IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received July 16, 2020, accepted July 27, 2020, date of publication August 3, 2020, date of current version August 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013639

Aspect-Based Fashion Recommendation With Attention Mechanism

WEIQIAN LI[1,2] AND BUGAO XU[2]
[1]School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China
[2]Department of Merchandising and Digital Retailing, University of North Texas, Denton, TX 76203, USA

Corresponding author: Bugao Xu (bugao.xu@unt.edu)

ABSTRACT With the rapid growth of fashion e-commerce, fashion recommendation has become a main digital marketing tool that is built on customer reviews and ratings. Online review is a powerful source for understanding users' shopping experiences, preferences and feedbacks on product/item performances, and thus is useful for enhancing personalized recommendations for future purchases. However, most extant fashion recommendation methods lack effective frameworks to integrate local and global aspect representations extracted from customers' ratings and reviews. In this paper, we proposed an aspect-based fashion recommendation model with attention mechanism (AFRAM) to predict customer ratings based on online reviews of fashion products. This model can extract latent aspect features about users and items separately through two parallel paths of convolutional neural networks (CNN), long short-term memory networks (LSTM), and attention mechanisms. One path processes user reviews and the other copes with item reviews. On each path, CNN and LSTM are both coupled with an attention mechanism to capture local aspect features and global aspect features respectively, which are combined through a mutual operation module. The mutual operations on both paths can enhance the generalization of the AFRAM model. The extracted features from the two paths are further merged to predict users' ratings. Real-world customer reviews and ratings collected from two renowned business websites were used to train and test AFRAM. The experiment results demonstrate that AFRAM is more effective in customer rating predictions, as compared to several state-of-the-art fashion recommenders.

INDEX TERMS Aspect-based fashion recommendation, attention mechanism, CNN, LSTM.

I. INTRODUCTION

Recommendation is a major digital marketing means used in the e-commerce business. Most recommendation systems are created based on users' reviews and ratings of previously purchased products. Generally, the reviews made by one user for all sought/bought items (e.g., fashion products) are regarded as a user review document, while the reviews associated with one item across all users are referred to as an item review document. A user review document contains personal preferences towards some aspects of an item. For example, if a female user prefers a fit style in her fashion selection, she may frequently use specific phrases to describe fashion products, such as a bit small, skinny body, and loose fit, in her review document. On the other hand, an item review document provides a sum of comments from all users about a specific item, which hopefully can lead to exposing some important aspects useful for understanding the item [1]. A rating score is an indicator of the overall merit of an item. To better understand rating behaviors, it is necessary to find out more definitive information on whether a user likes or dislikes particular aspects of a fashion product. Many studies have focused on improving the interpretability of users' intents by incorporating semantic information of reviews and ratings into a personalized recommendation system [1]–[7]. In these existing solutions, including D-Attn [2], TransNet [3], and ANR [4], topic modeling or nonnegative matrix factorization were adopted to detect users' preferences or products' features in a review, and the interactions of user and item data were analyzed to establish a text feature matrix for recommendation models. Although these models are effective to a certain extent, they do not integrate different polarities of the same word in a sentence (local aspect representation) and in the whole review (global aspect representation). Currently,

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

no recommendation model has been widely accepted for fashion marketing.

In this paper, we propose an aspect-based fashion recommendation model with attention mechanism (AFRAM) based on users' reviews and ratings to target local and global aspect representations. This model is constructed with two independent paths to process user/item reviews simultaneously, and each path has a convolutional neural network (CNN), a long-short time memory network (LSTM) with attention mechanism to capture local aspect features and global aspect features separately. To enhance the generalization of the AFRAM model, the local and global aspect features from both user and item reviews are merged through mutual operations prior to the rating prediction. Two datasets, the Clothing, Shoes & Jewelry dataset from Amazon 5-core and the reviews of another national online retailer, are used to train and test AFRAM. The importance of extracting customers' preferences for different aspects of fashion and the effectiveness of our attention modules in AFRAM are examined as opposed to several state-of-the-art fashion recommenders. The main contributions of this paper may be stated as follows:

1) A new fashion recommendation adopts local and global aspects to learn the relevant latent semantic information, and to model the interactions between user and item reviews to enhance recommendation interpretability.

2) With aspect representation and attention mechanism, users' preferences in different fashion aspects can be grasped to assess the local and global importance of each word in a review.

3) Trained with the real-world datasets from two online retailers, AFRAM can outperform several other fashion recommendation models in predicting user's ratings.

II. RELATED WORK

This work is related to research in four areas: reviews-based recommendation, aspect-based recommendation, attention mechanism, and fashion recommendation. Recent advances in each of these areas are introduced below.

A. REVIEWS-BASED RECOMMENDATION

Many articles have reported methods of using user reviews to improve the accuracy and interpretability of recommendations. To solve data sparsity and cold-start problems in traditional collaborative filtering algorithms, some work used topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [5] or LDA-like algorithms [6]–[8], to extract latent semantic information from reviews and fused it with ratings to make recommendations more accurate and interpretable. Song *et al.* proposed two separate factor learning models by taking advantage of the sentiment-consistency and text-consistency of users and items, and combined the two views to make a rating prediction [9]. However, these studies ignored the word order and sentence characteristics in a review, missing contextual information.

Recently, deep learning techniques have been applied to recommender systems based on review contents with great successes. CNN [10], [11], Recurrent Neural Network (RNN) [12], [13], and Capsule Network [14] are widely used to extract the semantic contextual information [10], [15] by training the networks to learn the deep feature representation of reviews and probabilistic matrix factorization for the rating prediction. In [16] and [17], a word vector model and CNNs were used to learn users' behaviors and item's attributes. In [18], an RNN was combined with a factorization machine via a regularization term to predict the rating of an item by using item's latent factors learned from the RNN. In [15], a sentiment capsule architecture with a novel routing called the bi-agreement mechanism was proposed to identify the informative logic unit and the sentiment-based representations in the user-item level for rating predictions.

As mentioned in the above example, RNN has been widely used for recommendation [19]–[23], sentiment analysis [24], and text classification [25] because they can capture global long-term dependencies and temporal sentence semantics by the directed cycle connections between units. However, it is known that RNN has gradient vanishing or exploding problems [26], [27]. A special RNN, called the Long Short-Term Memory network (LSTM), has been developed to overcome the problems in performing natural language process (NLP) tasks for recent years [20], [23], [28], [29]. LSTM is empowered by the transferability of sequence information so that its hidden layer output at each position can be used as a contextual embedding for a current word [30]. For example, Heinz, et. al. proposed a feedforward network in the "fashion space" to generate article embeddings as an LSTM's input to predict a styled vector for each client based on client's past purchase sequence [28]. However, LSTM was not sufficient for learning long-term dependencies without the context information of future words [25]. This led to a bi-LSTM model [31] which was able to grasp the previous and future aspect context information and to get the expression of a word in the whole sentence. Thus, the bi-LSTM will be adopted to get latent global aspect information in reviews for this study.

B. ASPECT-BASED RECOMMENDATION

Although a review-based recommendation is based on the understanding of why users made their ratings, it lacks the ability to capture not only user preferences and item properties, but also dynamic and fine-grained interactions between users and items. Consequently, aspect-based recommendations became attractive.

An aspect-based recommender system intends to extract aspects from textual reviews, which can be generally divided into two groups. The first group relies on external NLP tools to analyze review contents to learn aspects for sentiment analysis [32]–[34]. For instance, using fine-grained aspect-level sentiment analysis can automatically discover the most valuable aspect to enhance future user experience [35]. The second group, such as LDA [5], [32], AARM [36], FLAME [37], and ASCF [38], builds an internal structure or framework to represent different aspects in a user or item review. Chin *et al.* [4] combined user and item aspect-level

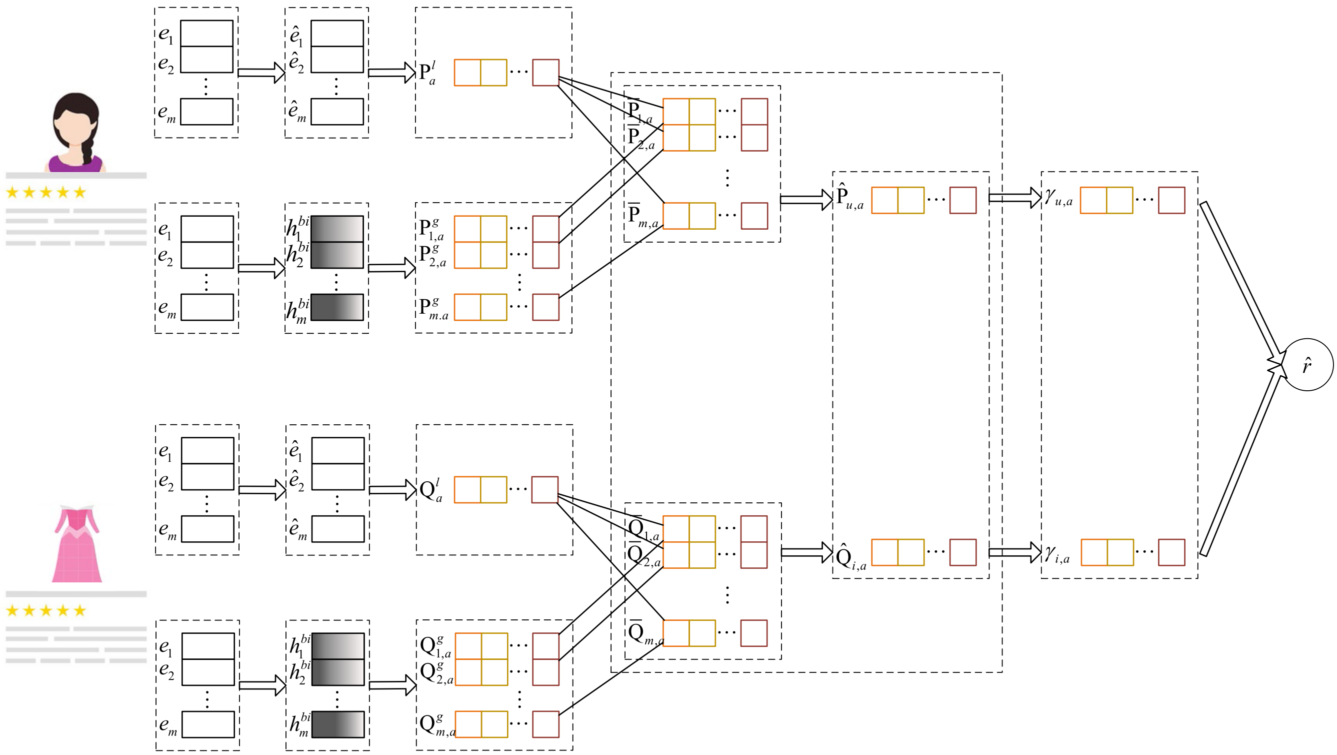


FIGURE 1. The architecture of aspect-based fashion recommendation with attention mechanism using user reviews and item reviews.

representations with aspect importance, and then estimated an overall rating for any user-item pair. Cheng *et al.* designed an aspect-aware topic model via a multinomial distribution over the reviews to learn different user and item aspects in the topic space [39]. Recently, Li *et al.* proposed an aspect-specific gating and projection to extract the viewpoint of relevant information precisely by disambiguating the semantics of each word [15].

C. ATTENTION MECHANISM

Nowadays, attention has been one of the powerful concepts in the deep learning field. As a vector, it is often the outputs of dense layers using the softmax function. Attention mechanism was first introduced to the NLP field in [40] and still maintains its activeness in this field because it can flexibly select a reference for context information to facilitate global learning [41]. The attention mechanism models include three main categories. The first category is a single attention model which is merely determined by textual level interactions among different parts of a content [42]. The second category, such as hierarchical attention [21], [43], [44] and co-attention (local & global [2], hard & soft [41], [45]–[47]), is a hybrid attention model which relies on CNN, RNN and other networks at different levels. In [43] and [44], the word-level and sentence-level hierarchical attention models were used, and the output of the first attention model was the input of the second attention model. A dual attention (D-Attn) model contains local attention (L-Attn) and global attention (G-Attn) models in parallel. L-Attn aims to find meaningful keywords in the sliding window of a user/item review, while G-Attn aims to capture the overall sentiment expression of the user or item review. The third category is a knowledge-based attention model which is similar to the hierarchical attention model except that its input is originated from the information of other field knowledge [48], [49]. Other attention-based work includes translation tasks [50] and text classification [51].

D. FASHION RECOMMENDATION

Users' preferences in fashion can be reflected by diverse fashion aspects in their reviews. Therefore, it is particularly essential to explore the aspects of fashion that users are interested. Most of the current fashion recommendations are based on visual images of clothing, consumers' purchase behaviors, sale data, and other relevant information [52]–[60]. However, few studies on fashion recommendation have paid enough attentions to users' reviews and ratings mainly because they are not easily interpretable. In this paper, we will use reviews and ratings from the Amazon 5-core database (Clothing, Shoes & Jewelry) and from another online retailer to build a model for fashion recommendations.

III. ASPECT-BASED FASHION RECOMMENDATION WITH ATTENTION MECHANISM (AFRAM)

Based on reference [2], a network structure with two parallel paths was proposed for AFRAM whose overall architecture is presented in Figure 1. The two parallel paths of AFRAM extract the aspect features from user reviews and item reviews, respectively, and each path contains side-by-side CNN and bi-LSTM modules combined with attention mechanism to capture local-aspect and global-aspect features simultaneously. These two kinds of aspect features are fused together by a mutual operation to detect the relevant semantic information and to enhance the robustness of the model, and then are merged with the semantic information detected from another path through mutual operation and fully connected layers as the final learned aspect representation for the user/item rating prediction. Since the modeling processes for user and item reviews are identical, we will just focus on the process of user review.

A. CONTEXT EMBEDDING MODULE

Assume that a user review text is $R_U = (r_1, r_2, \cdots, r_m)$, where m refers to the number of words in a review, and r_i denotes the i-th word in the review. We first map each word to its embedding representation $E_U = (e_1, e_2, \cdots, e_m) \in \mathbb{R}^{m \times d}$ via a context embedding layer, in which e_i is the embedding vector for the i-th word, and d is the number of dimensions for the embedding vector of each word. The embedding layer can be simply regarded as a look-up operation in a shared embedding matrix which can be initialized using word vectors that have been pre-trained on large corpora, and the embedding matrix can facilitate a better semantic representation of the user reviews. In this paper, we use the word2vec [61] for the pre-trained word embedding.

B. LOCAL ASPECT FEATURE EXTRACTION MODULE

Given an embedded user/item review representation, the goal of this module is to extract a set of local aspect user/item features. Figure 2 illustrates the process of local aspect feature extraction. Initially, we can encode the embedded context from the massive given review texts and ratings [4], and then use CNN to learn the importance of each word and its aspect representation. Finally, we can obtain the word aspect representation and extract latent local aspect semantic features by attention mechanism.

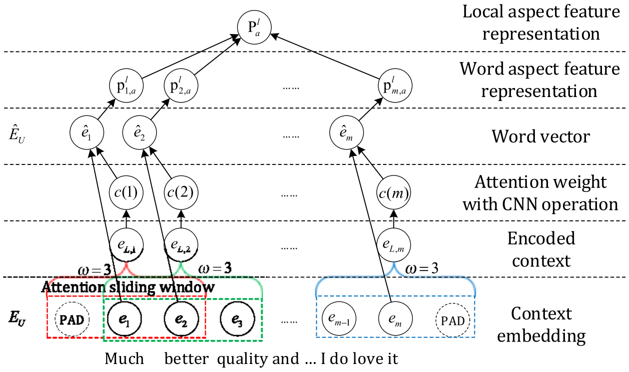


FIGURE 2. **Local aspect feature extraction module.**

In a sentence of a user/item review (e.g., "Much better quality and texture than expected for the price. The color is more of an off white, but I do love it."), aspect-related words (e.g., price, fit, color, material) and emotional words (e.g., affordable, comfortable, perfect, soft) usually occur close to each other. Essentially, the importance of the i-th word in a review depends on both the word itself and its surrounding words. We can infer the importance of the word by looking at its surrounding words. As shown in Figure 2, an attention sliding window with a width of ω is placed at the i-th word and then slid across its embedding vector E_U to learn the importance of the word [4]. Specifically, the review context is encoded by using a window spanning $(\omega - 1)/2$ words on both sides of each word. $c(i) \in \mathbb{R}^d$ is the attention weight of i-th word embedding, which judges the importance of the word in the sentence. The smaller the value of $c(i)$, the lower the importance of the word. $c(i)$ can be computed with convolution parameter matrix $\mathrm{W}_L \in \mathbb{R}^{\omega \times d}$ and bias b_L as follows [2], [62]:

$$c(i) = \delta(e_{L,i} * \mathrm{W}_L + \mathrm{b}_L) \tag{1}$$

$$e_{L,i} = (e_{i+\frac{-\omega+1}{2}}, e_{i+\frac{-\omega+3}{2}}, \cdots, e_i, \cdots, e_{i+\frac{\omega-3}{2}}, e_{i+\frac{\omega-1}{2}})^{\mathrm{T}} \tag{2}$$

where $*$ is the convolution operation, δ denotes a nonlinear activation function, which is the ReLU function in this study [17]. According to the attention weight, the word vector matrix with local attention weights can be expressed as follows:

$$\hat{E}_U = (\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_m) \in \mathbb{R}^{\omega \times d} \tag{3}$$

where the word vector $\hat{e}_i$ of the i-th word is computed as follows:

$$\hat{e}_i = c(i) e_i \tag{4}$$

In the fashion domain, possible aspect words can be price, category, color, texture, fabric, shape, part, style, etc., which are frequently mentioned in user reviews. In general, different users have different preferences over fashion, and a specific customer's attention may change with targeted fashion. For instance, a user may focus on the quality and fit, but does not care much about the price when selecting a suit. On the other hand, the same user may be more concerned about the style and color than the material when purchasing a dance skirt. For a specific item, different customers may purchase it with different intentions.

Since all words share the same d dimensions across the k aspects, we use the local aspect-specific word transform matrix $W_a^l \in \mathbb{R}^{d \times k}$ to change the word vector representation. W_a^l is initialized randomly by a uniform distribution $U(-0.01, 0.01)$. Thus, we can extract the local word aspect representation $\mathrm{p}_{i,a}^l \in \mathbb{R}^k$ from $\hat{e}_i$ as:

$$\mathrm{p}_{i,a}^l = \hat{e}_i W_a^l \tag{5}$$

It is common that the aspect-specific semantics of one word tend to express different polarities in the fashion domain. Therefore, we should capture this kind of words in the aspect representation. For example, many ‘great’ words in a review could lead to opposite sentiments in such phrases as “a great price” and “fits great”. Another word, ‘high,’ carries different sentiments in the contexts of “this clothing was made with high-quality materials and I would highly recommend” and “the price is too high.”

Considering the importance of learning each word in a review, the local aspect user representation $\mathrm{P}_a^l \in \mathbb{R}^k$ can be derived by an attention mechanism based on the following weighted sum:

$$\mathrm{P}_a^l = \sum\nolimits_i attn_{i,a}^l \mathrm{p}_{i,a}^l \tag{6}$$

$$attn_{i,a}^l = softmax((\mathrm{p}_{i,a}^l)^T \mathrm{v}_a^l) \tag{7}$$

where $softmax(\alpha_i) = \exp(\alpha_i)/\sum_i \exp(\alpha_i)$ (the same below), and $attn_{i,a}^l$ is defined as the i-th local attention vector (i.e. a probability distribution) over the review words for user u concerning aspect a. $\mathrm{v}_a^l \in \mathbb{R}^{k\times m}$ is the local aspect embedding matrix, and it is initialized randomly by a uniform distribution $U(-0.01, 0.01)$.

C. GLOBAL ASPECT FEATURE EXTRACTION MODULE

In addition to the local aspect feature extraction module, a parallel module is used to extract a set of global aspect user/item features. Inspired by the work in [25] and [30], we employed a bi-LSTM [31] to model the global long-term dependency and to find global aspect features by utilizing both previous and future contexts and by processing the sequence in both forward and backward directions. At each time step t, the output vectors of the two directions are concatenated. The global aspect feature extraction module is depicted in Figure 3. Firstly, the global long-term dependency of text sequence information is obtained based on the embedded context. Secondly, the aspect semantic meanings of the sentence are expressed by the global aspect-shared transform matrix. Finally, the global aspect representations are extracted by an attention mechanism.

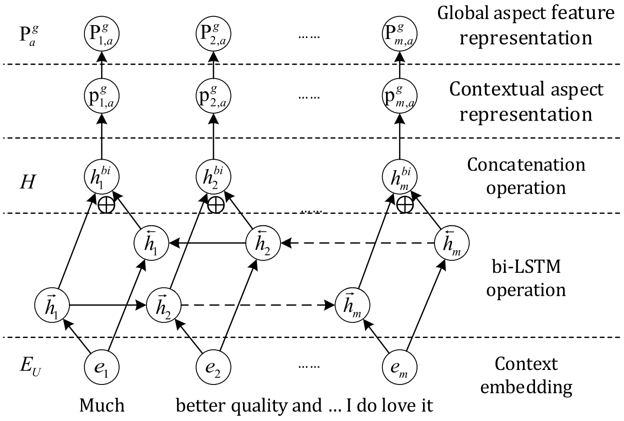


FIGURE 3. Global aspect feature extraction module.

Let d_h be the hidden size of a single direction LSTM. The hidden state $h_t \in \mathbb{R}^{d_h}$ of the LSTM at t can be updated in the following steps:

$$h_t = o_t \odot tanh(c_t) \tag{8}$$

$$o_t = \sigma(W_o \cdot E + b_o) \tag{9}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c \cdot E + b_c) \tag{10}$$

$$f_t = \sigma(W_f \cdot E + b_f) \tag{11}$$

$$i_t = \sigma(W_i \cdot E + b_i) \tag{12}$$

$$E = [\begin{matrix} h_{t-1} \\ e_t \end{matrix}] \tag{13}$$

where W_i, W_f, W_o and W_c are the weighted matrices; b_i, b_f, b_o and b_c are the biases to be learned during the training and parameterizing input gate i_t, forget gate f_t, output gate o_t, and cell state c_t, respectively; h_t is the hidden state of the LSTM; σ is the sigmoid function; $\odot$ stands for element-wise multiplication; and e_t is the input of the LSTM cell unit representing the word embedding vectors.

We then feed the input embedding text sequence $E_U = (e_1, e_2, \cdots, e_m)$ to the LSTM in the forward direction and obtain the forward hidden state $\vec{h}_t$. We also update the backward hidden state $\overleftarrow{h}_t$ by feeding the sequences into the LSTM in the reverse direction. The hidden states of the two directions are concatenated as follows:

$$h_t^{bi} = \vec{h}_t \oplus \overleftarrow{h}_t \tag{14}$$

where $t = 1, 2, \cdots, m$, and h_t^{bi} represents the global long-term dependency at t as it contains text sequence information from both directions. All the hidden states are collected into a matrix, which is denoted as:

$$H = [h_1^{bi}, h_2^{bi}, \cdots, h_m^{bi}] \tag{15}$$

where $H \in \mathbb{R}^{m\times 2d_h}$ and each row of H represents the global long-term dependency at the corresponding position of the input text sequence.

After that, we extract the contextual global aspect representation $\mathrm{p}_{i,a}^g$ from h_i^{bi} through a projection expressed as:

$$\mathrm{p}_{i,a}^g = h_i^{bi} W_a^g \in \mathbb{R}^k \tag{16}$$

where $W_a^g \in \mathbb{R}^{2d_h\times k}$ is the global aspect-shared transform matrix initialized randomly by $U(-0.01, 0.01)$., and k is the number of the aspects. The global aspect-specific representations can express the aspect semantic meanings of the sentence. $2d_h$ is the number of filters.

Because the importance of learning each word in the reviews varies, the global aspect user representation P_a^g can be derived with an attention mechanism:

$$\mathrm{P}_a^g = [\mathrm{P}_{1,a}^g, \mathrm{P}_{2,a}^g, \cdots, \mathrm{P}_{m,a}^g] \in \mathbb{R}^{m\times k} \tag{17}$$

$$\mathrm{P}_{i,a}^g = attn_{i,a}^g \mathrm{p}_{i,a}^g \tag{18}$$

$$attn_{i,a}^g = softmax((\mathrm{p}_{i,a}^g)^T \mathrm{v}_a^g) \tag{19}$$

where $attn_{i,a}^g$ is the global attention vector, which is defined over the words for user u concerning aspect a. $\mathrm{v}_a^g \in \mathbb{R}^{m\times k}$

is the global aspect embedding matrix, which is initialized randomly by $U(-0.01, 0.01)$.

D. MUTUAL OPERATION MODULE

To get the generalization capability for this model, we concatenate user's local aspects and global aspects extracted above with equation (20) and sum it using equation (21) with an attention mechanism in equation (22):

$$\bar{\mathrm{P}}_{i,a} = [(\mathrm{P}^{l}_{a} - \mathrm{P}^{g}_{i,a}) \odot (\mathrm{P}^{l}_{a} + \mathrm{P}^{g}_{i,a})] \in \mathbb{R}^{k} \tag{20}$$

$$\hat{\mathrm{P}}_{u,a} = \sum\nolimits_{i} attn_{i,a} \bar{\mathrm{P}}_{i,a} \in \mathbb{R}^{k} \tag{21}$$

$$attn_{i,a} = softmax(\bar{\mathrm{P}}_{i,a}) \tag{22}$$

where $\odot$ is an element-wise multiplication, and $attn_{i,a}$ is the mutual attention vector.

Then, we use fully connected layers W_1 and W_2 to calculate the user aspect mutual representation $\gamma_{u,a}$:

$$\gamma_{u,a} = \sigma(\hat{\mathrm{P}}_{u,a} W_1 + \mathrm{b}^1) W_2 + \mathrm{b}^2 \tag{23}$$

To improve the generalization performance, we adopt the dropout technique, which is widely used in existing neural models for recommendation [2], [4], [16].

The item mutual aspect representation $\gamma_{i,a}$ for item i and aspect a can be calculated similarly with equations from (1) to (23).

E. RATING PREDICTION AND OPTIMIZATION

Two parallel channels are used to learn representations of user aspect features and item aspect features, $\gamma_{u,a}$ and $\gamma_{i,a}$, which can be combined to create the overall rating, $\hat{r}$, as follows:

$$\hat{r} = \sum\nolimits_{a} (\gamma_{u,a})^{T} \gamma_{i,a} + b_u + b_i + b_0 \tag{24}$$

where b_u, b_i, and b_0 are the user, item, and global biases [4], respectively.

The above estimation can be considered as a regression problem in which all parameters are trained jointly through the backpropagation technique, in which the mean squared error (MSE) is used as a loss function. To learn the parameters of this model, the objective function, J, can be written as:

$$J = \sum (r - \hat{r})^2 + \lambda_{\Theta} \|\Theta\|^2 \tag{25}$$

where r is the known rating, $\hat{r}$ is the predicted rating, Θ denotes the set of all the parameters and $\lambda_{\Theta}\|\Theta\|^2$ is the regularization to prevent the model from overfitting. The stochastic gradient descent (SGD) algorithm and the backpropagation are used to optimize the parameters of the model, and the Adaptive Moment Estimation [62] is utilized over mini-batches.

IV. EXPERIMENTS

In this section, comprehensive experiments on two real-world review datasets are presented to evaluate the performance of AFRAM. The information about the datasets, the baseline methods, the experiment setup, and the results are elaborated.

A. DATASETS

The two review datasets used in the experiment are the Clothing, Shoes & Jewelry dataset (D1) from Amazon 5-core [63], and the dataset from one US online retailer (D2). The two datasets were filtered to ensure that each user or item review has at least one rating. The basic characteristics of these datasets are shown in Table 1 where #Rating, #User, and #Item are the numbers of ratings, users, and items in each dataset, Density = #Rating / (#User × #Item), and Sparsity = 1-Density. It can be seen that D2 has >10 times more the number of users that D1 has, while both have approximately equivalent numbers of items. D2's #Rating is almost twice D1's, and D2's Density is around 1/5 of D1's. D2's Sparsity is greater than D1's. We randomly partitioned each dataset into a training set (80%), a validation set (10%), and a test set (10%).

TABLE 1. Statistics of datasets used in this paper.

Dataset	#User	#Item	#Rating	Density (%)	Sparsity (%)
D1	39,387	23,033	278,677	0.0307	99.9693
D2	405,850	21,083	571,944	0.0067	99.9933

B. BASELINES

To verify the performance of the proposed model, AFRAM, we used the results of the following state-of-the-art rating prediction methods as the baselines.

1) DeepCoNN [17]: Deep collaborative neural network is based on two parallel CNNs to learn the latent feature vectors of user and item from reviews, and an FM to predict ratings.

2) D-Attn [2]: This model incorporates the local and global attention-based modules to select locally and globally informative words from reviews and to achieve the interpretability of latent features of user and item reviews.

3) NARRE [16]: The neural attentional regression model exploits two parallel CNNs and attention mechanisms to learn the latent features of users' and items' reviews to complete the rating prediction.

4) ANR [4]: ANR is based on the ideas of neural attention and co-attention by including an aspect-aware representation from a learning component and an estimator of aspect importance.

5) DAML [62]: The model utilizes local and mutual attentions of the attention mechanism of CNNs and the nonlinear of multi-layer perceptron (MLP) to achieve predictive rating for users.

C. EXPERIMENT SETUP

In the experiment with AFRAM, we set the dimension of the latent feature vector of user and item reviews at 300, the sliding window size at 5, the dropout rate at 0.5, the training batch size at 64, the number of aspects at 5 for both datasets, D1 and D2. For D1, the learning rate was set at 0.001, and the hidden size of the bi-LSTM at 5. For D2, the learning rate was set at 0.0008, and the hidden size of the bi-LSTM at 7.

The mean squared error (MSE) and the mean absolute error (MAE) were used as the evaluation metrics. The experiment was repeated five times to obtain stable results when the validation MSE was the lowest, and the average MSE and the average MAE of the five tests were calculated. Our experiment was executed with Pytorch 0.4.1 and Python 3.6 on a GPU machine of NVIDIA GeForce GTX 1080 Ti.

D. PERFORMANCE COMPARISON

The performance of AFRAM on the two given datasets was compared with those of five state-of-the-art recommendation models. Table 2 displays the experimental results from the DeepCoNN (a), D-Attn (b), NARRE (c), ANR (d), DAML (e), and AFRAM (f) models, in which the best result is highlighted by bold, and the second-best by underlined. Δ% denotes a relative difference in MSE or MAE between a baseline method and AFRAM, which measures the performance improvement by AFRAM (f).

TABLE 2. MSE and MAE performance on two datasets.

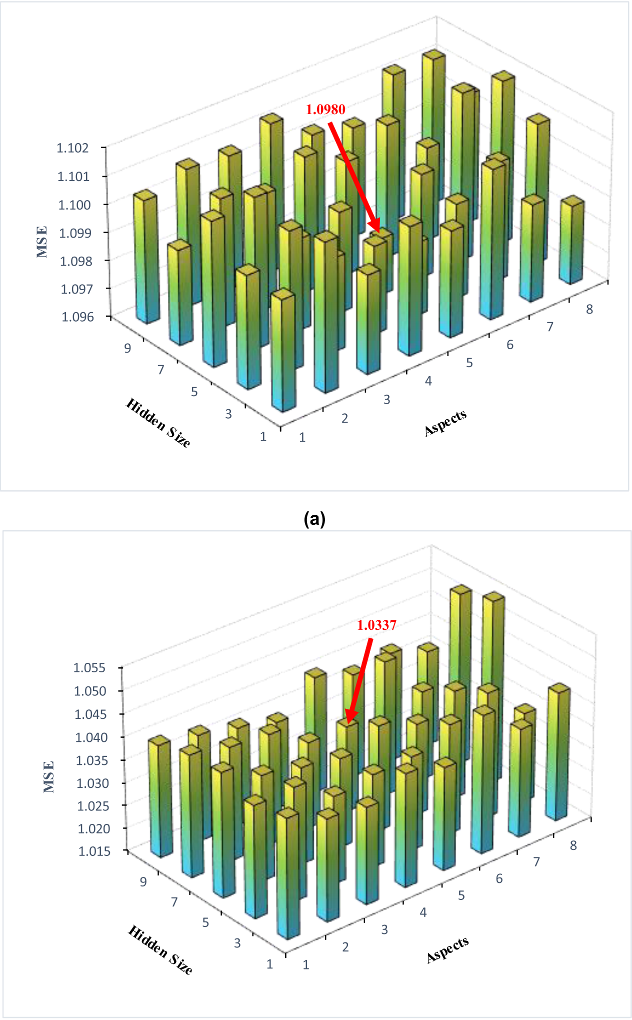
Methods		D1		D2	
		MSE	MAE	MSE	MAE
Error	(a)	1.1885	0.8409	1.2032	0.7820
	(b)	1.1644	0.8376	1.1640	0.7687
	(c)	1.1583	0.8222	1.1544	0.7993
	(d)	<u>1.1027</u>	<u>0.8139</u>	1.1397	<u>0.7651</u>
	(e)	1.1488	0.8402	<u>1.0967</u>	0.7871
	(f)	**1.0980**	**0.8016**	**1.0337**	**0.7258**
Δ%	(a)	7.61	4.67	14.09	7.19
	(b)	5.70	4.30	11.19	5.58
	(c)	5.21	2.51	10.46	9.20
	(d)	0.43	1.51	9.30	5.14
	(e)	4.42	4.59	5.74	7.79

The result shows that AFRAM (f) achieved the best MSE and MAE scores and outperformed the five baselines for both datasets. The second-best performer was ANR (d) on D1, from which AFRAM improved MSE by 0.43% and MAE by 1.51%, respectively. For D2, the second-best performer was DAML (e) if MSE was concerned or ANR (d) if MAE was concerned. But AFRAM improved MSE and MAE from the second-best performers by 5.74% and 5.14%, respectively. The experiment demonstrated that the aspect-based method with attention mechanism between the user and item reviews can offer more accurate rating predictions than the other five models on these two datasets.

V. MODEL ANALYSIS

A. INFLUENCES OF HIDDEN SIZE & NUMBER OF ASPECTS

The selections of the hidden size, d_h, and the number of aspects, k, have direct impacts on the rating predictions. In our experiment, the hidden size of AFRAM was chosen from {1,3,5,7,9}, and the number of aspects varied from 1 to 8. Figure 4 illustrates the effects of d_h and k on the predicted ratings for the reviews in these two datasets. The graphs show the optimal performances (smallest MSE) occurred on D1 (Clothing, Shoes & Jewelry dataset) when $d_h = 5$ and $k = 5$, and on D2 (2$^{\text{nd}}$ online retailer) when $d_h = 7$ and $k = 5$. $k = 5$ led the best performance in both cases. This is why we set k at 5 in our experiment.



(a)

(b)

FIGURE 4. Influences of hidden size and aspects. (a) Amazon dataset (D1). (b) 2nd online retailer's dataset (D2).

B. MODEL INTERPRETABILITY

Similar to what was found in [4], [6], the following five fashion aspects were most frequently captured among the reviews in the two datasets: $\mathcal{A} = \{$size/fit, color, fabric/texture, price, style$\}$. The 'background' distribution of each embedding word e_i is defined as $b_e = \sum_{a\in\mathcal{A}} \psi_a \big/ |\mathcal{A}|$ where $\psi_a = \sum_{u\in U} \psi_{u,a} \big/ |\mathcal{U}| + \sum_{i\in I} \psi_{i,a} \big/ |I|$. ψ_a refers to the importance of word e_i for aspect a, and $\psi_{u,a} = \sum_i attn_{i,a}$ is the importance of each word e_i by $attn_{i,a}$ with respects to user $u \in \mathcal{U}$ and aspect $a \in \mathcal{A}$ over vocabulary $\mathcal{V}$. Therefore, we can represent aspect a using its top words on $(\psi_a - b_e)$. Tables 3 and 4 list the top eight words used in the reviews for

TABLE 3. Top 8 words of each aspect in D1.

Size/fit	Color	Fabric/texture	Price	Style
length	colors	material	price	style
medium	black	shirt	deal	comfort
usually	white	soft	cost	classic
sizes	design	fabric	value	particular
normally	blue	cotton	paid	fashion
shape	red	lace	reasonable	unique
sizing	brown	sweater	bargain	statement
fitted	dark	smooth	unbeatable	styling

TABLE 4. Top 8 words of each aspect on D2.

Size/fit	Color	Fabric/texture	Price	Style
fit	color	soft	price	style
size	colors	smooth	quality	comfort
fits	shade	material	value	classic
usually	dark	shirt	deal	unique
medium	black	fabric	cost	styles
length	blush	thin	reasonable	particular
shape	pink	sheer	paid	styling
sizes	white	cotton	prices	fashion

TABLE 5. Influence of attention layers in AFRAM on MSE.

Dataset	AFRAM	AFRAM-Local	AFRAM-Global	AFRAM-Local&Global	AFRAM-Mutual	AFRAM-All
D1	1.0980	1.0994	1.0988	1.0984	1.1010	1.1020
D2	1.0337	1.0344	1.0344	1.0356	1.0359	1.0472

each aspect. These words in the five aspects properly reflect the relationships among the users, reviews, and ratings.

C. IMPACT OF ATTENTION LAYER

We analyzed the impacts of the local attention layer, global attention layer, and mutual attention layer. Table 5 provides the performance comparisons when attention layers in AFRAM changed and other parameter settings remained the same. AFRAM-local and AFRAM-global are AFRAM without the local or global attention layer. AFRAM-Local&global is AFRAM without local and global attention layers. AFRAM-mutual and AFRAM-All are AFRAM without the mutual attention layer or all attention layers.

AFRAM can obtain the best MSE on the two datasets when all the attention layers were included, and the worst MSE when all the attention layers were removed (AFRAM-All) on both datasets. In the other four scenarios (AFRAM-Local, Global, Local&Global, and Mutual), attention mechanisms performed differently on D1 and D2. Among them, the removal of the mutual attention mechanism had the greatest impact on MSE, indicating that the attention layers, especially the mutual attention layer, can improve the recommendation performance. This is because the mutual attention layer can combine different polarities of the same word in a sentence and the aspect semantic meanings of the whole sentence.

VI. CONCLUSION

In this paper, we presented an aspect-based fashion recommendation model with attention mechanism (AFRAM) to predict users' ratings based on users' reviews of purchased fashion products. This model used two parallel paths to extract latent aspect features about users and items separately and a mutual operation module to merge the two paths at the end for predicting users' ratings. On each path, there were a convolutional neural network (CNN) and a long short-term memory (LSTM) network, both having an attention mechanism, to capture local aspect features and global aspect features simultaneously. The mutual operations combining local and global aspect features in both user and item reviews greatly enhanced the generalization of the AFRAM model. As demonstrated in the experiment with real-world customer reviews and ratings collected from two renowned business websites, AFRAM outperformed the five state-of-the-art recommenders in terms of the accuracy of predicting customer ratings on fashion products.

ACKNOWLEDGMENT

The support provided by China Scholarship Council (CSC) to W. Li for his visit in University of North Texas is acknowledged.

REFERENCES

[1] Y. Lu, R. Dong, and B. Smyth, "Coevolutionary recommendation model: Mutual learning between ratings and reviews," in *Proc. World Wide Web Conf.*, 2018, pp. 773–782.

[2] S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *Proc. 11th ACM Conf. Recommender Syst.*, Aug. 2017, pp. 297–305.

[3] R. Catherine and W. Cohen, "TransNets: Learning to transform for recommendation," in *Proc. 11th ACM Conf. Recommender Syst.*, Aug. 2017, pp. 288–296.

[4] J. Y. Chin, K. Zhao, S. Joty, and G. Cong, "ANR: Aspect-based neural recommender," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 147–156.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[6] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 165–172.

[7] G. Ling, M. R. Lyu, and I. King, "Ratings meet reviews, a combined approach to recommend," in *Proc. 8th ACM Conf. Recommender Syst.*, 2014, pp. 105–112.

[8] Y. Bao, H. Fang, and J. Zhang, "Topicmf: Simultaneously exploiting ratings and reviews for recommendation," in *Proc. 28th AAAI Conf. Artif. Intell.*, vol. 2014, pp. 2–8.

[9] K. Song, W. Gao, S. Feng, D. Wang, K.-F. Wong, and C. Zhang, "Recommendation vs sentiment analysis: A text-driven latent factor model for rating prediction with cold-start awareness," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2744–2750.

[10] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[11] W. Ouyang, B. Xu, J. Hou, and X. Yuan, "Fabric defect detection using activation layer embedded convolutional neural network," *IEEE Access*, vol. 7, pp. 70130–70140, 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] T. Mikolov, M. Karafiát, L. Burget, J. Ernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. speech Commun. Assoc.*, 2010, pp. 1045–1048.

[14] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[15] C. Li, C. Quan, L. Peng, Y. Qi, Y. Deng, and L. Wu, "A capsule network for recommendation and explaining what you like and dislike," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 275–284.

[16] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in *Proc. World Wide Web Conf.*, 2018, pp. 1583–1592.

[17] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 425–434.

[18] A. Almahairi, K. Kastner, K. Cho, and A. Courville, "Learning distributed representations from reviews for collaborative filtering," in *Proc. 9th ACM Conf. Recommender Syst.*, 2015, pp. 147–154.

[19] Q. Cui, S. Wu, Q. Liu, W. Zhong, and L. Wang, "MV-RNN: A multi-view recurrent neural network for sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 317–331, Feb. 2020.

[20] C. He, Y. Liu, Q. Guo, and C. Miao, "Multi-scale quasi-RNN for next item recommendation," 2019, *arXiv:1902.09849*. [Online]. Available: http://arxiv.org/abs/1902.09849

[21] Q. Cui, S. Wu, Y. Huang, and L. Wang, "A hierarchical contextual attention-based network for sequential recommendation," *Neurocomputing*, vol. 358, pp. 141–149, 2019.

[22] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 197–206.

[23] R. Huang, S. McIntyre, M. Song, H. E, and Z. Ou, "An attention-based recommender system to predict contextual intent based on choice histories across and within sessions," *Appl. Sci.*, vol. 8, no. 12, p. 2426, Nov. 2018.

[24] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment analysis by capsules," in *Proc. World Wide Web Conf.*, vol. 2, 2018, pp. 1165–1174.

[25] Q. Ma, L. Yu, S. Tian, E. Chen, and W. W. Y. Ng, "Global-local mutual attention model for text classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2127–2139, Dec. 2019.

[26] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.

[27] R. Pascanu and T. Mikolov, "On the difficulty of training recurrent neural networks," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[28] T. Donkers, B. Loepp, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proc. 11th ACM Conf. Recommender Syst.*, Aug. 2017, pp. 152–160.

[29] S. Heinz, C. Bracher, and R. Vollgraf, "An LSTM-Based dynamic customer model for fashion recommendation," in *Proc. CEUR Workshop Proc.*, vol. 1922, 2017, pp. 45–49.

[30] J. Zeng, X. Ma, and K. Zhou, "Enhancing attention-based LSTM with position context for aspect-level sentiment classification," *IEEE Access*, vol. 7, pp. 20462–20471, 2019.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[32] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 83–92.

[33] N. Wang, H. Wang, Y. Jia, and Y. Yin, "Explainable recommendation via multi-task learning in opinionated text data," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 165–174.

[34] X. He, T. Chen, M.-Y. Kan, and X. Chen, "TriRank: Review-aware explainable recommendation by modeling aspects," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1661–1670.

[35] K. Bauman, B. Liu, and A. Tuzhilin, "Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 717–725.

[36] X. Guan, Z. Cheng, X. He, Y. Zhang, Z. Zhu, Q. Peng, and T.-S. Chua, "Attentive aspect modeling for review-aware recommendation," *ACM Trans. Inf. Syst.*, vol. 37, no. 3, pp. 1–27, Jul. 2019.

[37] Y. Wu and M. Ester, "FLAME: A probabilistic model combining aspect based opinion mining and collaborative filtering," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 199–208.

[38] J. Zhang, D. Chen, and M. Lu, "Combining sentiment analysis with a fuzzy kano model for product aspect preference recommendation," *IEEE Access*, vol. 6, pp. 59163–59172, 2018.

[39] Z. Cheng, Y. Ding, L. Zhu, and M. Kankanhalli, "Aspect-aware latent factor model: Rating prediction with ratings and reviews," in *Proc. World Wide Web Conf. World Wide Web*, 2018, pp. 639–648.

[40] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, Sep. 2014, pp. 3104–3112.

[41] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[42] X. Wang, L. Yu, K. Ren, G. Tao, W. Zhang, Y. Yu, and J. Wang, "Dynamic attention deep model for article recommendation by learning human Editors' demonstration," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 2051–2059.

[43] H. Li, M. R. Min, Y. Ge, and A. Kadav, "A context-aware attention network for interactive question answering," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 927–935.

[44] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, "Leveraging contextual sentence relations for extractive summarization using a neural attention model," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 95–104.

[45] M. Xu Chen, O. Firat, A. Bapna, M. Johnson, W. Macherey, G. Foster, L. Jones, N. Parmar, M. Schuster, Z. Chen, Y. Wu, and M. Hughes, "The best of both worlds: Combining recent advances in neural machine translation," 2018, *arXiv:1804.09849*. [Online]. Available: http://arxiv.org/abs/1804.09849

[46] J. Serrá, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4548–4557.

[47] Y. Tay, A. T. Luu, and S. C. Hui, "Multi-pointer co-attention networks for recommendation," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2309–2318.

[48] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with Item- and component-level attention," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 335–344.

[49] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 950–958.

[50] S. Shankar and S. Sarawagi, "Posterior attention models for sequence to sequence learning," in *Proc. 7th Int. Conf. Learn. Represent.*, 2019, pp. 1–11.

[51] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, "Retweet prediction with attention-based deep neural network," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 75–84.

[52] S. Kang, T. Phan, M. Bolas, and D. M. Krum, "Spatio-temporal wardrobe generation of actors' clothing in video content," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2016, pp. 448–459.

[53] M. He, S. Zhang, and Q. Meng, "Learning to style-aware Bayesian personalized ranking for visual recommendation," *IEEE Access*, vol. 7, pp. 14198–14205, 2019.

[54] H. Tuinhof, C. Pirker, and M. Haltmeier, "Image-based fashion product recommendation with deep learning," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, 2018, pp. 472–481.

[55] G.-L. Sun, Z.-Q. Cheng, X. Wu, and Q. Peng, "Personalized clothing recommendation combining user social circle and fashion style consistency," *Multimedia Tools Appl.*, vol. 77, no. 14, pp. 17731–17754, Jul. 2018.

[56] Y. Zhang, X. Liu, Y. Shi, Y. Guo, C. Xu, E. Zhang, J. Tang, and Z. Fang, "Fashion evaluation method for clothing recommendation based on weak appearance feature," *Sci. Program.*, vol. 2017, pp. 1–12, Oct. 2017.

[57] H. T. Nguyen, T. Almenningen, M. Havig, A. Kofod-petersen, H. Langseth, and H. Ramampiaro, "Learning to rank for personalised fashion recommender systems via implicit feedback," in *Proc. Mining Intell. Knowl. Explor.*, 2014, pp. 51–61.

[58] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli, "MMALFM: Explainable recommendation by leveraging reviews and images," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 1–28, Mar. 2019.

[59] Q. Liu, S. Wu, and L. Wang, “DeepStyle: Learning user preferences for visual recommendation,” in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 841–844.

[60] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, “Aesthetic-based clothing recommendation,” in *Proc. World Wide Web*, 2018, pp. 649–658.

[61] Googl. *word2vec*. Accessed: Jul. 29, 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/

[62] D. Liu, J. Li, B. Du, J. Chang, and R. Gao, “DAML: Dual attention mutual learning between ratings and reviews for item recommendation,” in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 344–352.

[63] Amazon. *Amazon Product Data*. Accessed: Feb. 14, 2020. [Online]. Available: http://jmcauley.ucsd.edu/data/amazon

WEIQIAN LI received the Ph.D. degree from the Xi’an University of Technology, Xi’an, China, in 2013. He is currently a Lecturer with the School of Computer Science, Xi’an Polytechnic University, in 2013. He is also a Visiting Scholar with the University of North Texas. His primary research interests include recommendation system and deep learning.

BUGAO XU received the Ph.D. degree from The University of Maryland at College Park, in 1992. He joined the faculty of The University of Texas at Austin, in 1993. Since 2016, he has been a Professor and Chair with the Department of Merchandising and Digital Retailing, and a Professor with the Department of Computer Science and Engineering, University of North Texas. His research interests include high-speed imaging systems, image and video processing, and AI in retailing.