DAML: Dual Attention Mutual Learning between Ratings and Reviews for Item Recommendation

Donghua Liu
School of Computer Science
Wuhan University
Wuhan, China
liudonghualdh@whu.edu.cn

Jing Li*
School of Computer Science
Wuhan University
Wuhan, China
leejingcn@whu.edu.cn

Bo Du*
School of Computer Science
Wuhan University
Wuhan, China
remoteking@whu.edu.cn

Jun Chang
School of Computer Science
Wuhan University
Wuhan, China
chang.jun@whu.edu.cn

Rong Gao
School of Computer Science
Hubei University of Technology
Wuhan, China
gaorong198149@163.com

ABSTRACT

Despite the great success of many matrix factorization based collaborative filtering approaches, there is still much space for improvement in recommender system field. One main obstacle is the cold-start and data sparseness problem, requiring better solutions. Recent studies have attempted to integrate review information into rating prediction. However, there are two main problems: (1) most of existing works utilize a static and independent method to extract the latent feature representation of user and item reviews ignoring the correlation between the latent features, which may fail to capture the preference of users comprehensively. (2) there is no effective framework that unifies ratings and reviews. Therefore, we propose a novel ***d**ual **a**ttention **m**utual **l**earning between ratings and reviews for item recommendation*, named DAML. Specifically, we utilize local and mutual attention of the convolutional neural network to jointly learn the features of reviews to enhance the interpretability of the proposed DAML model. Then the rating features and review features are integrated into a unified neural network model, and the higher-order nonlinear interaction of features are realized by the neural factorization machines to complete the final rating prediction. Experiments on the five real-world datasets show that DAML achieves significantly better rating prediction accuracy compared to the state-of-the-art methods. Furthermore, the attention mechanism can highlight the relevant information in reviews to increase the interpretability of rating prediction.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Computing methodologies** → *Neural networks; Factorization machines.*

KEYWORDS

Recommender systems; Neural network; Attention mechanism; Neural factorization machines; Rating prediction

ACM Reference Format:
Donghua Liu, Jing Li, Bo Du, Jun Chang, and Rong Gao. 2019. DAML: Dual Attention Mutual Learning between Ratings and Reviews for Item Recommendation. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD’19), June 22–24, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330906

1 INTRODUCTION

User ratings, as a kind of user feedback reflecting users’ interests, have been widely used to predict users’ preferences. Matrix factorization (MF) has achieved great success in learning user preferences [9, 24]. Relying on user-item interaction records, MF method represents users’ preferences and items’ features as latent factor vectors in a latent space for rating prediction. However, a rating only reflects a user’s overall satisfaction towards an item without explaining why the user assigns such a rating to the item, resulting in unexplained recommendations and cold-start problem [4, 22].

To tackle these limitations, researchers have paid attentions to the reviews. The reviews accompanying with the ratings typically contain a variety of information related to user preferences and item attributes. Recently, deep learning models have been proposed to extract effective latent features from the reviews for rating prediction [18, 30]. In these works, the convolutional neural networks (CNNs) are used to capture more effective contextual features of the reviews. Although these researches have achieved better recommendation performance than topic-based model, there are also some issues not thoroughly studied.

- Most of existing works have learned the latent features of users and items in a static and independent way [2, 18, 30]. However, the review text of users and items usually contains semantic information related to users and items, without considering the relevance of features between them. It may lead to great deviation to predict users’ preferences [26].
- There are two kinds of fusion methods. One is the traditional data fusion method based on MF or Factorization Machines (FMs) [16]. However, this method fails to capture the

*Jing Li and Bo Du are the corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD ’19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330906

complexity between features in different modalities [7, 21]. The other is a straightforward one that treats features extracted from different data sources equally, concatenating them sequentially into a feature vector for recommendation tasks. Nevertheless, simply a vector concatenation does not account for any interactions between latent features, which is insufficient for modelling the collaborative filtering effect[31].

To solve the above problems, we propose a DAML model. The model utilizes the attention mechanism of CNNs and the nonlinear of multi-layer perceptron (MLP) to achieve predictive rating for users. The contributions of this paper are summarized as follows:

- The proposed DAML model adopts dual attention layers: a local attention layer and a mutual attention layer. The former is used before the convolution layers to select informative words from a local window that contributes to the item attributes and user preferences. The latter is exploited after the convolution layer to learn the relevant semantic information between user review text and item review text that realize the dynamic interaction of users and items.
- We propose a unified neural network model to fuse ratings and reviews. Then neural factorization machines are used to realize the high-order nonlinear interaction of latent features. The model parameters are trained through end-to-end multitasking learning method.
- The experiments are performed on five real-world datasets and the experimental results show that the proposed DAML model achieves better rating prediction accuracy than the existing state-of-the-art methods. Further studies demonstrate that the words highlighted by DAML in reviews are extremely meaningful and can reveal users' specific preference towards an item, which helps to improve the interpretability of the recommender systems.

The remainder of this paper is organized as follows: Section 2 reviews the related work in the recommender systems. Section 3 presents the problem formulation; Section 4 introduces the overall framework of DAML model in details and together with it is the learning algorithm. The experimental settings and results are presented in Section 5. Section 6 concludes the paper.

2 RELATED WORK

Our work is related to two lines of literatures, the review text used for recommendation and neural networks. We review the recent advances in both areas.

2.1 Review Text for Recommendation

Traditional collaborative filtering recommendation algorithms have two significant drawbacks: one is data sparsity and the other is the cold-start problem. With the increase of interaction between users and system platforms, some auxiliary information closely related to users and items have been utilized for relieving the above drawbacks, especially, review text has become a research hotspot to improve the performance of recommender systems. Some works exploit topic modeling technique such as the Latent Dirichlet Allocation(LDA)[1] on the reviews and couple the latent topics and ratings [3, 4] which have shown significantly improvements over the baselines that exploits ratings or reviews alone. However, these studies ignore the word order and the local contextual information of the reviews, a lot of specific information in the form of phrases and sentences have been lost [9, 30]. Meanwhile, these studies adopt the linear methods rather than the nonlinear methods [8, 11, 19] to integrate reviews and ratings for rating prediction, which are not sufficient to capture the nonlinear and complex structure of feature interactions.

2.2 Deep Learning for Recommendation

Deep learning techniques have experienced great success in recommender systems. He et al. apply multiple layers of MLP to extract the high-level hidden feature by maximizing user-item interactions and realize the nonlinear interactions of user-item features [7, 8]. Wang et al. [14, 25] utilize neural networks to learn the deep feature representation of reviews and probabilistic matrix factorization (PMF) [17] to complete the rating prediction. However, this work still employs bag-of-words [5] representation to learn the latent topic. To improve the ability of deep feature representation, the word vector model and CNNs are used to learn user behaviors and item attributes representation [2, 9, 30]. However, they still learn the latent features of user and item in a static and independent way, which fails to learn the relevance of latent features. Inspired by the works of relationship prediction and deep learning [12, 18, 26], we propose a DAML model for item recommendation.

The proposed DAML model differs significantly from above models in several aspects. First, a dual attention mechanism is exploited to learn user-item interactions. The local attention layer focuses on the importance of different words in the sentences, while the mutual attention layer focuses on the learning of feature interactions. Second, instead of applying an attentive matrix to derive the relation of features, we propose a new method of defining a correlation scoring function to calculate the relevance of features, which can more intuitively capture the correlation between users and items. Then we exploit neural networks to unify ratings and reviews by some given rules for rating prediction. Our experimental results show that DAML delivers superior performance than the state-of-the-art methods.

3 PROBLEM FORMALIZATION

Figure 1 is the architecture of DAML model. Let u and $U = \{u_1, u_2, \cdots, u_i, \cdots, u_M\}$ denote a user and the whole user set respectively; similarly, i and $I = \{i_1, i_2, \cdots, i_j, \cdots, i_N\}$ are used to denote an item and the whole item set respectively. The rating of users to items $R \in \mathbb{R}^{M \times N}$ denotes their interactions, which can be real-valued explicit ratings or binary 0/1 implicit feedback. Here, we research the explicit ratings in which each $r_{u,i} \in R$ is a real-value.

Let x and $X = \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_i, \cdots, \mathrm{x}_M\}$ denote a user review text and the whole user review text set respectively; similarly, let s and $S = \{\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_j, \cdots, \mathrm{s}_N\}$ denote an item review text and the whole item review text set respectively. The DAML model can be formalized as follows:

Input: The input of interaction data is the identity of users and items. We use one-hot encoded sparse vector v_u^U and v_i^I that describe user $u \in U$ and item $i \in I$ respectively. The input of review

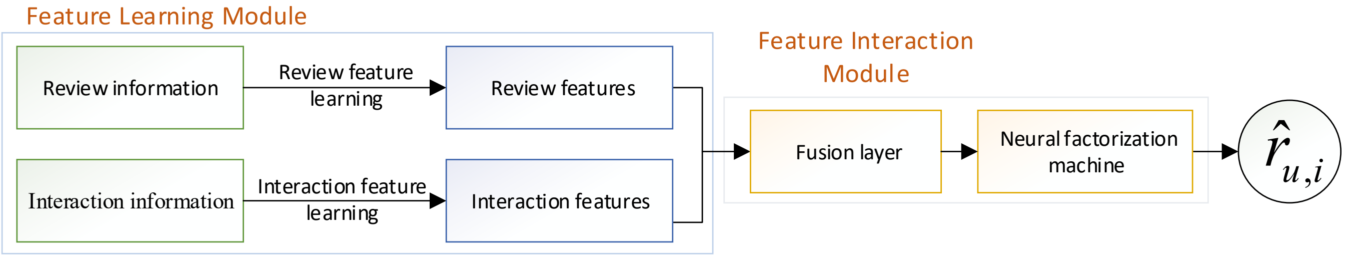


Figure 1: The architecture of DAML model.

data is $x_u \in X$ the review data set of user u and $s_i \in S$ the review data set of item i respectively.

Output: The whole training process can be expressed as a function: $f_u : U, X, I, S \rightarrow \hat{R}$. The output of the model is the final prediction rating $\hat{R}$. That is, for any user u, we can get the prediction rating $\hat{r}_{u,i}$ based on the function $f_u : v_u^U, v_i^I, x_u, s_i \rightarrow \hat{r}_{u,i}$.

4 THE PROPOSED MODEL

There are two parts in the model: one is feature learning module; the other is feature interaction module.

Feature learning module. The feature learning module consists of two components: review-based feature learning component and interaction-based feature learning component. The review-based feature learning component utilizes the attention mechanism of CNNs to learn the correlation between user preferences and item features. The interaction-based feature learning component is used to map users and items into low-dimensional dense vectors to capture the nonlinear interaction of features.

Feature interaction module. We integrate the rating and review features into a unified neural network model in the feature interaction module. Then the neural factorization machine is used to model the higher-order nonlinear interactions between latent feature vectors.

4.1 Feature Learning Module

Review-based Feature Learning. The review-based feature learning exploits the convolution operation and attention mechanism of CNNs to learn the relevant semantic information between user and item features. Figure 2 is the architecture of review-based feature learning.

Embedding look up layer: Given the review text $x_u \in X$, which is composed of l words and can express the complete meaning. In the embedding layer, the word vector model Glove [15] is used to map each word in x_u into word vector $\mathbf{w}_i$, then a review text matrix $\mathbf{D} \in \mathbb{R}^{d \times l}$ is formed by concatenating these words in order of their appearances in the review text and the order of words is preserved.

$$\mathbf{D} = [\cdots, \mathbf{w}_{i-1}, \mathbf{w}_i, \mathbf{w}_{i+1}, \cdots] \quad (1)$$

where d is the embedded dimension of each word, $\mathbf{w}_i$ represents the word vector of the $i{-}th$ word in x_u.

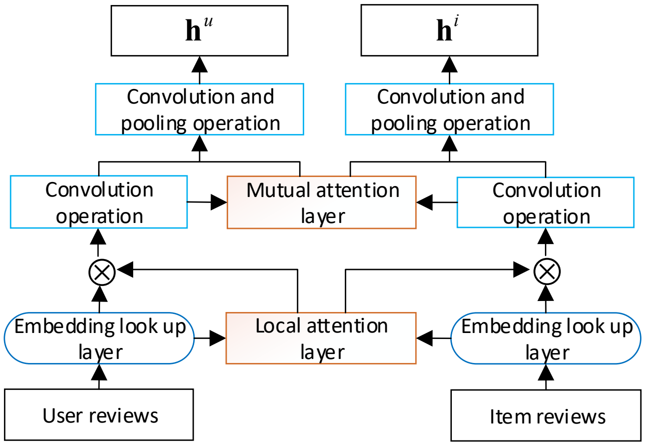


Figure 2: The architecture of dual attention mutual learning the relevance of user and item features. The ⊗ denotes element-wise product, $\mathbf{h}^u$ and $\mathbf{h}^i$ denote user feature and item feature respectively.

Local attention layer: Inspired by the work [18], we utilize an attention sliding window to learn the weight of each word in the reviews. We take the $i{-}th$ word in the word vector matrix $\mathbf{D} \in \mathbb{R}^{d \times l}$ as the center word and ω is the width of sliding window. The attention weight $s(i)$ of the $i{-}th$ word can be computed by the parameter matrix $\mathbf{W}_{L\_a}$ and bias b_{L_a} as follows:

$$\begin{aligned} \mathbf{w}_{L-a,i} &= (\mathbf{w}_{i+\frac{-\omega+1}{2}}, \mathbf{w}_{i+\frac{-\omega+3}{2}}, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_{i+\frac{\omega-3}{2}}, \mathbf{w}_{i+\frac{\omega-1}{2}}) \\ s(i) &= \delta(\mathbf{w}_{L-a,i}\mathbf{W}_{L\_a} + b_{L\_a}) \end{aligned} \quad (2)$$

where δ denotes the nonlinear activation function, we use the sigmoid function as the activation function. $s(i)$ is the attention weight of $i{-}th$ word embedding which can be used to judge the importance of the word in the sentence. According to the attention weight, the word vector $\hat{\mathbf{w}}_i$ of $i{-}th$ word can be computed as follows:

$$\hat{\mathbf{w}}_i = s(i)\mathbf{w}_i \quad (3)$$

where $s(i)$ and $\mathbf{w}_i$ are the attention weight and the word vector of the $i{-}th$ word respectively. The word vector matrix with local attention weights can be expressed as follows:

$$\hat{\mathbf{D}} = [\cdots, \hat{\mathbf{w}}_{i-1}, \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_{i+1}, \cdots] \quad (4)$$

Convolution operation: Given the word vector matrix $\hat{\mathbf{D}}$, the convolution operation is used to extract semantic information. Specifically, the sliding window size ω of the $j{-}th$ convolution filter is exploited to extract the local contextual feature c_i^j.

$$c_i^j = \mathbf{W}_c^j * \hat{\mathbf{D}}_{(:,\, i:(i+\omega-1))} \tag{5}$$

where $*$ is the convolution operation; $\mathbf{W}_c^j$ denotes the convolution weight vector for the $j-th$ convolution filter and $\hat{\mathbf{D}}_{(:,\, \mathrm{i}:(\mathrm{i}+\omega-1))}$ is the slice of matrix $\hat{\mathbf{D}}$ within the sliding window starting at $i-th$ position. To produce l contextual features, we first pad $\omega - 1$ zero vectors at the end of matrix $\hat{\mathbf{D}}$ before the convolution operation. Since shared weights in a convolution window can only capture one type of contextual feature, we use multiple convolution filters with different convolution weights to capture the context features for each word. After convolution operation, the contextual features at $i-th$ position can be expressed as $\mathbf{c}_i$:

$$\mathbf{c}_i = [c_i^1,\ c_i^2,\ \cdots, c_i^i,\ \cdots,\ c_i^f] \tag{6}$$

where c_i^f is the contextual feature produced by convolution filter f . Therefore, exploiting Equation (5), we can capture the contextual features of user review text and item review text respectively:

$$\begin{aligned} \mathbf{U} &= [\mathbf{c}_1^u, \mathbf{c}_2^u, \mathbf{c}_3^u, \cdots, \mathbf{c}_{l_u}^u] \\ \mathbf{V} &= [\mathbf{c}_1^i, \mathbf{c}_2^i, \mathbf{c}_3^i, \cdots, \mathbf{c}_{l_i}^i] \end{aligned} \tag{7}$$

where $\mathbf{c}_k^u$ and $\mathbf{c}_j^i$ are the contextual feature vectors for the $k-th$ word and the $j-th$ word in the user and item review text respectively. l_u and l_i are the lengths of the user and item review text respectively.

Mutual attention layer: Similar to [23, 26], we exploit a mutual attention layer to study the correlation between user review and item review. Specifically, we define a correlation scoring function $f_{relation-score}$ which exploits the Euclidean distance to calculate the relevance between the user contextual features $\mathbf{U}$ and and the item contextual features $\mathbf{V}$, the correlation scoring can be calculated as follows:

$$f_{relation-score} = 1/(1 + |\mathbf{c}_k^u - \mathbf{c}_j^i|) \tag{8}$$

According to the correlation scoring function, the user-item pair mutual attention matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ can be obtained by Equation (8):

$$\mathbf{A} = f_{relation-score}(\mathbf{U}, \mathbf{V}) \tag{9}$$

Each element in $\mathbf{A}$ denotes the correlation of user-item feature pair, and each row $\mathbf{A}_{k,*}$ in the correlation matrix denotes the correlation of each contextual feature $\mathbf{c}_k^u$ in $\mathbf{U}$ with the contextual features in $\mathbf{V}$. Similarly, $\mathbf{A}_{*,j}$ denotes the correlation between each contextual feature $\mathbf{c}_j^i$ in $\mathbf{V}$ with the contextual features in $\mathbf{U}$. The correlation weight g_k^u and g_j^i of contextual feature $\mathbf{c}_k^u$ and $\mathbf{c}_j^i$ can be computed as follows:

$$\begin{aligned} g_k^u &= \sum \mathbf{A}[k, :] \\ g_j^i &= \sum \mathbf{A}[:, j] \end{aligned} \tag{10}$$

Local pooling layer: Inspired by the work [28], we introduce attention weighting as an alternative, but exploit average pooling as a baseline as follows. Concretely, for the output feature map of the mutual attention layer, we conduct a row-wise averaging over slide window consecutive row to generate more abstract features of higher granularity, and we can get the contextual features with the weight of mutual attention:

$$\begin{aligned} \mathbf{t}_k^u &= \sum_{k=k:k+\omega} g_k^u \cdot \mathbf{c}_k^u \\ \mathbf{t}_j^i &= \sum_{j=j:j+\omega} g_j^i \cdot \mathbf{c}_j^i \end{aligned} \tag{11}$$

where $\mathbf{t}_k^u \in \mathbf{U}^u = [\mathbf{t}_1^u, \mathbf{t}_2^u, \mathbf{t}_3^u, \cdots, \mathbf{t}_{l_u}^u]$ and $\mathbf{t}_j^i \in \mathbf{V}^i = [\mathbf{t}_1^i, \mathbf{t}_2^i, \mathbf{t}_3^i, \cdots, \mathbf{t}_{l_i}^i]$ denote the contextual features of the $k-th$ position and the $j-th$ position of the user review and the item review respectively. In order to extract more abstract features and reduce the noise of irrelevant aspects, we stack a convolutional layer and an average pooling layer on the local pooling layer. The abstract features can be calculated as follows:

$$\begin{aligned} h_h^j &= \delta(\mathbf{W}_a^j \mathbf{U}_{h:h+\omega-1}^u + b_a^j) \\ h_h &= mean(h_1^j, \cdots, h_{l_u-\omega+1}^j) \\ \mathbf{h}^u &= [h_1, \cdots, h_f] \end{aligned} \tag{12}$$

where $\mathbf{W}_a^j$ denotes the convolution weight of the $j-th$ convolution filter, the size of the sliding window is ω, b_a^j is the bias, the feature of the $j-th$ convolution filter at position h is h_h^j. The final contextual feature $\mathbf{h}^u$ with user-item correlation is obtained. Similarly, we can obtain the item contextual feature $\mathbf{h}^i$. At last, we use a nonlinear function to map the contextual feature to dimensional space to complete recommendation tasks.

$$\begin{aligned} \mathbf{h}^u &= \delta(\mathbf{W}^u \mathbf{h}^u + b^u) \\ \mathbf{h}^i &= \delta(\mathbf{W}^i \mathbf{h}^i + b^i) \end{aligned} \tag{13}$$

where $\mathbf{W}^u$ and $\mathbf{W}^i$ denote the weight matrix of user and item contextual features respectively, b^u and b^i denote the bias.

Rating-based feature learning. The one-hot encoded item and user identity are taken as item feature vector v_i^I and user feature vector v_u^U to describe user and item respectively. Then the feature vector v_i^I and v_u^U are mapped to low-dimensional dense latent factor vectors through the latent factor matrix $\mathbf{P}_u \in \mathbb{R}^{M \times K}$ and $\mathbf{Q}_i \in \mathbb{R}^{N \times K}$ in the embedded layer, which are expressed as follows:

$$\begin{aligned} \mathbf{p}_u &= \mathbf{P}_u^T v_u^U \\ \mathbf{q}_i &= \mathbf{Q}_i^T v_i^I \end{aligned} \tag{14}$$

where $\mathbf{p}_u$ and $\mathbf{q}_i$ denote the interaction features of user and item respectively.

4.2 Feature Interaction Module

To capture comprehensive user preferences and item characteristics, we first fuse the two heterogeneous information features in the feature interaction module. The user and item features after fusion can be denoted as:

$$\begin{aligned} \mathbf{u}^u &= \mathbf{h}^u + \mathbf{q}_u \\ \mathbf{v}^i &= \mathbf{h}^i + \mathbf{p}_u \end{aligned} \tag{15}$$

where $\mathbf{u}^u$ and $\mathbf{v}^i$ denote the final user and item features respectively. We combine the features by concatenating them.

$$\mathbf{z} = \delta\, [\mathbf{u}, \mathbf{v}] \tag{16}$$

Where $\mathbf{z}$ denotes the user-item features, and δ is the activation function, [,] combines the two features by concatenating them in the hidden layer. Inspired by the work [7], we use the neural factorization machines to capture the high-order nonlinear interaction of features, the objective function is expressed as:

$$\hat{r}_{u,i}(\mathbf{z}) = m_0 + \sum_{j=1}^{|\mathbf{z}|} m_j z_j + f(\mathbf{z}) \tag{17}$$

In which $\mathbf{z} \in \mathbb{R}^{|\mathbf{z}|}$ denotes the user-item feature vector, $z_j \in \mathbf{z}$ is the value of the feature j, m_0 denotes the global bias, m_j denotes the coefficient for latent feature vector, $f(\mathbf{z})$ models the high-order interaction of features which can be expressed as:

$$f(\mathbf{z}) = \frac{1}{2}[(\sum_{j=1}^{|\mathbf{z}|} z_j \mathbf{v}_j)^2 - (\sum_{k}^{|\mathbf{z}|} z_k \mathbf{v}_k)^2] \tag{18}$$

where $z_j, z_k \in \mathbf{z}$ denote the $j-th$ and the $k-th$ user-item feature value, $\mathbf{v}_j, \mathbf{v}_k \in \mathbb{R}^s$ denote the embedding vector of feature j and k, and s is the embedding dimension. Similar to [27, 29], we capture the higher-order interaction between features by stacking the multi-layer full connection layer, the final predictive objective function can be expressed as follows:

$$\hat{r}_{u,i}(\mathbf{z}) = m_0 + \sum_{j=1}^{|\mathbf{z}|} m_j z_j + \mathrm{h}^{\mathrm{T}} \delta_{\mathrm{L}}(\mathrm{W}_{\mathrm{L}}(\cdots \delta_1(\mathrm{W}_1 f(\mathbf{z}) + \mathrm{b}_1) \cdots) + \mathrm{b}_{\mathrm{L}}) \tag{19}$$

where $\hat{r}_{u,i}(\mathbf{z})$ denotes the prediction score, model parameter $\Theta = \{m_0, \{m_j, \mathbf{v}_j\}, \mathrm{h}, \{\mathrm{W}_{\mathrm{L}}, \mathrm{b}_{\mathrm{L}}\}\}$, and δ_{L} represents the activation function. The added parameter $\{\mathrm{W}_{\mathrm{L}}, \mathrm{b}_{\mathrm{L}}\}$ is mainly used for the higher-order interaction of learning features compared with FM.

4.3 Joint Training

To learn the parameters of DAML model, we exploit the regression with squared loss as the objective function:

$$J = \sum_{(u,i) \in R} (\hat{r}_{u,i} - r_{u,i})^2 + \lambda_\Theta ||\Theta||^2 \tag{20}$$

where R is the user-item rating matrix, $r_{u,i}$ is the real rating of user u for item i, $\hat{r}_{u,i}$ is the prediction rating, Θ denotes all the parameters. $\lambda_\Theta ||\Theta||^2$ is used as regularization to prevent the model from overfitting. The entire framework can be effectively trained by using end-to-end paradigm reverse propagation. Algorithm 1 illustrates the training process of the DAML model.

Algorithm 1 Joint Training of DAML Model

```
Input: user set $U$, user reviews $X$, item set $I$, item reviews $S$. one-hot coding user rating features $v_u^U$ and item rating features $v_i^I$, user rating latent factor matrix $\mathbf{P}_u$ and item rating latent factor matrix $\mathbf{Q}_i$, the times of interaction $T$ and the size of each batch $B$.
Output: the prediction rating $\hat{r}_{u,i}$
begin
for $t \leftarrow 1$ to $T$ do
    Sample a batch $\beta_0$ of $(u, i, \mathrm{x}, \mathrm{s}, r_{u,i})$ of size $B$.
    $J = \sum_{(u,i) \in R} (\hat{r}_{u,i} - r_{u,i})^2 + \lambda_\Theta ||\Theta||^2$
    Take a gradient step to optimize $J$
end for
until $J$ converges or is sufficiently small
end
```

The parameters of DAML are optimized based on Equation (20) with stochastic gradient descent (SGD) and backpropagation. That is, the parameters for both two learning modules are jointly learnt. For parameter update, we utilize Adaptive Moment Estimation (Adam) [10] over mini-batches. Additionally, to prevent overfitting, we adopt dropout strategy [20] to the neural factorization machines.

4.4 Time Complexity Analysis

In the local attention layer, the time complexity is $O(p(l_u M + l_i N))$ where M and N are the number of user and item reviews, p is the embedded dimension, l_u and l_i are the length of user and item reviews respectively. In the convolutional layer, the time complexity of updating weights and bias variables is $O(p(ml_u M + nl_i N))$ where m and n denote the number of user and item contextual features respectively. In the mutual attention layer, the time complexity to calculate the user-item correlation is $O(p(ml_u M + nl_i N))$. For feature interaction module, the time complexity of feature interaction is $O(d(m + n))$, in which d is the dimension of interaction features. For hidden layer, the matrix-vector multiplication is the main operation which can be done in $O(d_{l-1} d_l)$, where d_l denotes the dimension of the $l-$th hidden layer. The complexity of prediction layer is $O(d_L)$. As a result, the total time complexity is $O(p(2ml_u M + 2nl_i N + l_u M + l_i N) + d(m + n) + \sum_{l=1}^{L} d_{l-1} d_l)$. We can see that the time complexity of the DACML model proposed in this paper is mainly related to the dimension and quantity of latent features when the number of users and items is fixed.

5 EXPERIMENTS

5.1 Datasets

We exploit five publicly available datasets that provide user reviews and ratings. The five datasets are from Amazon 5-core [6], which include reviews on Music Instruments, Office Products, Grocery and Gourmet Food, Video Games, Sports and Outdoors. Since the raw data is very large and sparse, we preprocessed it to ensure that all users and items have at least one rating. To alleviate the long tail effect of reviews, we follow the preprocessing steps used in [2] to adjust the length of reviews. The characteristics of these datasets are shown in Table 1.

In the experiments, we randomly split each dataset into three parts: training set (80%), validation set (10%) and test set (10%). The final performance comparison results derive from the test set.

5.2 Evaluation Metric

To evaluate the performance of DAML model, we exploit Mean Absolute Error (MAE) as evaluation metric,

$$MAE = \frac{1}{N} \sum_{(u,i) \in R} |r_{u,i} - \hat{r}_{u,i}| \tag{21}$$

where $\hat{r}_{u,i}$ denotes the prediction rating value, $r_{u,i}$ is the actual rating, and N denotes the number of tested ratings.

5.3 Experimental Scheme

To comprehensively evaluate the performance of our proposed DAML model, we design different strategies to evaluate the effectiveness of the model.

Table 1: Statistics of datasets used in this paper

Dataset	#users	#items	#ratings	Word per user	Word per item	density
Musical Instruments	1,429	900	10,261	170	163	0.798%
Office Products	4,905	2,420	53,228	203	172	0.449%
Grocery and Gourmet Food	14,681	8,713	151,254	177	155	0.118%
Video Games	24,303	10,672	231,577	148	135	0.089%
Sports and Outdoors	35,598	18,357	296,337	159	154	0.045%

Table 2: Performance comparison on five datasets for all methods. The best and the second best results are highlighted by boldface and underlined respectively. Δ% denotes the performance improvement of DAML over the best baseline.

Method	Musical Instruments	Office Products	Grocery and Gourmet Food	Video Games	Sports and Outdoors
PMF	1.137	1.265	1.397	1.395	1.203
NeuMF	0.7198	0.7301	0.9434	0.8693	0.7516
CDL	0.8336	1.062	0.9669	0.9018	0.8522
ConvMF	0.7860	0.7279	0.8634	0.8993	0.8235
DeepCoNN	0.7590	0.7109	0.8016	0.8752	0.7192
D-attn	0.7420	0.7161	0.8241	0.8422	0.7840
NARRE	0.6949	0.6807	0.7467	0.7991	0.6897
CARL	0.6766	0.6469	0.7534	0.7979	0.6864
DAML	**0.6510**	**0.6124**	**0.7354**	**0.7881**	**0.6676**
Δ%	3.78	5.33	1.51	1.23	2.74

1. **The comparison with existing state-of-the-art methods:** To evaluate the performance of the DAML model, we compare it with eight state-of-the-art recommendation models, so as to verify the effectiveness of the proposed model.

2. **The influence of components and parameters:** The important part of the DAML model: The dimension of the latent feature and impact of latent feature interactions in the feature interaction module are compared and analyzed.

3. **The analysis of review-based feature learning:** The local attention layer and the mutual attention layer are analyzed, and the latent features are visualized by heat maps to provide explanation for the recommendation results.

5.4 Baselines

To verify the performance of the DAML model proposed in this paper, we compared the model with the following state-of-art recommendation methods.

1. **PMF [17]:** Probabilistic matrix factorization is a standard matrix factorization that exploits ratings for prediction rating alone.

2. **NeuMF [8]:** Neural network-based Collaborative Filtering uses neural network to model the interaction between user and item latent features for prediction rating.

3. **CDL [25]:** Collaborative deep learning model exploits SDAE to extract effective deep features, and uses hierarchical Bayesian model to realize the integration of ratings and reviews for rating prediction.

4. **ConvMF [9]:** Convolutional matrix factorization model uses CNN to learn the contextual features of review text, then integrates the contextual features into the PMF for rating prediction.

5. **DeepCoNN [30]:** Deep collaborative neural network is based on two parallel CNNs to learn the latent feature vectors of user and item from user and item reviews respectively, and FM is used for rating prediction.

6. **D-attn [18]:** Dual attention mechanisms are used to achieve the interpretability of latent features of user and item.

7. **CARL [26]:** The context-aware user-item representation learning model exploits CNNs to learn the relevant features of user-item pairs. A dynamic linear fusion mechanism is utilized to derive the final rating score.

8. **NARRE [2]:** Neural attentional regression model exploits two parallel CNNs and attention mechanism to learn the latent features of reviews, and integrates the reviews and items to complete the rating prediction.

5.5 Parameter Setting

We exploit the grid search to tune the parameters for all the baselines based on the setting strategies reported by their papers. For PMF model, the Gaussian function is used to initialize user and item latent features. For deep learning model, the learning rate is tuned from [0.00001, 0.00002, 0.001, 0.002].The range of dropout ratio is searched in [0.1, 0.2, 0.3, 0.4]. The size of the training batch is tested in [50, 100, 150, 200], the dimension of latent features are turned in [8, 16, 32, 64, 128]. By adjusting, the number of latent features of CDL are set to 10 or 20, the noise size is set to 0.2. For the CNN text training model, this paper uses the hyperparameters set in the DeepCoNN model. The number of convolution kernels is 100, and the size of the sliding window is 3. In addition, the embedding dimensions of words in all models are 100. Without special statement, the number of latent features is 32 and the size of dropout ratio is 0.2.

As the proposed DAML model in this paper, the dimension of latent feature vector for user and item is 8, the sliding window size is set to 3, the dropout rate is set to 0.2, the learning rate is 0.00001, and the regularization parameter is tested in [0.001, 0.01,1,10,100]. All subsequently discussed components were implemented in Python3 using the TensorFlow library [1].

5.6 Performance Comparison

The overall performances of all contrast methods based on the Amazon 5-core datasets are shown in Table 2. We make the following observations from the results.

The MAE of the rating estimations are shown in Table 2 for the DAML model as well as the baseline models. We can see that the DAML outperforms the baselines on the five datasets. PMF performs the worst on these datasets. Similar method is NeuMF, yet it significantly outperforms the PMF. This indicates that the nonlinear interaction between latent features can offer better recommendation performance than the existing shallow models.

Among review-based baseline method. ConvMF, DeepCoNN and D-attn are three CNN-based methods that perform better than CDL, and such good performance should be attributed to CNNs, which ensure the contextual features in review text to be extracted successfully. The performance of D-attn model has great improvement than DeepCoNN, especially on the dataset: Musical Instruments and Video Games, which demonstrates that attention mechanism can help to capture users' preferences and items' properties. For other datasets, it delivers much worse performance when compared with DeepCoNN model. One possible explanation is that the D-attn model fails to model the high-order feature interaction which cannot fully capture the complexity of user-item feature interactions.

CARL achieves significantly improvement over DeepCoNN, D-attn and NARRE, which is consistent with the results reported in [26]. This indicates that learning the dynamic correlation of user and item reviews can yield a better understanding of user rating behaviors. The proposed DAML model achieves the best MAE scores on the five datasets. We can observe that the average improvement of DAML against the best baseline is 2.92%. This result demonstrates that DAML is effective for rating prediction on datasets with different features. Moreover, the significant performance gap between DAML and CARL validates that the distance metric method of calculating the relevance of features, which can more intuitively show the correlation between the features [13, 28]. And integrating ratings and reviews into a unify network and the high-order nonlinear interactions of DAML model can capture more knowledge about users' preferences.

5.7 Analysis of DAML Model

We mainly analyze: the dimension of latent factors and the interaction of latent features.

- **The impact of latent factor dimension number**

Figure 3 depicts the variation of latent factor dimension, which is tuned from [8, 16, 32, 64, 128]. We can see that with the change of the dimension number, the performance of the DAML model changes little. Even when the dimension number is much smaller, such as 8, the model achieves nearly optimal rating prediction accuracy in the five datasets, so we believe the latent factor dimensions have little effect on the experimental performance. With the increase of latent factor dimensions, the MAE of these five datasets shows a little improvement, while the time complexity of calculation has been increased. Besides, it may cause over-fitting and impact the performance of recommendation. Therefore, the dimension of the latent factor is set to 8.

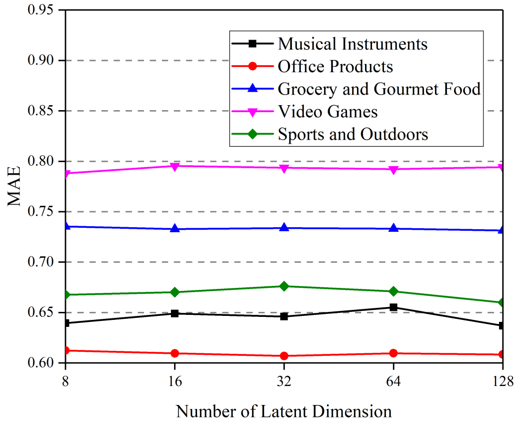


Figure 3: The plots of the impact of latent factor dimension number.

- **The impact of latent feature interactions**

Table 3 shows the impact of the high-order nonlinear feature interactions, we can see that the DAML model with the high-order nonlinear feature interactions outperforms the DAML with general factorization machines on the five datasets. This demonstrates that the high-order non-linear interaction function can capture the correlation between features in different modalities and improve the recommendation performance

5.8 Analysis of Review-based Feature Learning

This section analyzes the impact of attention layers. We use the heat map to visualize the latent features of attention layers.

- **The impact of attention layer**

Table 4 is the performance comparison with attention layers and without attention layers. We remove the attention layers and keep other parameter settings the same. We can see that the recommendation performance of the DAML is the best, the second is review-mutual, the third best is the review-local, review-outatt ranks the last in term of the recommendation list. The results indicate that the local attention layer can effectively distinguish the information words to reduce the noise disturbance and the mutual attention layer is able to improve the recommendation performance by identifying the relevance information for user-item pairs.

- **The visualization of Attention Layer**

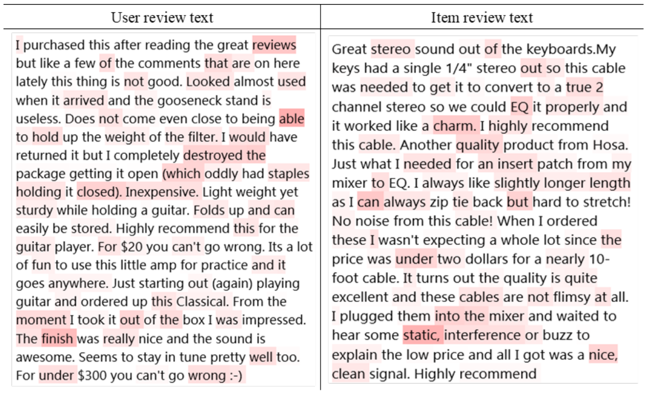
[1] https://www.tensorflow.org/

Table 3: The impact of latent feature interactions. DAML-FM: The DAML model with factorization machines. DAML: The DAML with neural factorization machines

Method	Musical Instruments	Office Products	Grocery and Gourmet Food	Video Games	Sports and Outdoors
DAML-FM	0.6818	0.6320	0.7437	0.7930	0.6828
DAML	0.6510	0.6124	0.7354	0.7881	0.6676

Table 4: The influence of attention layers. Review-outatt: review-based feature learning without local attention layer and mutual attention layer. Review-local: review-based feature learning with local attention layer and without mutual attention layer. Review-mutual: review-based feature learning with mutual attention layer and without local attention layer.

Method	Musical Instruments	Office Products	Grocery and Gourmet Food	Video Games	Sports and Outdoors
Review-outatt	0.7095	0.7540	0.9669	0.9018	0.8522
Review-local	0.6985	0.7138	0.7591	0.7958	0.6939
Review-mutual	0.6834	0.6945	0.7453	0.7864	0.6884
DAML	0.6510	0.6124	0.7354	0.7881	0.6676



User review text	Item review text
I purchased this after reading the great reviews but like a few of the comments that are on here lately this thing is not good. Looked almost used when it arrived and the gooseneck stand is useless. Does not come even close to being able to hold up the weight of the filter. I would have returned it but I completely destroyed the package getting it open (which oddly had staples holding it closed). Inexpensive. Light weight yet sturdy while holding a guitar. Folds up and can easily be stored. Highly recommend this for the guitar player. For $20 you can't go wrong. Its a lot of fun to use this little amp for practice and it goes anywhere. Just starting out (again) playing guitar and ordered up this Classical. From the moment I took it out of the box I was impressed. The finish was really nice and the sound is awesome. Seems to stay in tune pretty well too. For under $300 you can't go wrong :-)	Great stereo sound out of the keyboards.My keys had a single 1/4" stereo out so this cable was needed to get it to convert to a true 2 channel stereo so we could EQ it properly and it worked like a charm. I highly recommend this cable. Another quality product from Hosa. Just what I needed for an insert patch from my mixer to EQ. I always like slightly longer length as I can always zip tie back but hard to stretch! No noise from this cable! When I ordered these I wasn't expecting a whole lot since the price was under two dollars for a nearly 10-foot cable. It turns out the quality is quite excellent and these cables are not flimsy at all. I plugged them into the mixer and waited to hear some static, interference or buzz to explain the low price and all I got was a nice, clean signal. Highly recommend

Figure 4: The visualization of the local attention layer of user and item review text. The left is the user review text of a user, the right is the item review text of an item.

We randomly sample a user and an item review text from the Musical Instruments dataset to evaluation the performance of attention layers, the results are shown in Figure 4 and 5.

Figure 4 is the visualization of the local attention layer. We can see that the local attention layer can highlight some words with important information. Such as the word "weight", "guitar", "destroyed" and "Inexpensive" are highlighted in user review text, these words include the information about the item and the sentiment preference of the user for the item's different aspects. From the item review text, "stereo", "longer", "cable" and "quality" word which are related with the item have been selected. Hence, this case shows that the local attention layer can select words with important information about users and items from reviews.

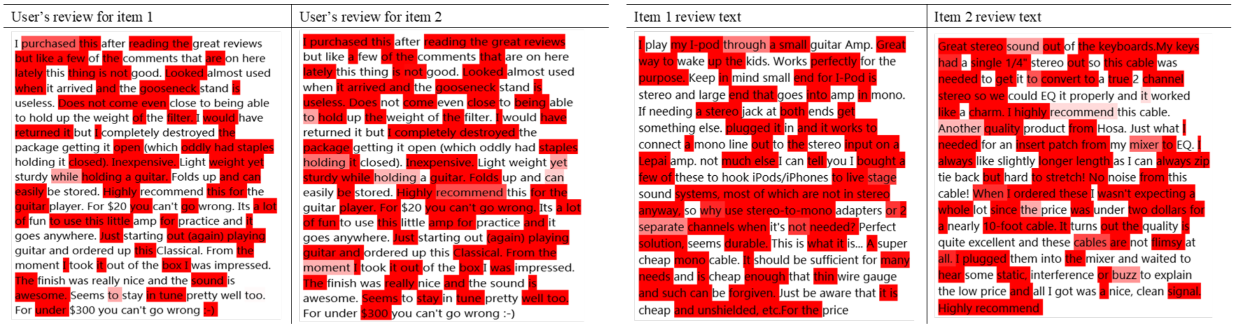
Figure 5 is the heat map of mutual attention layer. Figure 5(a) shows that the reviews of two different items written by the same user. For the review of item1, the words, such as "gooseneck", "weight" and "sound" are highlighted which indicate that the user pays more attention to the aspects of package, weight and sound. Some sentimental words like "useless", "destroyed" are found to be highlighted in the user review text for item2. From the highlighted information in the two heat maps of the same user, we can speculate that "weight", "sound" and "price" aspects about item1 are the focuses of the user, while "package" and "gooseneck" aspects about item2 could be more importance instead. In Figure 5(b), we then list the heat maps of the two item review texts confirm the speculations made above. For item1, it is not difficult to see that some aspects such as "small", "channels" are highlighted. From item2's review text, we can see that many aspects related to the keyboard, and the construction of the product are highlighted. The observed correlations between the heat map of user review text and item review text reveal that DAML can effectively capture relevant semantic information in the reviews of the user-item pair.

6 CONCLUSION

In this paper, we presented a novel dual attention mutual learning between ratings and reviews for item recommendation model. In this model, the local attention layer is developed to filter the review information and the mutual attention layer can learn the relevance of the user-item pairs. Besides, the ratings and reviews are unified into a neural network. Then the neural factorization machines are introduced in DAML to capture the high-order nonlinear interaction of the features. Experiments on five real-world datasets from Amazon show that our method consistently outperforms the existing state-of-the-art methods.

ACKNOWLEDGMENTS

The authors would like to thank anonymous referees and the editor for their thankful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China under Grants 61822113, 41871243, the Natural Science Foundation of Hubei Province under Grants 2018CFA050 and the National Key R&D Program of China under Grant 2018YFA0605501.



User's review for item 1

I purchased this after reading the great reviews but like a few of the comments that are on here lately this thing is not good. Looked almost used when it arrived and the gooseneck stand is useless. Does not come even close to being able to hold up the weight of the filter. I would have returned it but I completely destroyed the package getting it open (which oddly had staples holding it closed). Inexpensive. Light weight yet sturdy while holding a guitar. Folds up and can easily be stored. Highly recommend this for the guitar player. For $20 you can't go wrong. Its a lot of fun to use this little amp for practice and it goes anywhere. Just starting out (again) playing guitar and ordered up this Classical. From the moment I took it out of the box I was impressed. The finish was really nice and the sound is awesome. Seems to stay in tune pretty well too. For under $300 you can't go wrong :-)

User's review for item 2

I purchased this after reading the great reviews but like a few of the comments that are on here lately this thing is not good. Looked almost used when it arrived and the gooseneck stand is useless. Does not come even close to being able to hold up the weight of the filter. I would have returned it but I completely destroyed the package getting it open (which oddly had staples holding it closed). Inexpensive. Light weight yet sturdy while holding a guitar. Folds up and can easily be stored. Highly recommend this for the guitar player. For $20 you can't go wrong. Its a lot of fun to use this little amp for practice and it goes anywhere. Just starting out (again) playing guitar and ordered up this Classical. From the moment I took it out of the box I was impressed. The finish was really nice and the sound is awesome. Seems to stay in tune pretty well too. For under $300 you can't go wrong :-)

(a) The review text of user for two items

Item 1 review text

I play my I-pod through a small guitar Amp. Great way to wake up the kids. Works perfectly for the purpose. Keep in mind small end for I-Pod is stereo and large end that goes into amp in mono. If needing a stereo jack at both ends get something else. plugged it in and it works to connect a mono line out to the stereo input on a Lepai amp. not much else I can tell you I bought a few of these to hook iPods/iPhones to live stage sound systems, most of which are not in stereo anyway, so why use stereo-to-mono adapters or 2 separate channels when it's not needed? Perfect solution, seems durable. This is what it is... A super cheap mono cable. It should be sufficient for many needs and is cheap enough that thin wire gauge and such can be forgiven. Just be aware that it is cheap and unshielded, etc.For the price

Item 2 review text

Great stereo sound out of the keyboards.My keys had a single 1/4" stereo out so this cable was needed to get it to convert to a true 2 channel stereo so we could EQ it properly and it worked like a charm. I highly recommend this cable. Another quality product from Hosa. Just what I needed for an insert patch from my mixer to EQ. I always like slightly longer length as I can always zip tie back but hard to stretch! No noise from this cable! When I ordered these I wasn't expecting a whole lot since the price was under two dollars for a nearly 10-foot cable. It turns out the quality is quite excellent and these cables are not flimsy at all. I plugged them into the mixer and waited to hear some static, interference or buzz to explain the low price and all I got was a nice, clean signal. Highly recommend

(b) The review text of different users for the same item

Figure 5: The visual results of word highlighting by mutual learning attentive weights in the review texts of user and item respecitvely. (a) is the same user's reviews on different items. (b) is the reviews of two different items.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (Jan. 2003), 993–1022.

[2] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1583–1592.

[3] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan Kankanhalli. 2018. A3NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 3748–3754.

[4] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 639–648.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2493–2537.

[6] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 507–517.

[7] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 355–364.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 173–182.

[9] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional Matrix Factorization for Document Context-Aware Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 233–240.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Weiwei Liu and Ivor W. Tsang. 2017. Making Decision Trees Feasible in Ultrahigh Feature and Label Dimensions. *Journal of Machine Learning Research* 18 (2017), 81:1–81:36.

[12] Weiwei Liu, Ivor W. Tsang, and Klaus-Robert Müller. 2017. An Easy-to-hard Learning Paradigm for Multiple Classes and Multiple Labels. *Journal of Machine Learning Research* 18 (2017), 94:1–94:38.

[13] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. 2019. Metric Learning for Multi-Output Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 408–422.

[14] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Coevolutionary Recommendation Model: Mutual Learning Between Ratings and Reviews. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 773–782.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*. 1532–1543.

[16] Steffen Rendle. 2010. Factorization Machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE Computer Society, 995–1000.

[17] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 1257–1264.

[18] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 297–305.

[19] Ying Shan, T. Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. Deep Crossing: Web-Scale Modeling Without Manually Crafted Combinatorial Features. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 255–262.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 1929–1958.

[21] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2222–2230.

[22] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-boosted Latent Topics: Understanding Users and Items with Ratings and Reviews. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2640–2646.

[23] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. CANE: Context-Aware Network Embedding for Relation Modeling. 1722–1731.

[24] Chong Wang and David M. Blei. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 448–456.

[25] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative Deep Learning for Recommender Systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1235–1244.

[26] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, and Xiangyang Luo. 2019. A Context-Aware User-Item Representation Learning for Item Recommendation. *ACM Trans. Inf. Syst.* 37, 2 (Jan. 2019), 1–29.

[27] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 3119–3125.

[28] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. *Transactions of the Association for Computational Linguistics* 4 (12 2015).

[29] Wei Zhang, Quan Yuan, Jiawei Han, and Jianyong Wang. 2016. Collaborative Multi-level Embedding Learning from Reviews for Rating Prediction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2986–2992.

[30] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 425–434.

[31] Yu Zheng. 2015. Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions on Big Data* 1 (03 2015), 1–1.