into consideration, we can find that top t_i matches bottom b_j better than negative bottom b_k.

To further illustrate how feature confidence affects compatibility modeling, we visualize the confidence of different features with several examples in Figure 7. It is worth noting that due to limited space, we only show the confidence of the first 10 dimensions of the 1,024. From this figure, we can gain the following insights:

(1) Interestingly, on the one hand, we found that the distributions of feature confidence learned by our model for similar bottom candidates are similar. For example, positive

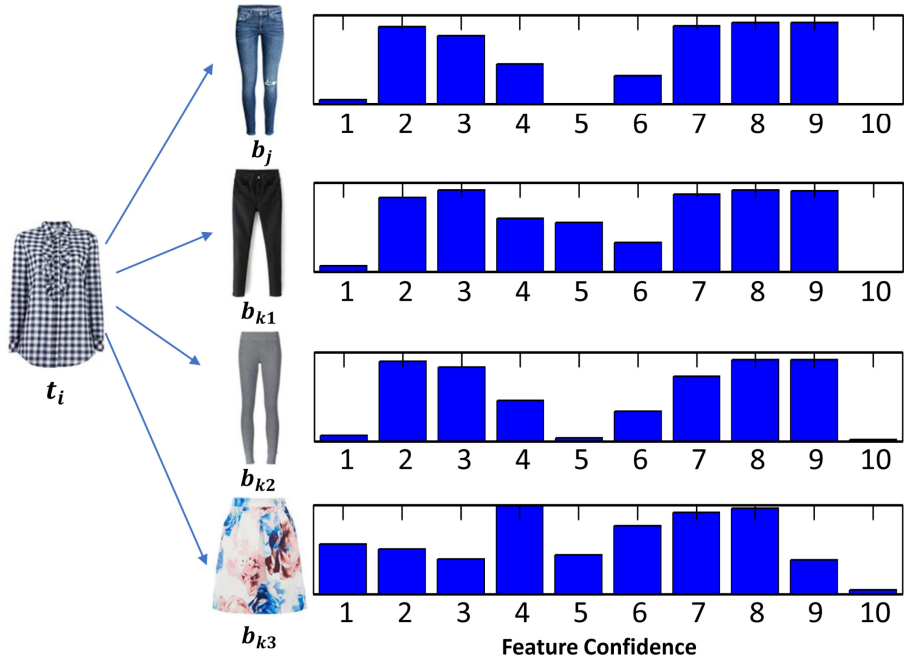

Fig. 7. Visualization of the feature confidences learned for different features of fashion items. Each row represents the feature confidence between top t_i and the corresponding bottom. The height of the blue bar indicates the value of the feature confidence. In addition, bottom b_j is the positive sample for top t_i, whereas the bottoms b_{k1}, b_{k2}, and b_{k3} are the negative ones randomly sampled from the dataset. Due to limited space, we only show the first 10 dimensions of feature confidence.

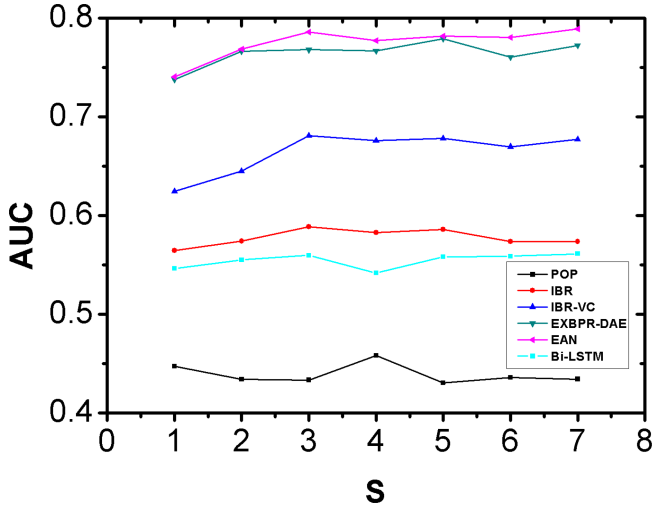

Fig. 8. AUC measure with respect to different numbers of negative examples of different models.

bottom b_j and negative bottom b_{k2} are both skinny trousers, and their confidence distributions for the top 10 features are similar.

(2) On the other hand, we noticed that feature confidences of bottoms with different categories (e.g., bottom b_j and bottom b_{k3}) vary greatly. This may be because fashion items with different categories have distinct visual or textural features.

(3) The confidence of different features are indeed different, which verifies the necessity of feature-level attention.

4.5 On the Number of Negative Examples (RQ3)

To evaluate the usefulness of EAN on a real-world dataset, where the proportion of the positive and negative top-bottom pairs can be diverse, we conducted experiments on different proportions of the negative samples. According to Equation (5), we randomly sampled S negative bottoms b_k to construct the triplets. Figure 8 shows the performance comparison of different models with respect to different numbers of negative examples, where we set $S = \{1, 2, \ldots, 7\}$. On the one hand, we find that our model EAN consistently achieves better performance compared to the baselines, which verifies the effectiveness of our model in real-world applications. On the other hand, overall, the proposed EAN model is not sensitive to the different proportions of negative examples.

5 CONCLUSION

In this work, we present an end-to-end attention-based neural model for complementary clothing matching. In particular, we present an end-to-end multi-modal deep neural framework to jointly learn the feature encodings of multiple modalities (i.e., the visual modality and textual modality) and map the different categories of fashion items into a unified compatibility space. Moreover, we propose a feature-level attention model to learn the confidence for different pairwise features of latent representations. We tested the proposed model on a real-world dataset. The experiments reveal that the proposed EAN outperforms state-of-the-art methods. The results show that the proposed feature-level attention model can effectively assign confidence for the pairwise features. We also obtained some interesting insights. For example, the distributions of the feature confidence for the similar bottom candidates are similar, whereas the bottoms of different categories vary greatly. As for future work, we plan to incorporate generative adversarial networks to generate clothing images and improve performance.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

[2] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*. 335–344.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 6298–6306.

[4] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2 (2018), 48.

[5] Xin Dong, Lei Yu, Zhonghuo Wu, Yuxia Sun, Lingfeng Yuan, and Fangxi Zhang. 2017. A hybrid collaborative filtering model with deep structure for recommender systems. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 1309–1315.

[6] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning fashion compatibility with bidirectional LSTMS. In *Proceedings of the ACM International Conference on Multimedia*. 1078–1086.

[7] Jing He, Xin Li, Lejian Liao, Dandan Song, and William K. Cheung. 2016. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 137–143.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 770–778.

[9] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 144–150.

[10] Xiangteng He and Yuxin Peng. 2017. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 4075–4081.

[11] Yang Hu, Xi Yi, and Larry S. Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the ACM International Conference on Multimedia*. 129–138.

[12] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 1. 2.

[13] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. 1925–1934.

[14] Shuhui Jiang, Yue Wu, and Yun Fu. 2018. Deep bidirectional cross-triplet embedding for online clothing shopping. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2018), Article 5.

[15] Yu-Gang Jiang, Minjun Li, Xi Wang, Wei Liu, and Xian-Sheng Hua. 2018. DeepProduct: Mobile product search with portable deep features. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2 (2018), 50.

[16] Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv:1408.5882.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems*. 1097–1105.

[18] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1419–1428.

[19] Xuelong Li, Lichao Mou, and Xiaoqiang Lu. 2016. Semantic video parsing by combining frame relevance and label propagation from images. *Multimedia Tools and Applications* 75, 19 (2016), 11961–11976.

[20] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. 2017. A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing* 26, 8 (2017), 3652–3664.

[21] X. Li, B. Zhao, and X. Lu. 2017. Key frame extraction in the summary space. *IEEE Transactions on Cybernetics* 48, 6 (2017), 1923–1934.

[22] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia* 19, 8 (2017), 1946–1955.

[23] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. 2018. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European Conference on Computer Vision*. 732–747.

[24] Shaohui Lin, Rongrong Ji, Chao Chen, and Feiyue Huang. 2017. ESPACE: Accelerating convolutional neural networks via eliminating spatial and channel redundancy. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 1424–1430.

[25] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2018. Explainable fashion recommendation with joint outfit matching and comment generation. arXiv:1806.08977.

[26] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. arXiv:1703.03130.

[27] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L. Yuille. 2017. Attention correctness in neural image captioning. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 4176–4182.

[28] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear! In *Proceedings of the ACM International Conference on Multimedia*. 619–628.

[29] Yong Liu, Peilin Zhao, Aixin Sun, and Chunyan Miao. 2015. A boosting algorithm for item recommendation with implicit feedback. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 15. 1792–1798.

[30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1096–1104.

[31] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. 2017. Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 38–44.

[32] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*. 43–52.

[33] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. 2017. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 1 (2017), 6.

[34] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In *Proceedings of the International Conference on Neural Information Processing Systems*. 2204–2212.

[35] Takuma Nakamura and Ryosuke Goto. 2018. Outfit generation and style extraction via bidirectional LSTM and autoencoder. arXiv:1807.03133.

[36] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the AAAI International Conference on Artificial Intelligence*. 2793–2799.

[37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*. 452–461.

[38] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the ACM International Conference on Recommender Systems*. 297–305.

[39] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. 2017. Compatibility family learning for item recommendation and generation. arXiv:1712.01262.

[40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[41] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, Vol. 1. 4263–4270.

[42] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. arXiv:1805.00313.

[43] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural compatibility modeling for clothing matching. In *Proceedings of the ACM International Conference on Multimedia*. 753–761.

[44] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, Vol. 4. 12.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1–9.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[47] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. 2018. Structure-aware multimodal feature fusion for RGB-D scene classification and beyond. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2s (2018), Article 39.

[48] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 5005–5013.

[49] Chenyan Xiong, Jimie Callan, and Tie-Yen Liu. 2017. Learning to attend and to rank with word-entity duets. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, Vol. 763. 772.

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

Received November 2018; revised March 2019; accepted September 2019