MGAR: A Multi-Grained Attention Review-Based Recommendation System

Zhenhua Huang[1,2,3], Xinyu Guo[1], Wenhao Song[1], Zhaohong Jia*[1,4]

[1]Anhui University, Hefei, China.
[2]University of Science and Technology of China, Hefei, China.
[3]Dataspace Institute, Hefei Comprehensive National Science Center, Hefei, China.
[4]Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Hefei, China.

Contributing authors: zhhuangscut@gmail.com; xinyuguoustc@163.com; songwenhaoo@gmail.com; zhjia@mail.ustc.edu.cn;

Abstract

Recommendation systems have traditionally employed one of two paradigms to harness review information: the document-level approach, which amalgamates all reviews of a user or item into a single document, potentially overlooking the distinct significance of each review; and the text-level approach, which analyzes reviews individually to extract features pertinent to users or items. Recognizing the complementary nature of these paradigms, we propose the Multi-Grained Attention mechanism Recommendation system (MGAR), a novel framework that synergistically learns both document-level and text-level representations. Our model features a document encoder for assimilating document-level features and a text encoder that distills text-level attributes from individual reviews. In MGAR, we introduce an asymmetric cross-attention mechanism that captures the varying relevance of reviews for users and items, acknowledging the unique characteristics of each item and the differential importance of reviews. This mechanism leverages both self-attention and cross-attention to discern these nuances. Additionally, we incorporate user and item ID information to refine our predictive performance. Extensive experimental evaluations across diverse datasets confirm the superior efficacy of MGAR in leveraging review information for enhanced recommendation accuracy.

Keywords: Multi-Grained, Recommendation System, Review Text, Cross Attention

1 Introduction

Recommendation systems have become an indispensable tool for managing online information overload, aiding users in navigating vast selections with ease [1]. A foundational technique in these systems is collaborative filtering (CF), which leverages an interaction matrix to identify patterns in user-item relationships [2]. Despite its widespread adoption, CF struggles with providing reliable recommendations for users or items with sparse historical data, a challenge known as the cold start problem [3].

To mitigate the limitations, semantic information extracted from user reviews has been leveraged. Research has demonstrated that semantic analysis of review texts can significantly enhance the accuracy and effectiveness of rating predictions [4]. Reviews reveal intricate details about user preferences and item characteristics that pure numerical ratings cannot capture [5]. By analyzing the content of reviews for an item, one can glean insights into its key features and how they align with user preferences, thus offering a viable solution to the data sparsity and cold start challenges.

The methodologies for harnessing review information in recommendation systems are diverse and can be broadly classified into three categories [6]: Document-level, Text-level, and Graph-level Approaches. Document-level approaches are coarse-grained methods that aggregate all reviews of a user or an item into a singular document to learn respective representations [7–9]. For instance, DeepCoNN [7] employs parallel convolutional neural networks to mine semantic features from these comprehensive documents for rating prediction. D-Attn [10] uses local attention to learn the user's preferences and properties of items. However, these approaches assume equal significance across all reviews, overlooking the varying relevance of individual texts [11, 12].

Text-level Approaches: Contrasting the document-level, these fine-grained methods recognize the varying informativeness of individual reviews [11–16]. They model each review independently, prioritizing those deemed most informative for constructing user and item representations. For example, NARRE [11] applies an attention mechanism to discern the utility of each review post-extraction of features via CNN. MPRS [16] regards recommendation as a text-matching problem and computes the matching score for the given user-item pair by CNNs. These methods are adept at capturing nuanced attributes but may fall short in encapsulating the overarching user and item characteristics conveyed through the collection of reviews.

Graph-level Approaches: Recent methods exploit the inherent graph structure of user-item interactions [17–19]. Nodes in this graph represent users or items, with edges denoting the associated reviews. For instance, RGCL [17] is a graph contrastive learning framework that enhances recommendation performance by constructing a bipartite graph with embedded review features and employing self-supervised contrastive learning tasks to refine user and item representations. KGL [19] designs a relationship-aware knowledge embedding network to reflect the heterogeneity of relationships in the knowledge graph structure when aggregating item knowledge.

To synthesize a more holistic understanding of users and items, we introduce the Multi-Grained Attention mechanism Recommendation system (MGAR), which integrates document-level and text-level modeling into a cohesive framework. MGAR features a document encoder for capturing broad user and item representations and

a text encoder for distilling text-level features from individual reviews. By fusing these two feature sets, and incorporating user and item ID information, we achieve a comprehensive representation that encapsulates both macro and micro-level insights.

Furthermore, we innovate by integrating an asymmetric cross-attention mechanism within the text encoder, specifically designed to discern latent features in user and item reviews. This approach acknowledges that not all reviews have equal bearing on the representation learning process. To quantify the significance of each text, we employ a self-attention mechanism [20], which effectively weighs the importance of various texts associated with a particular item or user. Recent studies [21] highlight thematic discrepancies between user and item review sets, suggesting that the relevance of a user's reviews can vary significantly depending on the item in question. For example, a user's reviews of a book may better indicate their likelihood to purchase another book than their reviews of a mobile phone. It is, therefore, crucial to tailor user representations to the target item, eschewing static, uniform profiles in favor of dynamic, context-sensitive ones. To address this, we apply cross-attention [22] to refine the learning process. This mechanism leverages the representation of a specific item to calculate attention weights, thereby intuitively capturing the relative importance of different reviews. By adopting this asymmetric approach, we can derive more nuanced and expressive representations for both users and items, enhancing the recommendation system's predictive prowess. The main contributions of this paper are as follows:

– We introduce a **Multi-Grained Attention Mechanism** that operates on dual-levels of granularity. By assimilating both document-level and text-level features, our model achieves a comprehensive representation of users and items, enhancing the accuracy of feature extraction.

– A novel **Asymmetric Cross-Attention Mechanism** is proposed to extract latent review features more effectively. This mechanism employs self-attention to ascertain the significance of reviews pertaining to the same item, while cross-attention is utilized to refine the user review features, informed by the representations of items.

– Comparative experiments on six different domain datasets of Amazon verify the effectiveness of MGAR. The results demonstrate that MGAR outperforms existing baselines, achieving lower mean square error (MSE) values, thereby confirming its superior predictive capability.

2 Related Work

Our work is related to two lines of literature: document-level review-based recommendation systems and text-level review-based recommendation systems. We retrospect the recent advances in both areas.

2.1 Document-level Review-based Recommendation Systems

Several studies have proposed approaches that operate at the document level, offering a coarse-grained analysis of review documents. DeepCoNN Zheng et al. [7] employs dual parallel symmetric CNN networks to extract features from user and item review documents, and a factorization machine to user latent preferences and item attribute

features is modeled. TransNets [23] builds upon DeepCoNN by introducing an additional potential layer representing the target user's target item and it normalizes this layer during training and simulates another potential representation of the target user's review of the target item. D-Attn Seo et al. [10] uses local attention to learn users' preferences and properties of items and uses global attention to help CNN focus on the overall semantic information of review documents, respectively. CARL [24] derives a joint representation for a given user-item pair based on their latent features and latent feature interactions. ANR [4] models the multi-faceted process behind how users rate items by estimating the aspect-level user and item importance by adapting the neural co-attention mechanism. A^3NCF [8] extracts the features from the user and item review information by the theme model, and the dynamic selection and important aspects related to the target user and item are integrated with the ID of the user and item respectively. DAML [9] utilizes local and mutual attention of the convolutional neural network to jointly learn the features of reviews. Then the rating features and review features are integrated into a unified neural network model, and the higher-order nonlinear interaction of features is realized by the neural factorization machines to complete the final rating prediction

2.2 Text-level Review-based Recommendation Systems

Serveral works have been proposed in terms of fine granularity at the review text level. SentiRec [25] constructs two networks to encode the emotional information into the vector of review text information, and the review written by users and the review of items were used as the initial representation of users and articles respectively for the final rating prediction. NARRE [11] introduces a review-level attention mechanism and combines ID features for fusion based on DeepCoNN to achieve user and item feature modeling. NRPA [13] learns the representations of reviews and the representations of users and items from their reviews by integrating the id information of users and items. HUITA [14] uses three levels of Attention (word level, sentence level, and comment level)to learn the representation of users and items, respectively. MPCN [12] proposed using multiple pointer networks to learn more critical comments on current user item reviews. CAML [15] uses a three-stage structure of encoding - selection - decoding to generate user, item representation, and corresponding interpretation of recommendations. NRCA Liu et al. [6] proposes a cross-attention model for user representation learning with query vector embeddings of target item IDs to select different information words and reviews. SSIR Liu et al. [26] augments review representation with user reviews and neighbor-assisted review features of the same rating. MAN [27] uses the auxiliary network to focus on the purification of RT at the word level and assists the main network in generating the predicted value of RT.

3 Problem Formulation

We employ $S = \{U, P, R, T\}$ to denote the dataset. The set of users is defined as $U = \{u_1, \ldots, u_i, \ldots, u_n\}$, $U \in R^n$, where n represents the number of users. The set of items is defined as $P = \{p_1, \ldots, p_j, \ldots, p_m\}$, $P \in R^m$, where m is the number of items. The user-item rating matrix is defined as $R = \{r_{1,1}, \ldots, r_{i,j}, \ldots, r_{n,m}\}$, $R \in \mathbb{R}^{n \times m}$,

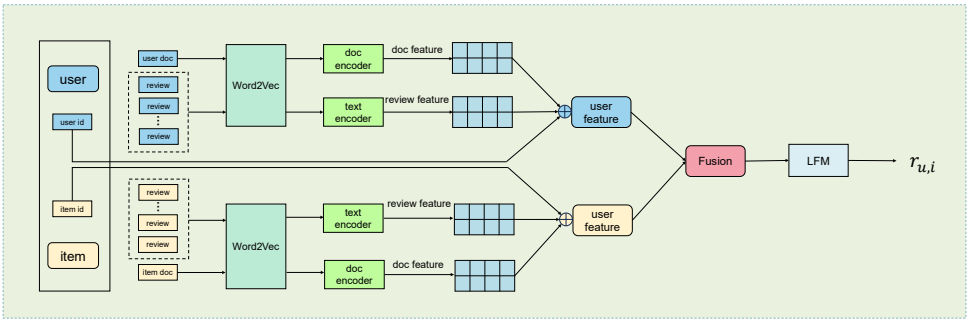

Fig. 1 The architecture of MGAR model.

where $r_{i,j}$ represents the real rating of item p_j by user u_i. The set of reviews is defined as $T = \{t_{1,1}, t_{1,2}, \ldots, t_{i,j}, \ldots, t_{n,m}\} \in R^{n \times m}$, where the element $t_{i,j}$ represents the review text posted by user i on item j.

Suppose that d_u and d_p denote the review documents of user u and item p, respectively. Here, d_u contains all review texts written by user u, and d_p contains all review texts belonging to item p, The main task of MGAR is to predict user ratings of items $\hat{r}_{u,p}$ from unseen history review information.

4 The Proposed Model

In this section, we describe MGAR in detail.

4.1 Framework

The MGAR framework is composed of four integral components:

Embedding Layer: This foundational layer utilizes the pre-trained Word2Vec model [28] to generate vector representations for each word within the review texts and documents, establishing the initial embedding space.

Encoding Layer: Comprising two specialized encoders, this layer processes review information at distinct granularities. The Review Encoder deals with document-level data, capturing the broader context of user and item reviews. The Text Encoder focuses on text-level data, distilling finer linguistic details from the reviews.

Fusion Layer: Here, the embeddings for user IDs and item IDs are amalgamated with the features extracted by the encoding layer. This synthesis creates a robust feature set that encapsulates both user and item characteristics.

Rating Prediction Layer: Employing the Latent Factor Model (LFM) [29], this layer synthesizes the inputs from previous layers to forecast user ratings for items.

Fig. 1 illustrates the structural design of MGAR, delineating the flow from raw input to predictive output.

4.2 Embedding Layer

The embedding layer in MGAR transforms the textual reviews into vectors by the following process:

1. Each word within a review is embedded into a d-dimensional vector space using a pre-trained model, resulting in word vectors that capture semantic meanings.

2. For a set of reviews T, this embedding process converts the reviews into a matrix of word vectors $D_{i,j} = \{w_1, w_2, ..., w_{n\times m}\}$, where each w_k is a d-dimensional word vector.

3. Considering the structure of the reviews, each $D_{i,j}$ is organized into a matrix with dimensions $s \times dm$, where s represents the fixed number of words in a review (the default review length) and dm is the dimensionality of the word embeddings.

4. Consequently, the embedding matrices for the review documents D_v and the review texts T_v are formed, with dimensions: $D_v \in \mathbb{R}^{n\times dm}$ for review documents, where n denotes the number of review texts within each document. $T_v \in \mathbb{R}^{s\times dm}$ for individual review texts.

These matrices serve as the input for the subsequent layers of the MGAR framework, enabling the system to interpret and analyze the reviews at both the document and text levels.

4.3 Encodering Layer

The encoding layer consists of two main parts: the document encoder and the text encoder.

4.3.1 Document Encoder

The document encoder operates in two phases, as depicted in Fig. 2: document word importance learning and text semantic information learning. First, it assesses the significance of each word within the document, and then it extracts semantic information from the text.

Document Word Importance Learning: To account for the varying significance of words and to capture long-range dependencies between them, we incorporate a self-attention mechanism [20]. This mechanism models the interrelationships within the sequence of document words to determine their respective weights.

Query vector Q, key vector K, and value vector V are obtained by element-wise multiplying the review document's word embedding matrix D_v.

$$Q = W^q D_v, \quad K = W^k D_v, \quad V = W^v D_v \tag{1}$$

where W^q, W^k, and W^v represent the transformation matrices for the query, key, and value, respectively. Subsequently, we apply scaled dot-product attention to ascertain the relevance scores for each word in the review document:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{ns}})V \tag{2}$$

To capture a richer representation of word information, we utilize a multi-head attention mechanism [30]. This approach performs attention operations in parallel, allowing the model to focus on different positions and representational subspaces:

The representation $head_i$ of the review document is calculated in each head vector space i.

$$head_i = Attention(Q_i, K_i, V_i) \tag{3}$$

Then the document representation $head_i$ in each subvector space is combined to get the review document vector. $d^* \in R^{s \times d}$.

$$Z = Concat(head_1, \ldots, head_h)W^O \tag{4}$$

$$d^* = LayerNorm(d_v + Dropout(Z)) \tag{5}$$

where h is the number of heads, $Concat$ is a concatenation operation, W^O is the multi-head attention layer parameter. $LayerNorm$ is a normalized layer. We use a fully connected feed-forward network with residual connections [31] to get semantic information D^* from d^*.

Text Semantic Information Learning: The semantic information learning module for document-level and text-level is the same. It consists of a convolutional layer, a max-pooling layer, and a fully connected layer. The review document sequence vector D^* is input to this module to extract the latent feature.

The convolutional layer consists of k convolutional kernels with different window sizes, denoted as $F = \{f_1, f_2, \ldots, f_k\}$. The padding method is used to ensure the length of the feature vector after convolution is the same as the length of the word sequence before convolution. $c_{j,i}$ denotes the contextual representation of the word w_i after the j-th convolution kernel operation.

$$c_{j,i} = \sigma\left(f_j * w_{i-\frac{t}{2}:i+\frac{t}{2}} + b_j\right) \tag{6}$$

where the symbol $*$ is the convolution operation, f_j and b_j are the convolution kernel parameters, t is the window size. σ is a nonlinear activation function such as $Relu$.

The second layer is the max-pooling layer. Based on the assumption that the most representative semantic feature information may exist in different positions of the sequence, the max-pooling operation is used to extract the feature information at other positions as the semantic information of the document.

$$c_j = max\{c_{j,1}, c_{j,2}, \ldots, c_{j,s}\} \tag{7}$$

where c_k symbolizes the feature representation of the review document post convolution and max pooling by the convolution kernel f_k. The outputs from the convolution and pooling of each word are aggregated to construct a holistic document representation ($c \in \mathbb{R}^k$):

$$c = [c_1 : c_2 : \ldots : c_k] \tag{8}$$

The fully-connected layer integrates and nonlinearly transforms the features extracted by the convolution and pooling layers. The final embedding of the user review

document and item review document features $\Theta_u \in R^k$ and $\Theta_p \in R^k$ can be obtained.

$$\Theta = \sigma(Wc + b) \tag{9}$$

where W is the weight matrix and b is the bias term, and σ is a nonlinear activation function such as $ReLU$.

4.3.2 Text Encoder

As shown in Fig. 3, the text encoder contains three stages: word importance learning, text semantic information learning and review texts importance learning.

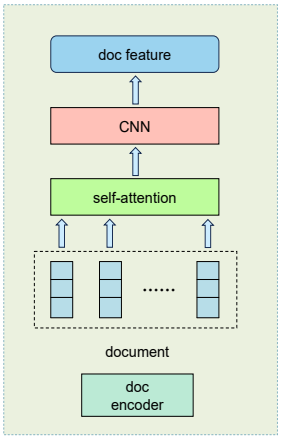

Fig. 2 The architecture of doc encoder.

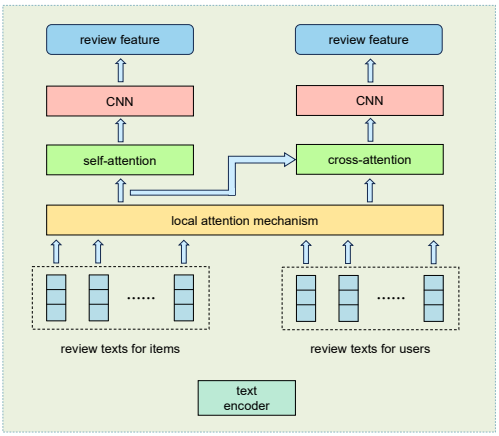

Fig. 3 The architecture of text encoder.

Word importance learning of texts: In contrast to review document sequences, word sequences in a single review text are short and have a single topic. Inspired by (D-Attn) work [10], a local attention mechanism with sliding windows is applied to learn the importance of each word. Sliding window sizes can be set differently to learn different ranges of local features.

$$\begin{aligned} t' &= \{w_1, \ldots, w_i, \ldots, w_s\} \\ \phi_i &= (w_{i+\frac{-w}{2}:i+\frac{-w}{2}}) \\ \alpha(i) &= \sigma(W\phi_i + b) \end{aligned} \tag{10}$$

where w_i is the embedding representation of the i-th word in the review text, Φ_i is a window of size ω, W is the learning parameter matrix, and b is the bias. σ is the $Relu$ activation function.

Based on the attention weight of each word, the embedding representation of the i-th word can be calculated as $\hat{w}_i$.

$$\hat{w}_i = \alpha\left(i\right) w_i \tag{11}$$

while $\alpha(i)$ is the attention weight of the i-th word and w_i is the corresponding embedding.

After learning by the local attention mechanism, each text can be represented as $t^* \in R^{s \times dm}$, where s is the sequence length and dm is the word embedding dimension.

$$t^* = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_s\} \tag{12}$$

Then semantic information module is used to kick out the semantic features of each review text. The feature vector for each review text is $\varepsilon \in R^k$.

$$\varepsilon = \{o_1, o_2, \ldots, o_k\} \tag{13}$$

where k is the number of convolution kernels.

$$\begin{aligned} \theta_u &= \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\} \\ \theta_p &= \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\} \end{aligned} \tag{14}$$

The review text feature of user u is denoted as $\theta_u \in R^{m \times k}$, where m is the number of reviews posted by the user. Similarly, $\theta_p \in R^{n \times k}$ denotes the feature representation extracted from the review text for each item, where n is the number of review texts for the item.

Review Texts Importance Learning: To discern the relative significance of each review text, we introduce an asymmetric cross-attention module. This module is bifurcated into a self-attention mechanism for item review texts and a cross-attention mechanism for user review texts.

(1) Self-Attention module: The reviews received by an item are topically related to the item itself, while not all review texts are equally important. The self-attention mechanism is used to learn the weights of different reviews of the same item. The review text feature of item p can be represented as $h_p \in R^k$.

$$\begin{aligned} g_j &= w_o \varepsilon_j \\ \alpha_j &= \frac{\exp(g_j)}{\sum_{i=1}^{n} \exp(g_i)} \\ h_p &= \sum_{j=1}^{n} \alpha_j \varepsilon_j \end{aligned} \tag{15}$$

where ε_j is the feature vector of j-th review text, w_o is the linear transformation matrix, g_j is the attention score, and α_j is the importance weight.

(2) Cross-Attention Module: When considering reviews authored by a user, the cross-attention mechanism comes into play to ascertain the significance of each text. In this context, the features extracted from item review texts serve a guiding role, steering the learning process to recognize the importance of user reviews concerning the items they discuss.

The query vector Q is calculated by dotting the item review text feature vector θ_p with the query transformation parameter W^Q. The guide vector G is obtained by

dotting the user review feature vector θ_u with the W^G parameter. And the value vector V is obtained by dotting the user review feature vector θ_u with the W^V parameter.

$$Q = W^Q \theta_p G = W^G \theta_u V = W^V \theta_u \tag{16}$$

Then, the similarity of Q and G is compared by Dot-Product Attention calculation and normalized by the softmax function.

$$Attention(Q, G, V) = softmax(\frac{QG^T}{\sqrt{d_k}})V \tag{17}$$

where d_k is the dimension of Q.

The weighted sum of all values in V is calculated for the calculated weights to obtain the feature $head_i$ of the user review features in i-th subspaces.

$$head_i = Attention(Q_i, G_i, V_i) \tag{18}$$

A multi-head attention mechanism is introduced to capture the information of users' review text features in different subspaces. Multiple $head_i$ are linearly combined to represent the latent feature vector $h_u \in R^k$ of the end-user review text.

$$h_u = Concat(head_1, \ldots, head_h)W^o \tag{19}$$

where W^o is the linear transformation matrix.

Text Semantic Information Learning: The module also contains three parts the same as that in the document encoder: a convolutional layer, a max-pooling layer, and a fully connected layer.

4.4 Fusion Layer

The encoding process within our recommendation system operates at two distinct granularities. Firstly, the document encoder is tasked with capturing the global feature information inherent in the review document as a whole. Secondly, the text encoder delves into the minutiae, extracting fine-grained features from the individual review text. These dual-level features are then coalesced within the fusion layer, creating a comprehensive feature set. To further enrich the feature representation, we integrate user and item identification information. This integration serves to contextualize the global features within the specific user-item interaction framework. The ultimate feature representations for user u and item p are symbolized as z_u and z_p, respectively.

$$\begin{aligned} z_u &= W_u^T(\Theta_u \oplus h_u \oplus ID_u) + b_u \\ z_p &= W_p^T(\Theta_p \oplus h_p \oplus ID_p) + b_p \end{aligned} \tag{20}$$

where the symbol $\oplus$ is the concatenation operation, Θ_u and Θ_p are the review document features of the user and item, respectively. h_u and h_p are the review text features

Table 1 Statistical details of datasets

Datasets	User	Items	Ratings	sparsity
Digital Music	5541	3568	64706	0.33%
Video Games	24303	10672	231780	0.09%
Tools Improvement	10217	3568	134476	0.08%
Toys and Games	11924	3568	167597	0.07%
Office Products	2420	3568	53258	0.45%
Automotive	2928	1835	20473	0.38%

of the user and item, respectively. $ID_u \in R^k$ and $ID_p \in R^k$ are the vector representation of the user ID and item ID after embedding. W_u^T and W_p^T are the linear transformation matrix. b_u and b_p are the bias.

4.5 Rating Prediction Layer

We use the Latent Factor Model (LFM)[29] approach for user and item feature fusion to achieve rating prediction. The LFM prediction rating formula is as follows:

$$\hat{r}_{u,p} = W^T(z_u + z_p) + b_u + b_p + \mu \tag{21}$$

where W^T is the linear transformation matrix, b_u is the user bias, b_p is the item bias, and μ is the global bias.

5 Experiments

Our experimental evaluation aims to assess the performance of the Multi-Grained Attention Recommendation (MGAR) system against established baselines using six authentic datasets. We juxtapose MGAR with a contemporary baseline to underscore its efficacy and detail the outcomes of our performance analysis. Subsequently, we dissect the influence of various model parameters and components on the system's performance.

5.1 Experiment Datasets

The datasets are from the 5-core dataset provided by Amazon: Digital Music, Video Games, Tools Improvement, Toys and Games, Office Products, and Automotive. Each user and item in the dataset has at least five reviews. Table 1 shows the statistics of the datasets.

5.2 Evaluation Metrics

Mean squared error (MSE) is used to evaluate models' performance [32].

$$MSE = \frac{1}{|S_t|} \sum_{u,p \in S_t} (r_{u,p} - \hat{r}_{u,p})^2 \tag{22}$$

where S_t is the instance in the test set, $r_{u,p}$ denotes the actual rating, and $\hat{r}_{u,p}$ denotes the model prediction rating.

Table 2 MSE comparison between MGAR with baselines

	Document-Feature				Review-Feature			Both
Datasets	DeepCoNN	DAML	D-Attn	SSIR	MAN	MPCN	NARRE	MGAR
Digital Music	1.053	0.8261	0.8365	0.8120	0.8701	0.9385	0.8120	**0.7910**
Video Games	1.201	1.1335	1.1241	1.1134	1.1458	1.2680	1.1115	**1.0952**
Tools Improvement	1.0360	0.9394	0.9529	0.9587	0.9734	1.0035	0.9511	**0.9241**
Toys and Games	0.8890	0.8488	0.8903	0.8043	0.8384	0.9024	0.7801	**0.7710**
Office Products	0.8607	0.7293	0.7404	0.7006	0.7727	0.7679	0.7207	**0.6812**
Automotive	0.8734	0.8336	0.8246	0.8672	0.8701	0.8344	0.8323	**0.8242**

5.3 Baselines

We considered the following strong baselines:

DeepCoNN [7]: It uses two parallel symmetric convolutional neural networks to extract potential feature vectors from user and item reviews, and the fusion layer uses FM to achieve rating prediction.

D-Attn [10]: It applies feature interaction components to learn correlations between user preferences and item features and uses a neural factorization machine (LFM) to achieve rating predictions.

MPCN [12]: This method proposes a review-based learning solution that extracts critical reviews from user and item reviews. It proposes a multi-pointer learning scheme for learning multiple views of combined user-item interactions.

NARRE [11]: This method uses the review-level attention mechanism to learn the weights of different review features based on DeepCoNN, fuses review text features with id features and finally uses FM to achieve rating prediction.

DAML [9]: It uses the local attention layer and the mutual attention layer to learn the comment features, and captures the higher-order interaction between features by stacking multiple full-connection layers to achieve score prediction.

SSIR [26]: It predicts ratings from reviews and ratings from specific-view, and shared-view, and uses auxiliary reviews to handle review sparsity.

MAN [27]: It uses two different parallel networks, the main network, and the auxiliary network to process specific review information.

5.4 Parameter Setting

To ensure an equitable performance comparison, we aligned the model parameters of the baseline models with those specified in their original formulations. For the MGAR model, we configured the following parameters. The embedding representation dimension of each word is set to 300. The convolutional kernel size is chosen from [3,4,5,6], and the number of convolutional kernels is set to 50. The dimensionality of the feature encoding representation of users and items is set as 128. The learning rate is set as 0.003 with the Adam optimizer. The batch size is set to 128.

5.5 Performance Evaluation

Table 2 shows the final results of the baselines and the proposed MGAR after optimization.

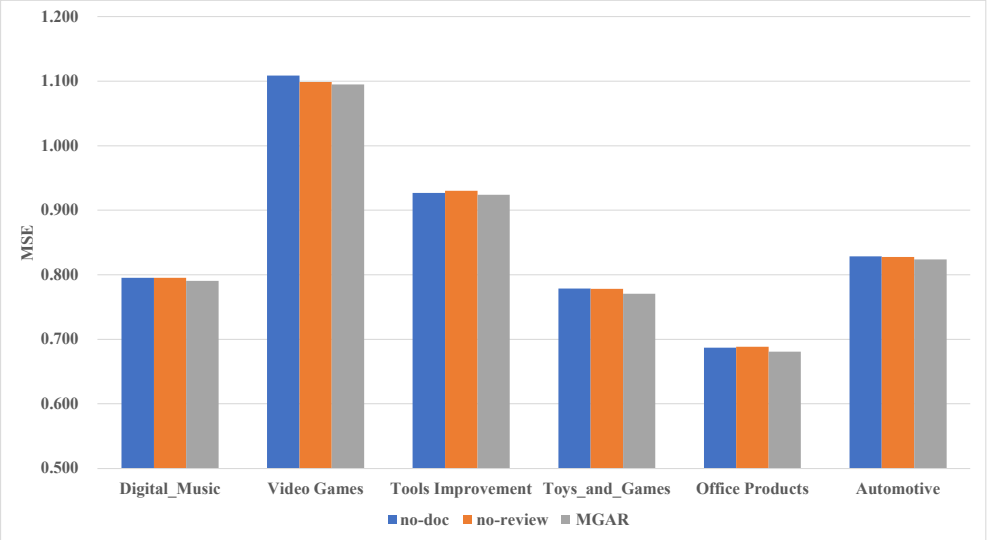

Fig. 4 Comparisons between MGAR and its variants.

To ensure the reliability of our experimental results, each test was conducted ten times, and the outcomes were averaged. The top-performing model's results are accentuated in bold, while the second-best are underlined for clarity. The MGAR model consistently outperformed the baseline models across all six datasets by a margin of over 2.0%, affirming its robust capability in user and item representation. Notably, MGAR exceeded the performance of the second-best baseline by a substantial margin of over 5.0% on the Digital Music and Office Products datasets, highlighting its exceptional effectiveness in these specific contexts.

5.6 Ablation Study

In this section, we perform ablation experiments to discover the impact of each component in the model.

Impact of Multi-Granularity Review Text Feature Fusion: To evaluate the benefits of integrating features from reviews at different granularities, we constructed two model variants: 'no-doc', which excludes the document encoder, and 'no-review', which omits the text encoder. The experimental outcomes, depicted in Fig. 4, reveal that both variants underperform compared to the full MGAR model on all datasets, as evidenced by higher Mean Squared Error (MSE) values. This degradation in performance upon the removal of either encoder underscores the significance of each feature set. The results demonstrate that the fusion of document-level and text-level review information is crucial for achieving optimal accuracy in rating prediction.

Evaluating the Word Local Attention Module: To investigate the contribution of the word local attention mechanism in the MGAR model, we developed a variant, MGAR-NG, which does not incorporate the attention module for processing

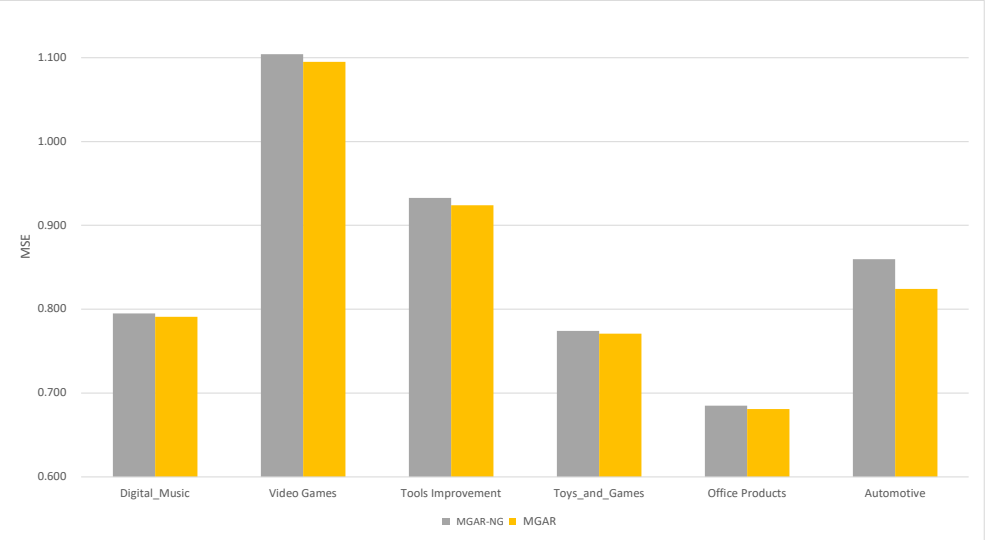

Fig. 5 Comparisons between MGAR and its variants.

long word sequences. In this variant, all words are considered to have equal significance. The performance of MGAR-NG, as shown in Fig. 5, is diminished across all six Amazon datasets when compared to the original MGAR model. This decline in performance suggests that treating every word in a review with equal weight is suboptimal. Conversely, MGAR's ability to discern and emphasize the most informative words within a review document is validated, highlighting the importance of the attention mechanism in enhancing the model's predictive accuracy.

Assessing the Asymmetric Review Attention Mechanism: We introduced three distinct variants of the MGAR model—MGAR-NOSACA, MGAR-SS, and MGAR-SC—to examine the efficacy of the asymmetric attention mechanism in handling review data. MGAR-NOSACA eliminates both the self-attention and cross-attention mechanisms, thereby disregarding the heterogeneity of users' reviews as well as the homogeneity of items' reviews. By doing so, it serves as a baseline to understand the importance of differentiating between review types. To affirm the heterogeneous nature of users' reviews, the MGAR-SS variant employs self-attention mechanisms for both user and item review features independently, replacing the cross-attention mechanism. This setup allows us to investigate the impact of treating user and item reviews as distinct entities that do not interact. MGAR-SC inverses the application of the attention mechanisms used in the original model. It applies the self-attention mechanism to user reviews and the cross-attention mechanism to item reviews, thereby assuming homogeneity within user reviews and heterogeneity within item reviews. This configuration tests the hypothesis that different types of reviews contribute asymmetrically to the recommendation process.

The experimental results are shown in Fig.6, where all three model variants perform worse than the MGAR on the six datasets. These findings substantiate our proposed

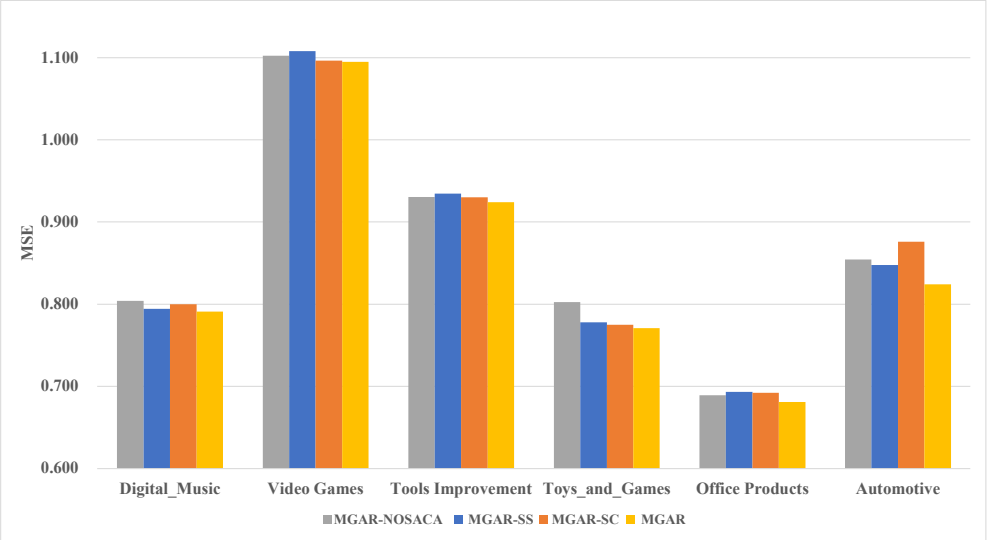

Fig. 6 Comparisons between MGAR and its variants.

hypothesis that user reviews and item reviews exhibit distinct characteristics, necessitating the application of disparate attention mechanisms. The inferior results of using two structurally identical attention modules, as seen in the MGAR-SS variant, further reinforce the argument for the necessity of asymmetric attention mechanisms to accurately capture the nuanced differences between user and item review dynamics.

6 Conclusion and Future Work

In this paper, we propose a unified neural recommendation method MGAR by fusing the document-level and text-level features of users and items, which captures more comprehensive representations for them. Both two types of features are learned under two different attention mechanisms: the self-attention mechanism and the asymmetric cross-attention mechanism. The core of our asymmetric cross-attention is using self-attention to learn the importance of the review texts for the same item and use the learned feature to guide the importance of the review texts for the same user by the cross-attention mechanism. And the ID information is also fused to enhance the performance. Experiment results on six real-world datasets from Amazon validate that our approach can effectively achieve rating prediction.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 71971002), the Major Project of Scientific Research in Higher Education Institutions in Anhui Province (Grant No. 2024AH040011), and the Anhui Postdoctoral Scientific Research Program Foundation (Grant No. 2024B828).

References

[1] Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook, 1–35 (2010)

[2] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295 (2001)

[3] Cheng, Z., Ding, Y., Zhu, L., Kankanhalli, M.: Aspect-aware latent factor model: Rating prediction with ratings and reviews. In: Proceedings of the 2018 World Wide Web Conference, pp. 639–648 (2018)

[4] Chin, J.Y., Zhao, K., Joty, S., Cong, G.: Anr: Aspect-based neural recommender. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 147–156 (2018)

[5] Jakob, N., Weber, S.H., Müller, M.C., Gurevych, I.: Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In: Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, pp. 57–64 (2009)

[6] Liu, H., Wang, W., Xu, H., Peng, Q., Jiao, P.: Neural unified review recommendation with cross attention. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1789–1792 (2020)

[7] Zheng, L., Noroozi, V., Yu, P.S.: 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, pp. 425–434 (2017)

[8] Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., Kankanhalli, M.S.: Aˆ 3ncf: An adaptive aspect attention model for rating prediction. In: Proceedings of IJCAI, pp. 3748–3754 (2018)

[9] Liu, D., Li, J., Du, B., Chang, J., Gao, R.: Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 344–352 (2019)

[10] Seo, S., Huang, J., Yang, H., Liu, Y.: Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 297–305 (2017)

[11] Chen, C., Zhang, M., Liu, Y., Ma, S.: Neural attentional rating regression with review-level explanations. In: Proceedings of the 2018 World Wide Web Conference, pp. 1583–1592 (2018)

[12] Tay, Y., Luu, A.T., Hui, S.C.: Multi-pointer co-attention networks for recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2309–2318 (2018)

[13] Liu, H., Wu, F., Wang, W., Wang, X., Jiao, P., Wu, C., Xie, X.: Nrpa: neural recommendation with personalized attention. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1233–1236 (2019)

[14] Wu, C., Wu, F., Liu, J., Huang, Y.: Hierarchical user and item representation with three-tier attention for recommendation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)

[15] Chen, Z., Wang, X., Xie, X., Wu, T., Bu, G., Wang, Y., Chen, E.: Co-attentive multi-task learning for explainable recommendation. In: Proceedings of IJCAI, pp. 2137–2143 (2019)

[16] Dezfouli, P.A.B., Momtazi, S., Dehghan, M.: Deep neural review text interaction for recommendation systems. Applied Soft Computing **100**, 106985 (2021)

[17] Shuai, J., Zhang, K., Wu, L., Sun, P., Hong, R., Wang, M., Li, Y.: A review-aware graph contrastive learning framework for recommendation, 1283–1293 (2022)

[18] Cai, X., Huang, C., Xia, L., Ren, X.: Lightgcl: Simple yet effective graph contrastive learning for recommendation. In: Proceedings of International Conference on Learning Representation (2023)

[19] Zhang, X., Ma, H., Yang, F., Li, Z., Chang, L.: Kgcl: A knowledge-enhanced graph contrastive learning framework for session-based recommendation. Engineering Applications of Artificial Intelligence **124** (2023)

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[21] Dong, X., Ni, J., Cheng, W., Chen, Z., Zong, B., Song, D., Liu, Y., Chen, H., De Melo, G.: Asymmetrical hierarchical networks with attentive interactions for interpretable review-based recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7667–7674 (2020)

[22] Sun, L., Liu, B., Tao, J., Lian, Z.: Multimodal cross-and self-attention network for speech emotion recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4275–4279 (2021)

[23] Catherine, R., Cohen, W.: Transnets: Learning to transform for recommendation.

In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 288–296 (2017)

[24] Wu, L., Quan, C., Li, C., Wang, Q., Zheng, B., Luo, X.: A context-aware user-item representation learning for item recommendation. ACM Transactions on Information Systems (TOIS) **37**(2), 1–29 (2019)

[25] Hyun, D., Park, C., Yang, M.-C., Song, I., Lee, J.-T., Yu, H.: Review sentiment-guided scalable deep recommender system. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 965–968 (2018)

[26] Liu, H., Zhao, J., Li, P., Zhao, P., Wu, X.: Shared-view and specific-view information extraction for recommendation. Expert Systems with Applications **186**, 115752 (2021)

[27] Yang, P., Xiao, Y., Zheng, W., Jiao, X., Zhu, K., Sun, C., Liu, L.: Man: Main-auxiliary network with attentive interactions for review-based recommendation. Applied Intelligence, 1–16 (2022)

[28] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Proceedings of ICLR (Workshop Poster) (2013)

[29] Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)

[30] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv preprint arXiv:1905.09418 (2019)

[31] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[32] Da'u, A., Salim, N.: Recommendation system based on deep learning methods: a systematic review and new directions. Artificial Intelligence Review **53**(4), 2709–2748 (2020)