# The Recommendation of Social-Contextual Images Using a Hierarchical Attention Model

[1]Anishetty Shiva Rama Krishna,Associate Professor,
[2]Chandeeshwar Gunde,Assistant Professor,
[3]Firdose Fathima,Associate Professor,
Department of CSE Engineering,
Pallavi Engineering College,
Kuntloor(V),Hayathnagar(M), Hyderabad,R.R.Dist.-501505.

**Abstract**—*Image based social networks are among the most popular social networking services in recent years. With tremendous images uploaded everyday, understanding users' preferences on user-generated images and making recommendations have become an urgent need. In fact, many hybrid models have been proposed to fuse various kinds of side information (e.g., image visual representation, social network) and user-item historical behavior for enhancing recommendation performance. However, due to the unique characteristics of the user generated images in social image platforms, the previous studies failed to capture the complex aspects that influence users' preferences in a unified framework. Moreover, most of these hybrid models relied on predefined weights in combining different kinds of information, which usually resulted in sub-optimal recommendation performance. To this end, in this paper, we develop a hierarchical attention model for social contextual image recommendation. In addition to basic latent user interest modeling in the popular matrix factorization based recommendation, we identify three key aspects (i.e., upload history, social influence, and owner admiration) that affect each user's latent preferences, where each aspect summarizes a contextual factor from the complex relationships between users and images. After that, we design a hierarchical attention network that naturally mirrors the hierarchical relationship (elements in each aspects level, and the aspect level) of users' latent interests with the identified key aspects. Specifically, by taking embeddings from state-of-the-art deep learning models that are tailored for each kind of data, the hierarchical attention network could learn to attend differently to more or less content. Finally, extensive experimental results on real-world datasets clearly show the superiority of our proposed model.*

## 1   INTRODUCTION

There is an old saying "a picture is worth a thousand words". When it comes to social media, it turns out that visual images are growing much more popularity to attract users [14]. Especially with the increasing adoption of smartphones, users could easily take qualified images and upload them to various social image platforms to share these visually appealing pictures with others. Many image-based social sharing services have emerged, such as *Instagram*[1], *Pinterest*[2], and *Flickr*[3]. With hundreds of millions of images uploaded everyday, image recommendation has become an urgent need to deal with the image overload problem. By providing personalized image suggestions to each active user in image recommender system, users gain more satisfaction for platform prosperity. E.g., as reported by *Pinterest*, image recommendation powers over 40% of user engagement of this social platform [30].

Naturally, the standard recommendation algorithms provide a direct solution for the image recommendation task [2]. For example, many classical latent factor based Collaborative Filtering (CF) algorithms in recommender systems could be applied to deal with user-image interaction matrix [26], [40], [26]. Successful as they are, the
extreme data sparsity of the user-image interaction behavior
 limits the recommendation performance [2], [26]. On one hand
,

 some recent works proposed to enhance recommendation performance with visual contents learned from a (pre-trained) deep neural network [18], [49], [5]. On the other hand, as users perform image preferences in social platforms, some social based recommendation algorithms utilized the social influence among users to alleviate data sparsity for better recommendation [33], [24], [3]. In summary, these studies partially solved the data sparsity issue of social-based image recommendation. Nevertheless, the problem of how to better exploit the unique characteristics of the social image platforms in a holistical way to enhance recommendation performance is still under explored.

In this paper, we study the problem of understanding users' preferences for images and recommending images in social image based platforms. Fig. 1 shows an example of a typical social image application. Each image is associated with visual information. Besides showing likeness to images, users are also creators of these images with the upload behavior. In addition, users connect with others to form a social network to share their image preferences. The rich heterogeneous contextual data provides valuable clues to infer users' preferences to images. Given rich heterogeneous contextual data, the problem of how to summarize the heterogeneous *social contextual aspects* that influence users' preferences to these highly subjective content is still unclear. What's more, in the preference decision process, different users care about different social contextual aspects for their personalized image preference. E.g. *Lily* likes images that are similar to her uploaded images, while *Bob* is easily swayed by social neighbors to present similar preference as her
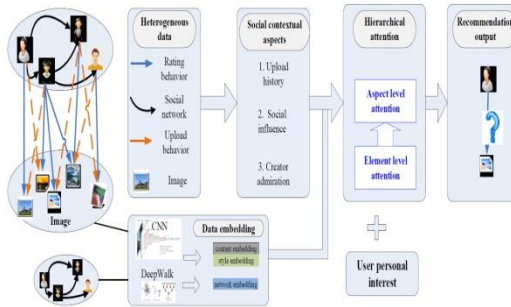
Fig. 1. An overall framework of social contextual image recommendation,
 where the left part shows the data characteristics of the
platform, and the right part shows our proposed model.

social friends. In other words, the unique user preference for balancing these complex social contextual aspect makes the recommendation problem more challenging.

To address the challenges mentioned above, in this paper, we design a hierarchical attention model for social image recommendation. The proposed model is built on the popular latent factor based models, which assumes users and items could be projected in a low latent space [34]. In our proposed model, for each user, in addition to basic latent user interest vector, we identify three key aspects (i.e., upload history, social influence and owner admiration) that affect each user's preference, where each aspect summarizes a contextual factor from the complex relationships between users and images. Specifically, the upload history aspect summarizes each user's uploaded images to characterize her interest. The social influence aspect characterizes the influence from the social network structure, and the owner admiration aspect depicts the influence from the uploader of the recommended image. The three key aspects are combined to form the auxiliary user latent embedding. Furthermore, since not all aspects are equally important for personalized image recommendation, we design a hierarchical attention structure that attentively weight different aspects for each user's auxiliary embedding. The proposed hierarchical structure aims at capturing the following two distinctive characteristics. First, as social contextual recommendation naturally exhibits the hierarchical structure (various elements from each aspect, and the three aspects of each user), we likewise construct user interest representation with a hierarchical structure. In the hierarchical structure, we first build auxiliary aspect representations of each user, and then aggregate the three aspect representations into an auxiliary user interest vector. Second, as different elements within each aspect, and different aspects are differentially informative for each user in the recommendation process, the hierarchical attention network builds two levels of attention mechanisms that apply at the element level and the aspect level.

We summarize the contributions of this paper as follows:

1)  We study the problem of image recommendation in social image based platforms. By considering the uniqueness of these platforms, we identify three so-
     cial contextual aspects that affect users' preferences from heterogeneous data sources.
2)  We design a hierarchical attention network to model the hierarchical structure of social contextual recommendation. In the attention networks, we feed embeddings from state-of-the-art deep learning models that are tailored for each kind of data into the attention networks. Thus, the attention networks could learn to attend differently based on the rich contextual information for user interest modeling.
3)  We conduct extensive experiments on real-world datasets. The experimental results clearly show the effectiveness of our proposed model.

## 2   RELATED WORK

We summarize the related work in the following four categories.

**General Recommendation.** Recommender systems could be classified into three categories: content based methods, *C*ollaborative*F*iltering (CF) and the hybrid models [2]. Among all models for building recommender systems, latent factor based models from the CF category are among the most popular techniques due to their relatively high performance in practice [40], [34], [39]. These latent factor based models decomposed both users and items in a low latent space, and the preference of a user to an item could be approximated as the inner product between the corresponding user and item latent vectors. In the real-world applications, instead of the explicit ratings, users usually implicitly express their opinions through action or inaction. Bayesian Personalized Ranking (BPR) is such a popular latent factor based model that deals with the implicit feedback [40]. Specifically,

BPR optimized a pairwise based ranking loss, such that the observed implicit feedbacks are preferred to rank higher than that of the unobserved ones. As users may simultaneously express their opinions with several kinds of feedbacks (e.g., click behavior, consumption behavior). SVD++ is proposed to incorporate users' different feedbacks by extending the classical latent factor based models, assuming each user's latent factor is composed of a base latent factor, and an auxiliary latent factor that can be derived from other kinds of feedbacks [26]. Due to the performance improvement and extensibility of SVD++, it is widely studied to incorporate different kinds of information, e.g., item text [58], multi-class preference of users [36].

**Image Recommendation.** In many image based social networks, images are associated with rich context information, e.g., the text in the image, the hashtags. Researchers proposed to apply factorization machines for image recommendation by considering the rich context information [6]. Recently, deep Convolutional Neural Networks(CNNs) have been successfully applied to analyzing visual imagery by automatic image representation in the modeling process [27]. Thus, it is a natural idea to leverage visual features of CNNs to enhance image recommendation performance [18], [28], [17], [5]. E.g., VBPR is an extension of BPR for image recommendation, on top of which it learned an additional visual dimension from CNN that modeled users' visual preferences [18]. There are some other image recommendation models that tackled the temporal dynamics of users' preferences to images over time [17], or users' location preferences for image recommendation [35], [49], [35]. As well studied in the computer vision community, in parallel to the visual content information from deep CNNs, images convey rich style information. Researchers showed that many brands post images that show the philosophy and lifestyle of a brand [14], images posted by users also reflect users' personality [13]. Recently, Gatys et al. proposed a new model of extracting image styles based on the feature maps of convolutional neural networks [10]. The proposed model showed high perceptual quality for extracting image style, and has been successfully applied to related tasks, such as image style transfer [11], and high-resolution image stylisation [12]. We argue that the visual image style also plays a vital role for evaluating users' visual experience in recommender systems. Thus, we leverage both the image content and the image style for recommendation.

**Social Contextual Recommendation.**

Social scientists have long converged that a user's preference is similar to or influenced by her social connections, with the social theories of homophily and social influence [3]. With the prevalence of social networks, a popular research direction is to leverage the social data to improve recommendation performance [33], [23], [24], [51]. E.g., Ma et al. proposed a latent factor based model with social regularization terms for recommendation [33]. Since most of these social recommendation tasks are formulated as non-convex optimizing problems, researchers have designed an unsupervised deep learning model to initialize model parameters for better performance [9]. Besides, ContextMF is proposed to fuse the individual preference and interpersonal influence with auxiliary text content information from social networks [24]. As the implicit influence of trusts and ratings are valuable for recommendation, TrustSVD is proposed to incorporate the influence of trusted users on the prediction of items for an active user [16]. The proposed technique extended the SVD++ with social trust information. Social recommendation has also been considered with social circle [38], online social recommendation [59], social network evolution [50], and so on.

Besides, as the social network could be seen as a graph, the recent surge of network embedding is also closely related to our work [8]. Network embedding models encode the graph structural information into a low latent space, such that each node is represented as an embedding in this latent space. Many network embedding models have been proposed [37], [44], [48], [47]. The network embedding could be used for the attention networks. We distinguish from these works as the focus of this paper is not to advance the sophisticated network embedding models. We put emphasis on how to enhance recommendation performance by leveraging various data embeddings.

**Attention Mechanism.** Neural science studies have shown that people focus on specific parts of the input rather than using all available information [22]. Attention mechanism is such an intuitive idea that automatically models and selects the most pertinent piece of information, which learns to assign attentive weights for a set of inputs, with higher (lower) weights indicate that the corresponding inputs are more informative to generate the output. Attention mechanism is widely used in many neural network based tasks, such as machine translation [4] and image captioning [53]. Recently, the attention mechanism is also widely used for recommender systems [19], [52], [43], [41]. Given the classical collaborative filtering scenario with user-item interaction behavior, NAIS extended the classical item based recommendation models by distinguishing the importance of different historical items in a user profile [19]. With users' temporal behavior, the attention networks were proposed to learn which historical behavior is more important for the user's current temporal decision [31], [32]. A lot of attention based recommendation models have been developed to better exploit the auxiliary information to improve recommendation performance. E.g., ANSR is proposed with a social attention module to learn adaptive social influence strength for social recommendation [43]. Given the review or the text of an item, attention networks were developed to learn informative sentences or words for recommendation [15], [41]. While the above models perform the standard vanilla attention to learn to attend on a specific piece of information, the co-attention mechanism is concerned to learn attention weights from two sequences [21], [56], [46]. E.g., in the hashtag recommendation with both text and image information, the co-attention network is designed to learn which part of the text is distinctive for images, and simultaneously the important visual features for the text [56]. Besides, researchers have made a comprehensive survey the attention based recommendation models [57]. In some realworld applications, there exists hierarchical structure among the data, several pioneering works have been proposed to deal with this kind of relationship [54], [29]. E.g., a hierarchical attention model is

proposed to model the hierarchical relationships of word, sentence and document for document classification [54]. Our work borrows ideas from the attention mechanism, and we extend this idea by designing a hierarchical structure to model the complex social contextual aspects that influence users' preferences. Nevertheless, different from the natural hierarchical structure of words, sentences and documents in natural language processing, the hierarchial structure that influences a user's decision from complex heterogenous data sources is summarized by our proposed model. Specifically, our proposed model has a two-layered hierarchical structure with the bottom layer attention network that summarizes each aspect from the various elements of this aspect. By taking the output of each aspect from the bottom layer, the top-layer attention network learns the importance of the three aspects.

The work that is most similar to ours is the *Attentive Collaborative Filtering* (ACF) for image and video recommendation [5]. By assuming there exists item level and component level implicitness that underlines a user's preference, an attention based recommendation model is proposed with the component level attention and the item level attention. Our work borrows the idea of applying attention mechanism for recommendation, and it differs from ACF and previous works from both the research perspective and the application point. From the technical perspective, we model the complex social contextual aspects of users' interests from heterogeneous data sources in a unified recommendation model. In contrast, ACF only leverages the image (video) content information. From the application view, our proposed model could benefit researchers and engineers in related areas when heterogeneous data are available.

## HETEROGENEOUS    DATA  EMBEDDING   AND PROBLEM DEFINITION

In a social image platform, there are a set of users $U$ ($|U|=M$) and a set of images $V$ ($|V| = N$). Besides *rating images* as standard recommender systems, users also perform two kinds of behaviors: *uploading images* and *building social links*. We represent users' three kinds of behaviors with three matrices: a rating matrix $R \in R^{M \times N}$, an upload matrix $L \in R^{N \times M}$, and a social link matrix $S \in R^{M \times M}$. Each element $r_{ai}$ in the rating matrix $R$ represents the implicit rating preference of user $a$ to image $i$, with $r_{ai} = 1$ denotes user $a$ likes image $i$, otherwise it equals 0. $s_{ba} = 1$ if user $a$ follows (connects to) user $b$, otherwise it equals 0. If the social platform is undirected, $a$ connects to $b$ means $s_{ab} = 1$ and $s_{ba} = 1$. We use $s_a = [s_{1a}, s_{2a}, ..., s_{Ma}]$ to denote the social connections of $a$, i.e., the $a$-th column of $S$. Please note that different from traditional social networking platforms (e.g., the social movie sharing platform), users in these platforms are both image consumers (i.e., reflected in the rating behavior) and image creators (reflected in the upload behavior). Each element $l_{ia}$ in the upload matrix $L$ denotes whether the image $i$ is uploaded (created) by user $a$. In other words, if $a$ is the creator of image $i$, then $l_{ia}=1$, otherwise it equals 0. Since each image can be uploaded by only one user, we have $\sum_{a=1}^{M} l_{ia} = 1$. For ease of explanation, we use $C_i$ to denote the creator of image $i$. And the image upload history of $a$ is denoted as $l_a$, i.e., the $a$-th column of $L$. Without confusion, we use $a,b,c$ to represent users and $i,j,k$ to denote items.

### Heterogeneous Data Embedding

Since there are heterogeneous data sources in this platform, it is natural to adopt the state-of-the-art data embedding techniques to preprocess the social network $S$ and the visual images. The learned embeddings are easier to be exploited by the following proposed model than directly dealing with the heterogeneous data sources. Next, we would first briefly introduce the embedding models for the social network and
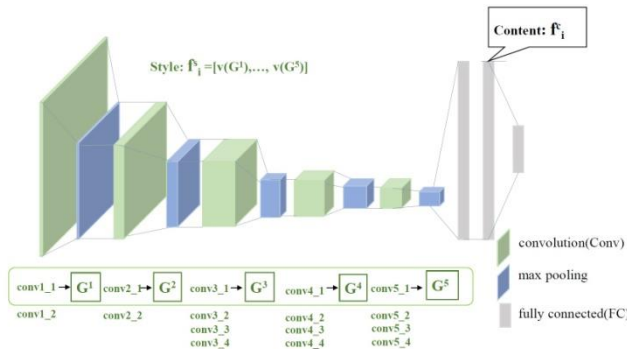


Fig. 2. The image embedding process. For each image $i$, the original image is passed through a VGG19 network. We use the vector of the last connected layer, i.e., $f_i^c$ as its content representation. The Gram matrices $G^l$ on the feature responses of a number of layers are computed. We concatenate the vectorized representations of the typical Gram matrix sequences as the image style representation, i.e., $f_i^s$.

the visual images, and then give the problem definition. Please note that, the problem of how to design sophisticated network embedding techniques, and the visual image features are well researched. Since the focus of this paper is not to advance these topics, we adopt state-of-the-art models and put emphasis on enhancing the recommendation performance with the rich social contextual information.

For the social network **S**, the social embedding part tries to learn the distributed representation of each user in the social network **S**, which encodes social relations in a continuous vector space. Since the focus of this paper is not to design more sophisticated models for network embedding, we exploit Deepwalk [37] for social embedding as it is time-efficient and shows high performance in many network based applications. Deepwalk takes **S** as input and outputs the social latent representation **E** $\in R^{d \times M}$, with the a-th column $e_a$ denotes the latent representation of user $a$.

For each image, it provides rich information including its content as well as its style. Traditionally, convolution neural networks have enjoyed great success for learning useful image visual content features in recent years [27], [42]. We choose *VGG19* for visual content feature extraction as it is a state-of-the-art convolutional neural network architecture that shows powerful capability to capture the image semantics [42]. As commonly adopted by many works, we use the 4096 dimensional representation in the last connected layer in VGG19 as the visual content representation, i.e., each image $i$'s visual feature $\mathbf{f}_i^c$ has 4096 dimensions [18], [55].

Besides image content representation, the image style also plays a vital role for users' visual experience. When users browse images in social platforms, their preferences are not only decided by "what is the content of the image?", but also "does the style of the image meets my preference?". To this end, for each image $i$, besides its content representation $\mathbf{f}_i^c$, we propose to borrow state-of-the-art image style representation models to capture its style representation $\mathbf{f}_i^s$. We choose a popular image style representation method proposed by Gatys et al. [10]. This method has shown high perceptual quality and is widely used in many image-style based tasks [11], [12]. This style describing model is based on the powerful feature spaces learned by convolutional neural networks, with the assumption that the styles are agnostic to the spatial information in the image hidden representations. With the trained VGG19 architecture, suppose a layer $l$ has $N_l$ distinct filter feature maps, each of which is vectorized into a size of $M_l$. Let $\mathbf{B^l} \in R^{Nl \times Ml}$ denotes the filter at layer $l$, with $b_{jk}^l$ is the activation of the j-th filter at position $k$. A summary *Gram* statistic is proposed to discard the spatial information in the feature maps by computing their relations as:

$$gijl = \textstyle\sum_{k} blikbljk$$

where $\mathbf{G^l} \in R^{Nl \times Nl}$ is the Gram matrix, with $g_{ij}^l$ denotes the correlation between feature map $i$ and $j$ in layer $l$. Naturally, the set of Gram matrices $\mathbf{G^1}, \mathbf{G^2}, ..., \mathbf{G^L}$ from different layers of VGG19 provides descriptions of the image style. In practice, researchers found that the style representations on layers 'con1 1', 'conv2 1', 'con3 1', 'con4 1' and 'con5 1' can well represent the textures of an image [11], [12]. As the sizes of these Gram matrices are very large, we downsample each Gram matrix into a fixed size of $32 \times 32$, and then concatenate the vector representation of the downsampled Gram matrices of the five layers. Since there are 5 Gram matrices and each each vectorized Gram matrix has 1024 dimensions, the style representation $\mathbf{f}_i^s$ of image $i$ has 5120 dimensions.

### 2.1 Problem Definition

Given the social matrix **S** and upload matrix **L**, we identify three key social contextual aspects, i.e., social influence, upload history, and the creator admiration that may influence users' preferences. Specifically, the *social influence* aspect from each user $a$'s social network structure $s_a$ is well recognized as an important factor in the recommendation process [33], [24]. The social influence states that, each active user is influenced by her social connections, leading to the similar preferences between social connections [3]. Besides, for each user-item pair $(a,i)$, we could get an upload history list $l_a$ of user $a$, and the creator $C_i$ of image $i$ from the upload matrix **L**. Based on this observation, we design the two contextual aspects in users' preference decision process: an *upload history* aspect that explains the consistency between her upload history $l_a$ and her preference for images, and the *creator admiration* aspect that shows the admiration from the creator $C_i$. These three contextual aspects characterise each user's implicit feedback to images from various contextual situations from the heterogeneous social image data. Now, we define the social contextual image recommendation problem as:

*Definition 1.* [**PROBLEM DEFINITION**]

Given the user rating matrix **R**, the upload matrix **L**, andthesocial network **S** in a social image platform, with the social embedding $\mathbf{e}_a$ of each user $a$, and the content representation $\mathbf{f}_i^c$ and style representation $\mathbf{f}_i^s$ of each image $i$, the social contextual recommendation task aims at: predicting each user $a$'s unknown preference for image $i$ with the three social contextual aspects $(\mathbf{s}_a, \mathbf{l}_a, C_i)$ and the heterogeneous data embeddings $(\mathbf{e}_a, \mathbf{f}_i^c \text{ and } \mathbf{f}_i^s)$ as $g(a,i,\mathbf{s}a,\mathbf{l}a,Ci,\mathbf{e}a,\mathbf{f}ac,\mathbf{f}as)$;

Specifically, in the above definition, $\mathbf{s}_a$, $\mathbf{l}_a$, and $C_i$ denotes the inputs of the three social contextual aspects, i.e., upload history aspect, social influence aspect and the creator admiration aspect.

In the following of this paper, we use bold capital letters to denote matrices, and small bold letters to denote vectors. For any matrix (e.g., social graph **S**), its $i$-th column vector is denoted as the corresponding small letter with a subscript index $i$(e.g., the $i$-th column of **S** is denoted as $\mathbf{s}_a$). We list some mathematical notations in Table 1.

TABLE 1
Mathematical Notations

TABLE 1
Mathematical Notations

| Notations | Description |
|---|---|
| U | userset, $|U| = M$ |
| V | imageset, $|V| = N$ |
| a,b,c,u | user |
| i,j,k,v | image |
| $\mathbf{R} \in \mathrm{R}M{\times}N$ | rating matrix, with $r_{ai}$ denotes whether $a$ likes image $i$ |
| $\mathbf{S} \in \mathrm{R}M{\times}M$ | social network matrix, with $s_{ba}$ denotes whether $a$ follows $b$ |
| $\mathbf{L} \in \mathrm{R}N{\times}M$ | upload matrix, with $l_{ia}$ denotes whether $a$ uploads image $i$ |
| $\mathbf{s}a \in \mathrm{R}M$ | the $a$-th column of **S**, which denotes the social connections of $a$ |
| $\mathbf{l}a \in \mathrm{R}N$ | the $a$-th column of **L**, which denotes the uploaded history of $a$ |
| $C_i \in U$ | the creator (owner) of image $i$, $C_i = [a : L_{ai} = 1]$ |
| $\mathbf{e}a$ | the social embedding of user $a$ from social embedding matrix $\mathbf{E} \in \mathrm{R}^{d{\times}M}$ |
| $\mathbf{f}c$ | the visual content representation of image $i$ |
| $\mathbf{f}_i^s$ | the visual style representation of image $i$ |

## THE PROPOSED MODEL

In this section, we present our proposed *Hierarchical Attentive Social Contextual* recommendation (HASC) model for image recommendation.

As shown in Fig. 3, HASC is a hierarchical neural network that models users' preferences for to unknown images from two attention levels with social contextual modeling. The top layered attention network depicts the importance of the three contextual aspects (i.e., upload history, social influence and creator admiration) for users' decision, which is derived from the bottom layered attention networks that aggregate the complex elements within each aspect. Given a user $a$ and an image $i$ with three identified social contextual aspects, we use $\gamma_{al}(l = 1,2,3)$ to denote $a$'s attentive degree for aspect $l$ on the top layer (denoted as the aspect importance attention with orange part in the figure). A large attentive degree denotes the current user cares more about this aspect in image recommendation process. Besides, as there are various elements within the upload history context $\mathbf{l}_a$ and social influence context $\mathbf{s}_a$. We use $\alpha_{aj}$ to denote $a$'s preference degree for image $j$ in the upload history context $\mathbf{l}_a$ ($l_{ja} = 1$), with a larger value of $\alpha_{aj}$ indicates that $a$'s current interest is more coherent with uploaded image $j$ by user $a$. Similarly, we use $\beta_{ab}$ to denote the influence strength of the $b$ to $a$ in social neighbor context $\mathbf{s}_a$($s_{ba} = 1$), with a larger value of $\beta_{ab}$ indicates that $a$ is more likely to be influenced by $b$. Please note that, for each user $a$ and image $i$, different from the upload history aspect and the social influence aspect, the creator admiration aspect is composed of one element $C_i$ (the creator). Thus, this aspect does not
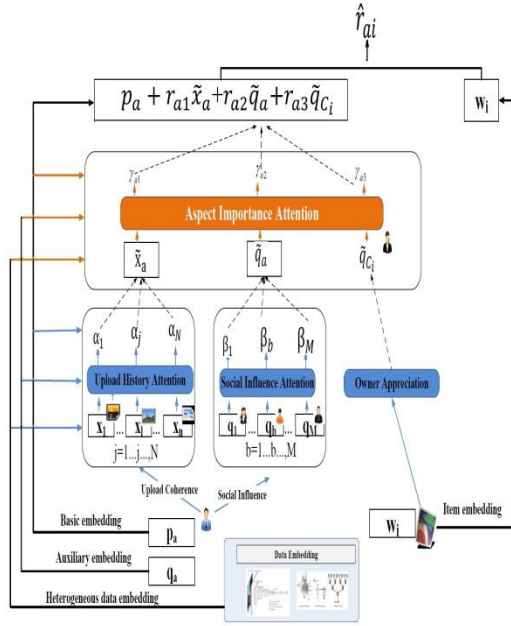
Fig. 3. The overall architecture of the proposed HASC model.

have any sub layers and it is directly sent to the top layer. We use three attention sub-networks to learn these attentive scores in a unified model.

**Objective Prediction Function.** In addition to parameterize each user $a$ with a base embedding $\mathbf{p}_a$ and each item $i$ with a base embedding $\mathbf{w}_i$ as many latent factor based models [40], [26], we also take the inputs of the three social contextual aspects: $\mathbf{s}_a$, $\mathbf{l}_a$, and $C_i$. To model the complex contextual aspects, we extend the classical latent factor models and assume each user and each item has two embeddings. Specifically, each user $a$ is associated with a base embedding $\mathbf{p}_a$ from the base embedding matrix $\mathbf{P}$ to denote her base latent interest in the standard latent factor based models, and an auxiliary embedding vector $\mathbf{q}_a$ from the auxiliary embedding matrix $\mathbf{Q}$. This auxiliary user embedding vector characterizes each user's preference from the social contextual aspects that could not be detected by standard user-image rating behavior. Similarly, each image $i$ is also associated with two embeddings: a base embedding $\mathbf{w}_i$ from the item base embedding matrix $\mathbf{W}$ to denote the basic image latent vector, and an auxiliary vector $\mathbf{x}_i$ from the item auxiliary embedding matrix $\mathbf{X}$ to characterize each image from the social contextual inputs. Thus, by combining the attention mechanism with the embeddings, we model each user $a$'s predicted preference to image $i$ as a hierarchical attention:

$$\hat{r}_{ai} = \mathbf{w}_i^T(\mathbf{p}_a + \gamma_{a1}\mathbf{x}_{ea} + \gamma_{a2}\mathbf{q}_{ea} + \gamma_{a3}\mathbf{q}C_i)$$

$$\text{where } \mathbf{x}_{ea} = X \sum_{j=1}^{N} l_{ja} \alpha_{aj} \mathbf{x}_j, \quad \mathbf{q}_{ea} = X \sum_{b=1}^{M} s_{ba} \beta_{ab} \mathbf{q}_b.$$

In the above prediction function, the representations of three contextual aspects are seamlessly incorporated in a holistic way. Specifically, the first line of Eq.(2) is a top layer attention network that aggregates the three contextual aspects for user embedding. The detailed attention subnetworks of the upload history attention and the social influence attention are listed in the second row. In fact, the attentive weights ($\gamma_{al}, \alpha_{aj}$, and $\beta_{ab}$) rely on our carefully designed attention networks that take various information as input. We leave the details of how to model these three attention networks in the following subsections. Next, we show the soundness of the objective predicted function.

**Relations to Other Models.** By rewriting the predicted preference score in Eq (2), we have:

Basic Latent Factor Model     Item Neighborhood Model

$$
\hat{r}_{ai} = \underbrace{\mathbf{z}_a^T \mathbf{w}_i}_{} \; \mathbf{p} + \underbrace{\gamma_{a1} \sum_{j=1}^{N} \alpha_{aj} l_{ja} \mathbf{x}_j^T \mathbf{w}_i}_{\{z\}} \\
+ \; \underbrace{\gamma_{a2} \sum_{b=1}^{M} \beta_{ab} \mathbf{q}_b^T \mathbf{w}_i}_{\text{Social Neighborhood Model}} \; + \; \underbrace{\gamma_{a3} \mathbf{q}_{C_i}^T \mathbf{w}_i}_{\text{Owner Admiration Bias}} \,,
$$

(3)

where the first part is a basic latent factor model, and the following three parts are extracted from the three contextual aspects. In the last three terms, $\mathbf{x}_j^T\mathbf{w}_i$ can be seen as the similarity function between image $i$ and the user's uploaded image $j$ in the neighborhood-based collaborative filtering from the upload history aspect [26]. $\mathbf{q}_b^T\mathbf{w}_i$ represents the social neighbor's preference to image $i$ with the social influence aspect. As each image is uploaded by a creator, the last term models the creator admiration aspect. This is quite natural in the real-world, as we always like to follow some specific creators' updates.

Please note that, if we replace all the attention scores with equal weights (i.e., $\alpha_{ai} = \frac{1}{\sum_{j=1}^{N} l_{ja}}, \beta_{ab} = \frac{1}{\sum_{b=1}^{M} s_{ba}}$, and $\gamma_{al} = \frac{1}{3}$), our model turns to an enhanced SVD++ model with rich social contextual information modeling [26], [58]. However, this fixed weight assignment treats each user, each aspect, and the elements in each aspect equally. This simply configuration neglects that each user has different considerations for these three contextual aspects. By using hierarchical attention networks, we could learn each user's attentive weights from their historical behaviors.

### 2.2   Hierarchical Attention Network Modeling

In this subsection, we would follow the bottom-up step to model the hierarchical attention networks in detail. Specifically, we would first introduce the two bottom layered attention networks: the upload history attention network and the social influence attention network, followed by the top layered aspect importance attention network that is based on the bottom layered attention networks.

**Upload History Attention.** The goal of the upload history attention is to select the images from each user $a$'s upload history that are representative to $a$'s preferences, and then aggregate this upload history contextual information to characterize each user. Given each image $j$ that is uploaded by $a$, we model the upload history attentive score $\alpha_{aj}$ as a three-layered attention neural network:

$$
\alpha_{aj} = \mathbf{w}^1 \times \sigma(\mathbf{W}^1[\mathbf{p}_a, \mathbf{q}_a, \mathbf{x}_j, \mathbf{w}_j, \mathbf{e}_a, \mathbf{W}^c\mathbf{f}_j^c, \mathbf{W}^s\mathbf{f}_j^s, \mathbf{W}^c\mathbf{f}_a^c, \mathbf{W}^s\mathbf{f}_a^s])
$$

(4)

where $\Theta_u = [\mathbf{W}^c, \mathbf{W}^s, \mathbf{W}^1, \mathbf{w}^1]$ is the parameter set in this three layered attention network, and $\sigma(x)$ is a non-linear activation function. Specifically, as the dimensions of visual content embeddings (i.e., $\mathbf{f}_j^c$ and $\mathbf{f}_a^c$) and the visual style embeddings (i.e., $\mathbf{f}_j^s$ and $\mathbf{f}_a^s$) are much higher than the dimensions of other kinds of embeddings, $\mathbf{W}^c \in \mathbb{R}^{D\times4096}$ and $\mathbf{W}^s \in \mathbb{R}^{D\times5120}$ are the parameters of the bottom layer that performs dimension reduction of the visual content and style representations. $\mathbf{W}^1 \in \mathbb{R}^{(8D+d)\times d1}$ denotes the matrix parameter of the second layer in the attention network, as all data embedding vectors has $D$ dimensions except the social embedding $\mathbf{e}_a$ has $d1$ dimensions. And $\mathbf{w}^1 \in \mathbb{R}^{d1}$ is the vector parameter of the third layer in the attention network. In this attention modeling process, we take three different kinds of embeddings as input:

- **Latent Embedding**: the latent embedding includes
  $[\mathbf{p}_a, \mathbf{q}_a, \mathbf{x}_j, \mathbf{w}_j]$, where $\mathbf{p}_a$ and $\mathbf{q}_a$ are the basic and auxiliary embeddings of user $a$, and $\mathbf{x}_j$ and $\mathbf{w}_j$ are the basic and auxiliary embeddings of item $j$.
- **Social Embedding**: the social embedding part contains the learned social embedding $\mathbf{e}_a$ of each user, which models the global and local structure of each user in the social network $\mathbf{S}$.
- **Visual Embedding**: the visual embedding part includes the visual representations of user $a$ and item $j$. Specifically, each image is characterized by content representation $\mathbf{f}_j^c$ and style representation $\mathbf{f}_j^s$. Besides, as users show their preferences for images from their historical implicit feedbacks, each user $a$'s visual content representation and style representation can

summarized as: $\mathbf{f}^c = \frac{\sum_i^N r_{ai}\mathbf{f}_i}{\sum_{i=1} r_{ai}}$ $\mathbf{f}_a^s = \frac{=1}{N}$ $c$,

also be $\frac{\sum_{i=1}^N r_{ai}\mathbf{f}_i^s}{\sum_{i=1}^N r_{ai}}$. By feeding all the sophisticated designed embeddings from heterogeneous data sources as the input, the upload history attention network learns to focus on the specific information. Please note that, we omit the bias terms in the attention network without confusion. In the following of this paper, for ease of explanation, we also omit the dimension reduction for the visual embeddings (i.e., $\mathbf{W}^c$ and $\mathbf{W}^s$) whenever they are appeared for the attention modeling. Then, the final attentive upload history score $\alpha_{aj}$ is obtained by normalizing the above attention scores as:

$$\alpha_{aj} = \frac{exp(\alpha_{aj})}{\sum_{k=1}^N exp(l_{ka}\alpha_{ak})}. \qquad (5)$$

After we obtain the attentive upload history score $\alpha_{aj}$, the upload history context of user $a$, denoted as $\widetilde{x}_a$, is calculated as a weighted combination of the learned attentive upload history scores:

$$x_{e}a = \sum_{j=1}^N lj a\alpha a j \mathbf{x}j. \qquad (6)$$

**Social Influence Attention.** The social influence attention module tries to select the influential social neighbors from each user $a$'s social connections, and then summarizes these social neighbors' influences into a social contextual vector. If user $a$ follows $b$, we use $\beta_{ab}$ to denote the social influence strength of $b$ to $a$. Then, the social attentive score $\beta_{ab}$ could be calculated as:

$$\beta_{ab} = \mathbf{w}^2\sigma(\mathbf{W}^2[\mathbf{p}_a, \mathbf{p}_b, \mathbf{q}_a, \mathbf{q}_b, \mathbf{e}_a, \mathbf{e}_b, \mathbf{f}_a^c, \mathbf{f}_a^s]), \qquad (7)$$

where $\Theta_s = [\mathbf{W}^2, \mathbf{w}^2]$ are the parameters in the social influence attention network. This social influence attention part also contains three kinds of data embeddings: the user interest embeddings of $\mathbf{p}_a, \mathbf{p}_b, \mathbf{q}_a, \mathbf{q}_b$, the social embeddings of $\mathbf{e}_a$ and $\mathbf{e}_b$, and the visual embeddings of user $a$ with content representation $\mathbf{f}_a^c$ and style representation $\mathbf{f}_a^s$.

Then, the final attentive social influence score $\beta_{ab}$ is obtained by normalizing the above attention scores as:

$$\beta_{ab} = \frac{exp(\beta_{ab})}{\sum_{c=1}^M exp(s_{ca}\beta_{ac})}. \qquad (8)$$

After we obtain the attentive social influence score $\beta_{ab}$, the social context of user $a$, denoted as $^q_e a$, is calculated as the a weighted combination as:

$$q_{e}a = \sum_{b=1}^M sb a\beta ab \mathbf{q}b. \qquad (9)$$

Since each image is uploaded by one creator, for each image $i$, the corresponding uploader is represented as $C_i$. Correspondingly, the owner appreciation context could be simply represented as the the auxiliary embedding $q_{C_i}$ from the user auxiliary embedding matrix $\mathbf{Q}$.

**Aspect Importance Attention Network.** The aspect importance attention network takes the contextual representation of each aspect from the bottom layered attention networks as input, and models the importance of each aspect in the user's decision process. Specifically, for each pair of user $a$ and image $i$, we have two contextual representations from the bottom layer of HASC as: upload history contextual representation $^x_e a$, the social influence contextual representation $^q_e a$, and the owner appreciation contextual representation $\mathbf{q}_{C_i}$. Then, the aspect importance score $\gamma_{al}$(l=1, 2, 3) is modeled with an aspect importance attention network as:

$$\gamma_{al} = \mathbf{w}^3\sigma(\mathbf{W}^3\mathbf{a}_l), \qquad (10)$$

where $\Theta_a = [\mathbf{W}^3, \mathbf{w}^3]$ is the parameter set of this attention network, and $\mathbf{a}_l (l = 1,2,3)$ denotes the input of the top layered attention network, which is the output of the bottom layered attention networks, i.e., $\mathbf{a}_1 = \mathbf{x}_f a$ is the upload history contextual representation, $\mathbf{a}_2 = \mathbf{q}_f a$ is the social influence contextual representation, and $\mathbf{a}_3 = \mathbf{q}_a$ denotes the representation of current active user $a$.

Then, the final aspect importance score $\gamma_{al}$ is obtained by normalizing the above attention scores as:

$$\gamma_{al} = \frac{exp(\gamma_{al})}{\sum_{k=1}^{3} exp(\gamma_{ak})}. \tag{11}$$

For each user $a$, the learned aspect importance scores are tailored to each user, which distinguish the importance of the three social contextual aspects in the user's decision process. For all learned aspect importance scores, the larger the value, the more likely the user's decision is influenced by this corresponding social contextual aspect.

### 2.3   Model Learning

As we focus on implicit feedbacks of users, similar as the widely used ranking based loss function in ranking based latent factor models [40], we also design a ranking based loss function as:

$$M \text{ minL} = {}^{X\,X}s(\hat{r}_{ai} - \hat{r}_{aj}) + \lambda \|\Theta_1\|^2 \tag{12}$$
$$\Theta$$
$$a{=}1(i,j){\in}Da$$

where $s(x)$ is a sigmoid function that transforms the input into range $(0,1)$. $\Theta=[\Theta_1,\Theta_2]$, with $\Theta_1=[\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{X}]$ denotes the embedding matrices and $\Theta_2 = [\Theta_u,\Theta_s,\Theta_a]$ denotes the parameters in each attention network. $\lambda$ is a regularization term that regularizes the user and image embeddings. $D_a =$
$\{(i,j)|i{\in}R_a{\wedge}j{\in}V - R_a\}$ is the training data for $a$ with $R_a$ the imageset that $a$ positively shows feedback.

All the parameters in the above loss function are differentiable. In practice, we implement HASC with TensorFlow to train model parameters with mini batch Adam. The detailed training algorithm is shown in Algorithm 1. In practice, we could only observe positive feedbacks of users with huge missing unobserved values, similar as many implicit feedback works, for each positive feedback, we randomly sample 5 missing unobserved feedbacks as pseudo negative feedbacks at each iteration in the training process [50], [49], [5]. As each iteration the pseudo negative samples change, each missing value gives very weak negative signal.

### 3   EXPERIMENTS

In this section, we show the effectiveness of our proposed HASC model. Specifically, we would answer the following questions: Q1: How does our proposed model perform compared to the baselines (Sec. 5.2)? Q2: How does the model perform under different sparsity (Sec. 5.3)? Q3: How does the proposed social contextual aspects and the hierarchical attention perform (Sec. 5.4)?

### 3.1   Experimental Settings

**Dataset.** To the best of our knowledge, there is no public available dataset that contains heterogenous data sources in a social image based network as described in Fig. 1. To show the effectiveness of our proposed model, we crawl a large

**Algorithm 1** The learning algorithm of HASC

**Input:** Rating matrix $\mathbf{R}$, social matrix $\mathbf{S}$, Uploader matrix $\mathbf{L}$; batch size $m$; max epoch $T$;

**Output:** Latent embedding matrix $\Theta_1 = [\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{X}]$ and parameters in the attention networks $\Theta_2$;

1: Initialize $\Theta$ with a Gaussian distribution with a mean of 0 and a standard variation of 0.1;
2: **for** epoch $\leftarrow$ 1 to $T$ **do**
3: Get training data D with randomly selected 5 times negative feedbacks $<a,i,j> (a\in U, i\in R_a, j\in V-R_a)$;
4:     **for** mini epoch $\leftarrow$ 1 to $\frac{|\mathcal{D}|}{m}$ **do**
5:     Get mini batch : randomly select m pairs
6:     **for** Each pair $<a^k, i^k, j^k>$ in the mini batch **do**
7:       Compute predicted rating of positive item $\hat{r}_{ai}$(Eq.(2));
8:       Compute predicted rating of negative item $\hat{r}_{aj}$(Eq.(2));
9:       Compute the loss $L^k$(Eq.(12));
10:     **end for**
$<a^k, i^k, j^k>_{k=1}^{m}$ in the training data;
11:     Update $\Theta$ with loss as $\frac{1}{m}\sum_{k=1}^{m}\mathcal{L}^k$;
12:     **end for**
13: **end for**
14: Return $\Theta_1 = [\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{X}]$ and parameters in the attention $\Theta_2$.

$\Theta_2$.

dataset from one of the largest social image sharing platform *Flickr*, which is extended from the widely used NUS-WIDE dataset [7], [45]. NUS-WIDE contains nearly 270,000 images with 81 human defined categories from Flickr. Based on this initial data, we get the uploader information according to the image IDs provided in NUS-WIDE dataset from the public APIs of Flickr. We treat all the uploaders as the initial userset, and the associated images as the imageset. We then crawl the social network of the userset, and the implicit feedbacks of the userset to the imageset.

After data collection, in data preprocessing process, we filter out users that have less than 2 rating records and 2 social links. We also filter out images that have less than 2 records. We call the filtered dataset as *F L*. As shown in Table 2, this dataset is very sparse with about 0.15% density. Besides, we further filter *F L* dataset to ensure each user and each image have at least 10 rating records. This leads to a smaller but denser dataset as *F S*. Table 2 shows the statistics of the two datasets after pruning. Please note that the number of images is much more than that of the users. This is consistent with the observation that the number of images usually far exceeds that of users in social image platforms [1], as each user could be a creator to upload multiple images. In data splitting process, we follow the leave-one-out procedure in many research works [5], [20]. Specifically, for each user, we select the last rating record as the test data, and the remaining data are used as the training data. To tune model parameters, we randomly select 5% of the training data to constitute the validation dataset

TABLE 2

The statistics of the two datasets.

**Evaluation Metrics** Since we focus on recommending images to users, we use two widely adopted ranking metric for *top-K* recommendation evaluation: the Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [18], [5]. HR measures

| Dataset | Users | Images | Ratings | Social Links | Rating Density |
|---|---|---|---|---|---|
| F S | 4,418 | 31,460 | 761,812 | 184,991 | 0.55% |
| F L | 8,358 | 105,648 | 1,323,963, | 378,713 | 0.15% |



(a) HR@K on F S

(a) HR@K on F S

(a) HR@K on F S

(a) HR@K on F S                                    (b) NDCG@K on F S

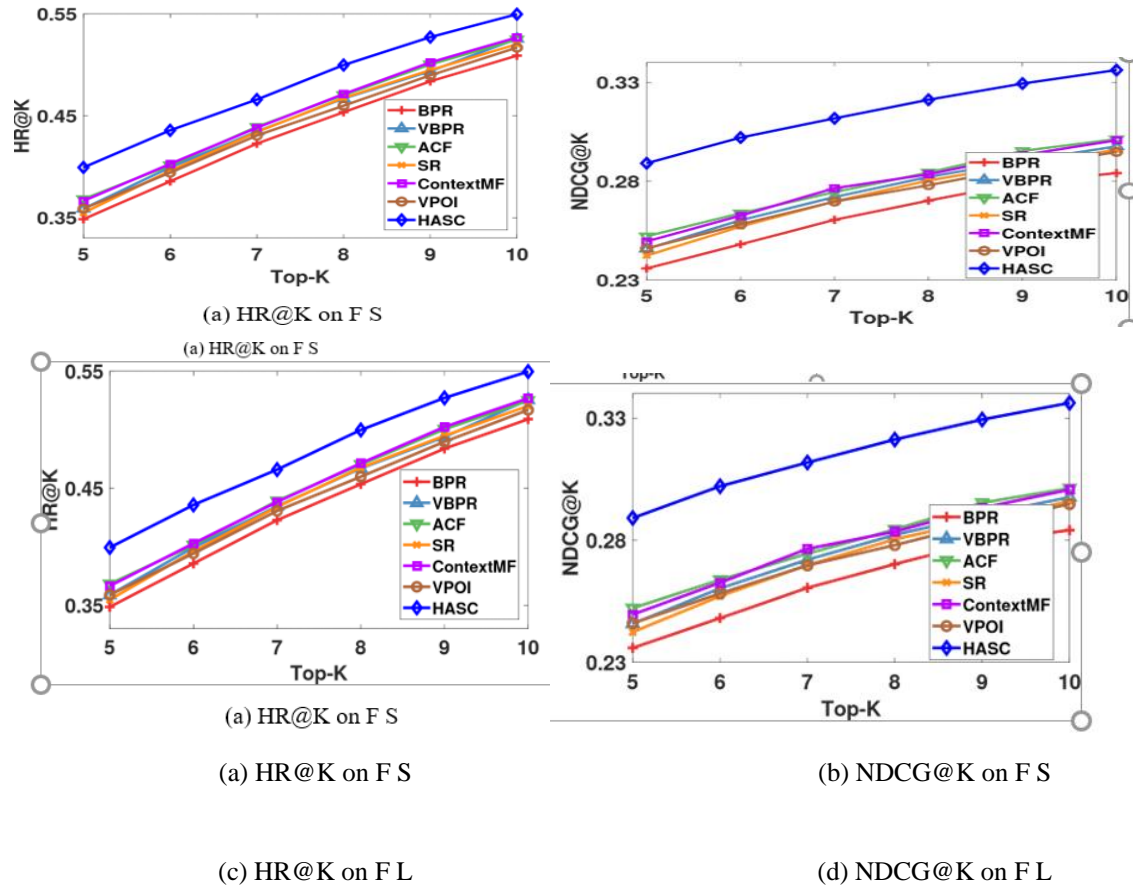(c) HR@K on F L                                    (d) NDCG@K on F L

Fig. 4. Overall performance of different models on the two datasets. (Better viewed in color.)

the percentage of images that are liked by users in the top-K list, and NDCG gives a higher score to the hit images that are ranked higher in the ranking list. As the image size is huge, it is inefficient to take all images as candidates to generate recommendations. For each user, we randomly select 100 unrated images as candidates, and then mix them with the records in the validation and test data to select the top-K results. This evaluation process is repeated for 10 times and we report the average results [18], [5]. For both metrics, the larger the value, the better the ranking performance.

**Baselines.** We compare our proposed HASC model with the following baselines:

- *BPR*: it is a classical ranking based latent factor based model for recommendation with competing performance. This method has been well recognized as a strong baseline for recommendation [40].
- *SR*: it is a social based recommendation model that encodes the social influence among users with social regularization in classical latent factor based models [33].
- *ContextMF*: this method models various social contextual factors, including item content topic, user personal interest, and inter-personal influence in a unified social contextual recommendation framework [24].
- *VBPR*: it extends BPR by modeling both the visual and latent dimensions of users' preferences in a unified framework, where the visual content dimension is derived from a pre-trained VGG network.

- *ACF*: it models the item level and component level attention for image recommendation with two attention networks. For fair comparison, we enrich this baseline by leveraging the upload history as users' auxiliary feedback in this model [5].
- *VPOI*: it is a visual based POI recommendation algorithm. This algorithm relies on the collective matrix factorization to consider the associated images with each POI and the uploaded images of each user. To adapt the POI recommendation to image recommendation, we treat each image as a POI and the uploaded images of each user as the associated images of her. [49].

**Parameter setting.** In the social embedding process with Deepwalk [37], we set the parameters as: the window size $w = 10$ and walks per vertex $\rho = 80$. The social embedding size $d$ is set in the range [32,64,128]. We find when $d = 128$, the social embedding reaches the best performance. Hence, we set $d = 128$ in Deepwalk. There are two important parameters in our proposed model: the dimension $D$ of the user and image embeddings, and the regularization parameter $\lambda$ in the objective function (Eq.(12)). We choose $D$ in [10,15,20,30] and $\lambda$ in [0.001,0.01,01], and perform grid search to find the best parameters. The best setting is $D = 15$ and $\lambda = 0.01$. We find the dimension of the attention networks does not impact the results much. Thus, we empirically set the dimensions of the parameters in the attention networks as 20 (i.e., parameters in $\Theta_2$). The activation function $\sigma(x)$ is set as the LeaklyReLU. To initialize the model, we randomly set the weights in the attention networks with a Gaussian distribution of mean 0 and standard deviation 0.1. Since the objective function of HASC is non-convex, we initialize **P** and **W** from the basic BPR model, and **Q** and **X** with the same Gaussian distribution as the parameters of the attention networks to speed up convergence. We use mini-batch Adam to optimize the model, where the batch size is set as 512 and the initial learning rate is set as 0.0005. There are several parameters in the baselines, for fair comparison, all the parameters in the baselines are also tuned to have the best performance. For all models, we stop model training when both the HR@5 and NDCG@5 on the validation dataset begins to decrease.

### 3.2 Overall Performance

Fig. 4 shows the overall performance of all models on HR@K and NDCG@K on the two datasets with varying sizes of $K$, where the top two subfigures depict the results on F S dataset and the bottom two subfigures depict the results on F L dataset. As shown in this figure, our proposed HASC model always performs the best. With the increase of the top-K list size, the performance of all models increase. The performance trend is consistent over different top-K values and different metrics. We find that considering either the social network or the visual image information could alleviate the data sparsity problem and improve recommendation performance. E.g., VBPR improves over BPR about 3% by incorporating the visual information in the modeling process. ACF further improves VBPR by assigning the attentive weights to different images the user rated and uploaded in the past. SR also has better performance as it leverages the social network information, and ContextMF further improves the performance with content modeling. On average, our proposed model shows about 20% improvement over BPR baseline, and more than 10% improvement over the best baselines on both datasets with regard to the NDCG@5 metric. Last but not the least, by comparing the results of F S and F L, we observe that for each method, the results on F L always outperform F S. We guess a possible reason is that, though F S is denser than F L, the larger F L has nearly two times as many records as F S for training. As the overall trend is similar on the two metrics with different values of $K$, in the following of the subsections, for page limit, we only show the top-5 results.

### 3.3 Performance under Different Data Sparsity

A key characteristic of our proposed model is that it alleviates the data sparsity issue with various social contextual aspects modeling. In this subsection, we investigate the performance of various models under different data sparsity. We mainly focus on the F L dataset as it is more challenging with sparser user rating records compared to the denser F S dataset. Specifically, we bin users into different groups based on the number of the observed feedbacks in the training data, and then show the performance under different groups. Fig. 5 shows the results, where the left part summarizes the user group distribution of the training data and the right part depicts the performance with different data sparsity. As shown in the left part, more than 5% users have less than 4 ratings, and 20% users have less than 16 ratings with more than 100 thousand images on the F L dataset. When the rating scale is very sparse, the BPR baseline can not work well under this situation as it only modeled the sparse user-image implicit feedbacks. Under this situation, the improvement is significant for all models over BPR as these models utilized different auxiliary data for recommendation. E.g., when users have less than 4 ratings, our proposed HASC model improves over BPR by more than 35%. As user rating scale increases, the performance of all models increase quickly with more training rating records, and HASC still consistently outperforms the baselines.

TABLE 3

| Bottom Layer Attention | Top Layer Attention | F S | | F L | |
|---|---|---|---|---|---|
| | | HR | NDCG | HR | NDCG |
| AVG | AVG | 6.44% | 10.28% | 5.54% | 9.02% |
| MAX | MAX | 5.82% | 9.55% | 4.98% | 8.10% |
| AVG | ATT | 7.33% | 11.15% | 5.95% | 9.93% |
| MAX | ATT | 6.84% | 10.96% | 5.72% | 9.55% |
| ATT | AVG | 12.75% | 19.23% | 8.30% | 13.28% |
| ATT | MAX | 12.20% | 18.56% | 8.02% | 12.85% |
| ATT | ATT | **14.57%** | **22.55%** | **10.67%** | **16.70%** |



**The improvement of using different attention mechanism compared to BPR**.

TABLE 4
**The improvement of modeling different contextual aspects with our proposed model compared to BPR.(U:upload history, S:**
**social influence, C: creator admiration)**

| Aspects | F S | | F L | |
|---|---|---|---|---|
| | HR | NDCG | HR | NDCG |
| U | 8.70% | 16.52% | 6.44% | 11.03% |
| S | 9.63% | 16.78% | 5.29% | 9.65% |
| C | 8.57% | 14.53% | 4.37% | 7.93% |
| U+S+C | **14.57%** | **22.55%** | **10.67%** | **16.70%** |

**3.4**                                      **Attention Analysis**

In this part, we conduct experiments to give more detailed analysis of the proposed attention network. We would evaluate the soundness of the designed attention structure and the superiority of combining the various data embeddings for attention modeling.

In the experiments, we use the LeaklyReLU as the activation function $\sigma(x)$ for attention modeling, and then attentively combine the elements of each set with a soft attention. Alternately, instead of attentively combining all the elements, a direct solution is to use the hard attention with MAX operation that selects the element with the largest attentive score at each layer of the hierarchical attention network. E.g., for the upload history aspect, *Max* learns the attentive upload history score in Eq.(6) as: $\mathbf{x}e_a = \mathbf{x}_j$, where $l_{ja}=1 \wedge (\forall l_{ka}= 1, \alpha_{ja} \geq \alpha_{ka})$. Particularly, if we simply set the attentive scores with the average pooling (i.e., $\alpha_{ai} = \frac{1}{|L_a|}, \beta_{ab} = \frac{1}{|S_a|}, \gamma_{al} = \frac{1}{3}$), our model degenerates to an enhanced SVD++ with social contextual modeling but without any attentive modeling. If we do not model any social contextual aspects, our model degenerates to the BPR model [40]. Table 3 shows the results of different attention mechanism. As shown in this table, the best results are achieved by using our proposed attention mechanism, followed by AVG and MAX. We guess a possible reason is that: each user's interests are diversified, and it is challenging to infer each user's interests from the limited training data. If we simply using a hard attention with the maximum value or adopting average aggregation, many valuable contextual information is neglected in this process. Besides, we
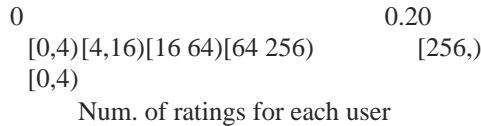
0                                         0.20
[0,4)[4,16)[16 64)[64 256)            [256,)
[0,4)

Num. of ratings for each user

**Fig. 5. Performance under different sparsity.**

observe that ATT that operates at the bottom layer achieves much better performance than its counterparts that operates on the top layer (e.g., the comparison results between the fourth row and the sixth row). Since each aspect at the bottom layer usually contains much more elements than the top layer, attentively summarizing each contextual aspect at the bottom layer would provide valuable information for the top layer. In contrast, if we use AVG or MAX at the bottom layer, the results are not satisfactory when we use "ATT" at the second layer, since the input of the second layer lacks many important information.

After showing the soundness of our proposed attention structure, Table 4 presents the performance of using different contextual aspects with our proposed hierarchical attention. As shown in this table, each aspect improves the performance. By combining all social contextual aspects with hierarchical attention, the model reaches the best performance.

**TABLE 5**
Performance of different kinds of inputs for attention modeling
(Base: base embedding, Aux: auxiliary embedding, Soc: social embedding, Vis C(Vis C): visual content(style) embedding). with "Base" denotes the base embedding, "Aux" denotes the
auxiliary embedding, "Soc" denotes the social embedding, and
"Vis C", "Vis S", "Vis CS" denotes the visual content feature, visual style feature, and both visual features.

| Input Embedding | F S | | F L | |
|---|---|---|---|---|
| | HR | NDCG | HR | NDCG |
| Base | 0.358 | 0.257 | 0.439 | 0.319 |
| Base+Aux | 0.366 | 0.264 | 0.445 | 0.323 |
| Base+Aux+Soc | 0.367 | 0.270 | 0.450 | 0.331 |
| Base+Aux+Vis C | 0.388 | 0.278 | 0.453 | 0.335 |
| Base+Aux+Vis S | 0.383 | 0.275 | 0.451 | 0.332 |
| Base+Aux+Vis CS | 0.393 | 0.282 | 0.464 | 0.342 |
| Base+Aux+Soc+Vis CS | **0.400** | **0.289** | **0.475** | **0.347** |

TABLE 6
Performance of different kinds of social embedding techniques for the attention modeling.

| Input Embedding | F S | | F L | |
|---|---|---|---|---|
| | HR | NDCG | HR | NDCG |
| Base+Aux+DeepWalk | 0.367 | 0.270 | 0.450 | 0.331 |
| Base+Aux+LINE | 0.369 | 0. 273 | 0.452 | 0.334 |
| Base+Aux+GCN | 0.371 | 0.276 | 0.459 | 0.340 |
| Base+Aux+DeepWalk+Vis CS | 0.400 | 0.289 | 0.475 | 0.347 |
| Base+Aux+LINE+Vis CS | 0.400 | 0.289 | 0.474 | 0.345 |
| Base+Aux+GCN+Vis CS | 0.401 | 0.290 | 0.475 | 0.348 |

Besides, in the attention modeling process, we also learn the attentive weights by modeling different kinds of input embeddings from the heterogeneous data sources. For each attention layer, it consists the following kinds of inputs: the latent interest representations of base embeddings (i.e., $\mathbf{p}_a$ and $\mathbf{w}_i$) and auxiliary embeddings (i.e., $\mathbf{q}_a$ and $\mathbf{x}_i$), the social embeddings (i.e., $\mathbf{e}_a$), and the visual embeddings with content representations (i.e., $\mathbf{f}_i^c$ of image $i$ and $\mathbf{f}_a^c$ of user $a$) and style representations (i.e., $\mathbf{f}_i^s$ of image $i$ and $\mathbf{f}_a^s$ of user $a$ ). Table 5 shows the performance of HASC with different kinds of input embeddings. From this table, we have several observations. First, as the auxiliary latent embedding representation could model each user and each item from the rich social contextual information, taking the auxiliary embeddings could improve the performance than solely feeding the base embeddings for attention modeling. Second, the improvement of social embeddings is not very significant. We guess a possible reason is that, the social influence aspect already considers the social neighborhood information for users' interest modeling. As the social embeddings represent the overall social network with both local and global structure, the improvement is limited with the additional global network structure modeling. Third, we observe that the improvement of the visual embeddings is very significant. Both the content and the style information could enhance the recommendation performance. By combining content and style embeddings, the performance further improves. This observation empirically shows the complementary relationship of content and style in visual images. Last but not least, by feeding the three different kinds of data embeddings into the attention network embedding, the proposed HASC could achieve the best performance.

In the previous experiments, we use the DeepWalk as the social network embedding model to obtain the social network embedding vector of each user. Now we would show the effectiveness of adopting different network embedding techniques. We choose two state-of-the-art network embedding models: LINE [44] and GCN [25], and compare the performance. The results are shown in Table 6. As can be seen from this table, when the item visual embeddings are not incorporated, using the advanced graph embedding techniques (e.g., GCN), could partially improve the recommendation performance, as these advanced models could better capture the social network structure. When all the input embeddings are incorporated, these advanced graph embedding models show similar performance compared to the DeepWalk based network embedding model. We guess the reason is that, as stated as Table 5, the improvement of the social embedding is not as significant as the visual based input for attention modeling when all the input embeddings are considered.

**Attention Weights Visualization.** Besides given the overall results of different attention modeling setting, we give a visualization of the learned attention weights of users from the F L dataset. Firstly, for each user, we group her into three categories according the aspect that has the
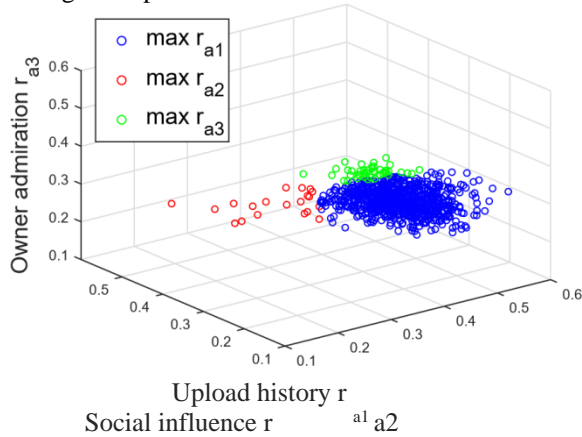


Fig. 6. Visualization of aspect weights of randomly sampled users.

largest attention value. In other words, for each user $a$ in the first group, she has the largest aspect weight for upload history, i.e., $\gamma_{a1} > \gamma_{a2} \wedge \gamma_{a1} > \gamma_{a3}$. Then, for each group, we randomly select 10% of users and visualize them in Fig. 6. As observed in this figure, each randomly sampled user has her own attentive weights for balancing the three contextual aspects. Besides, most of the users belong to the first group that has the largest value of the upload history aspect, which empirically shows that many users show similar preferences between their uploaded images and the liked images. This observation is also consistent with Table 4 that shows leveraging the upload history has the largest performance gain compared to the remaining two aspects on F L dataset.

## 3.5 Case Study

In order to better understand the proposed model, we visualize several typical users and the experimental results of different recommendation models in Fig.7. In this figure, each row represents a user. The first column shows the images liked by the user in the training data, and the second column shows the test image of each user in the test data. Please note that, due to page limit, we only show six typical training images of each user if she has rated more than 6 images in the training data. The third column shows the NDCG@5 results of different models. Specifically, to validate the effectiveness of different aspects in the modeling process, we use $U$, $S$, and $C$ to denote the three simplified versions of our proposed HASC model that only consider the upload history aspect (i.e., $\gamma_{a2} = \gamma_{a3} = 0$), the social influence aspect(i.e., $\gamma_{a1} = \gamma_{a3} = 0$), and the owner admiration aspect(i.e., $\gamma_{a1} = \gamma_{a2} = 0$). We present the learned attention weights of different aspects of our proposed HASC model in the fourth column. The last column gives some intuitive explanations of the experimental results. As shown in this figure, by learning the importance of different aspects with attentive modeling, HASC could better learn each user's preference from various social contextual aspects. Thus, it shows the the best performance for the users in the first three rows. In the fourth row, we present a case that all the models do not perform well expect than the simplified $C$ model from HASC that leverages the single creator admiration aspect into consideration. We carefully analyze this user's records and guess a possible reason is that: the style and the content of the test image has rarely appeared in the user's training data. As this test image differs from the distribution of the training images of this user, most models could not perform well. However, the $C$ model that leverages the owner admiration shows better results than the remaining models, as this user has liked several images uploaded by the owner. This example gives us an intuitive explanation that shows when our proposed model may not perform very well. Nevertheless, we must notice that this case is caused by the situation that the test

pattern is not consistent with the patterns in the training data, which is uncommon. Therefore, we could empirically conclude that our proposed model shows the best results for most cases.

## 4    CONCLUSIONS

In this paper, we have proposed a hierarchical attentive social contextual model of HASC for social contextual image recommendation. Specifically, in addition to user interest modeling, we have identified three social contextual aspects that influence a user's preference to an image from heterogeneous data: the upload history aspect, the social influence aspect, and the owner admiration aspect. We designed a hierarchical attention network that naturally mirrored the hierarchical relationship of users' interest given the three identified aspects. In the meantime, by feeding the data embedding from rich heterogeneous data sources, the hierarchical attention networks could learn to attend differently to more or less important content. Extensive experiments on real-world datasets clearly demonstrated that our proposed HASC model consistently outperforms various state-of-theart baselines for image recommendation.

## REFERENCES

[1]  Flickr Statistics. https://expandedramblings.com/index.php/ flickr-stats/, 2017. [Online; accessed 20-Jan-2018].

[2]  G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *TKDE*, 17(6):734–749, 2005.

[3]  A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, pages 7–15. ACM, 2008.

[4]  D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[5]  J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua. Attentive collaborative filtering: Multimedia recommendation with itemand component-level attention. In *SIGIR*, pages 335–344. ACM, 2017.

[6]  T. Chen, X. He, and M.-Y. Kan. Context-aware image tweet modelling and recommendation. In *MM*, pages 1018–1027. ACM, 2016.

[7]  T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nuswide: a real-world web image database from national university of singapore. In *MM*, page 48. ACM, 2009.

[8]  P. Cui, X. Wang, J. Pei, and W. Zhu. A survey on network embedding. *TKDE*, 2018.

[9]  S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu. On deep learning for trust-aware recommendations in social networks. *TNNLSS*, 28(5):1164–1177, 2017.

[10]  L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, pages 262–270, 2015.

[11]  L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[12]  L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, pages 3985–3993, 2017.

Fig. 7. The case study of several typical users. In this figure, each row represents a user.
The first and the second column are the training and test images of the user.
The Top-5 recommendation results of NDCG@5 are shown in the third column. In the third column, the left three models are simplified versions of our proposed HASC model that only leverage one aspect, and the model
with best performance is shown with bold italic letters.

[13] F. Gelli, X. He, T. Chen, and T.-S. Chua. How personality affects our likes: Towards a better understanding of actionable images. In *MM*, pages 1828–1837. ACM, 2017.

[14] F. Gelli, T. Uricchio, X. He, A. Del Bimbo, and T.-S. Chua. Beyond the product: Discovering image posts for brands in social media. In *MM*. ACM, 2018.

[15] Y. Gong and Q. Zhang. Hashtag recommendation using attentionbased convolutional neural network. In *IJCAI*, pages 2782–2788, 2016.

[16] G. Guo, J. Zhang, and N. Yorke-Smith. A novel recommendation model regularized with user trust and item ratings. *TKDE*, 28(7):1607–1620, 2016.

[17] R. He, C. Fang, Z. Wang, and J. McAuley. Vista: a visually, socially, and temporally-aware model for artistic recommendation. In *Recsys*, pages 309–316. ACM, 2016.

[18] R. He and J. McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[19] X. He, Z. He, J. Song, Z. Liu, Y.-G. Jiang, and T.-S. Chua. Nais: Neural attentive item similarity model for recommendation. *TKDE*, 2018.

[20] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[21] B. Hu, C. Shi, W. X. Zhao, and P. S. Yu. Leveraging metapath based context for top-n recommendation with a neural coattention model. In *SIGKDD*, pages 1531–1540. ACM, 2018.

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[23] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, pages 135–142. ACM, 2010.

[24] M. Jiang, P. Cui, F. Wang, W. Zhu, and S. Yang. Scalable recommendation with social contextual information. *TKDE*, 26(11):2789– 2802, 2014.

[25] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[26] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434. ACM, 2008. [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[28] C. Lei, D. Liu, W. Li, Z.-J. Zha, and H. Li. Comparative deep learning of hybrid representations for image recommendations. In *CVPR*, pages 2545–2553, 2016.

[29] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv:1506.01057*, 2015.

[30] D. C. Liu, S. Rogers, R. Shiau, D. Kislyuk, K. C. Ma, Z. Zhong, J. Liu, and Y. Jing. Related pins at pinterest: The evolution of a real-world recommender system. In *WWW*, pages 583–592, 2017.

[31] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang. Stamp: short-term attention/memory priority model for session-based recommendation. In *SIGKDD*, pages 1831–1839. ACM, 2018.

[32] P. Loyola, C. Liu, and Y. Hirate. Modeling user session and intent with an attention-based encoder-decoder architecture. In *RecSys*, pages 147–151. ACM, 2017.

[33] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM*, pages 287–296. ACM, 2011.

[34] A. Mnih and R. R. Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2008.

[35] W. Niu, J. Caverlee, and H. Lu. Neural personalized ranking for image recommendation. In *WSDM*, pages 423–431. ACM, 2018.

[36] W. Pan and Z. Ming. Collaborative recommendation with multiclass preference context. *IEEE Intelligent Systems*, 32(2):45–51, 2017.

[37] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *KDD*, pages 701–710. ACM, 2014.

[38] X. Qian, H. Feng, G. Zhao, and T. Mei. Personalized recommendation combining user interest and social circle. *TKDE*, 26(7):1763– 1777, 2014.

[39] S. Rendle. Factorization machines with libfm. *TIST*, 3(3):57, 2012.

[40] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461. AUAI Press, 2009.

[41] S. Seo, J. Huang, H. Yang, and Y. Liu. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Recsys*, pages 297–305. ACM, 2017.

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] P. Sun, L. Wu, and M. Wang. Attentive recurrent social recommendation. In *SIGIR*, pages 185–194. ACM, 2018.

[44] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *WWW*, pages 1067–1077, 2015.

[45] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain. Triclustered tensor completion for social-aware image tag refinement. *PAMI*, 39(8):1662–1674, 2017.

[46] Y. Tay, A. T. Luu, and S. C. Hui. Multi-pointer co-attention networks for recommendation. In *SIGKDD*, pages 2309–2318. ACM, 2018.

[47] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. In *ICLR*, 2018.

[48] D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In *KDD*, pages 1225–1234. ACM, 2016.

[49] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *WWW*, pages 391–400, 2017.

[50] L. Wu, Y. Ge, Q. Liu, E. Chen, R. Hong, J. Du, and M. Wang. Modeling the evolution of users' preferences and social links in social networking services. *TKDE*, 29(6):1240–1253, 2017.

[51] L. Wu, P. Sun, R. Hong, Y. Ge, and M. Wang. Collaborative neural social recommendation. *TSMC: Systems*, 2019.

[52] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, and T.-S. Chua. Attentional factorization machines: Learning the weight of feature interactions via attention networks. In *IJCAI*, pages 3119–3125.

[53] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048– 2057, 2015.

[54] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489, 2016.

[55] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma. Collaborative knowledge base embedding for recommender systems. In *KDD*, pages 353–362. ACM, 2016.

[56] Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong. Hashtag recommendation for multimodal microblog using co-attention network. In *IJCAI*, pages 3420–3426, 2017.

[57] S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *CSUR*, 52(1):5, 2019.

[58] S. Zhang, L. Yao, and X. Xu. Autosvd++: An efficient hybrid collaborative filtering model via contractive auto-encoders. In *SIGIR*, pages 957–960. ACM, 2017.

[59] Z. Zhao, H. Lu, D. Cai, X. He, and Y. Zhuang. User preference learning for online social recommendation. *TKDE*, 28(9):2522– 2534, 2016.