Information Processing and Management 60 (2023) 103166

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Analyzing review sentiments and product images by parallel deep nets for personalized recommendation

Zhu Zhan [a,b], Bugao Xu [b,c,*]

[a] College of Textiles, Donghua University, Shanghai, China
[b] Department of Merchandising and Digital Retailing, University of North Texas, Denton, United States
[c] Department of Computer Science and Engineering, University of North Texas, Denton, United States

ARTICLE INFO

*Keywords:*
Recommendation
Rating inference
Semantic analysis
Visual analysis
Aspect-sentiment
Scale-aware attention

ABSTRACT

Recommendation is an effective marketing tool widely used in the e-commerce business, and can be made based on ratings predicted from the rating data of purchased items. To improve the accuracy of rating prediction, user reviews or product images have been used separately as side information to learn the latent features of users (items). In this study, we developed a hybrid approach to analyze both user sentiments from review texts and user preferences from item images to make item recommendations more personalized for users. The hybrid model consists of two parallel modules to perform a procedure named the multiscale semantic and visual analyses (MSVA). The first module is designated to conduct semantic analysis on review documents in various aspects with word-aware and scale-aware attention mechanisms, while the second module is assigned to extract visual features with block-aware and visual-aware attention mechanisms. The MSVA model was trained, validated and tested using *Amazon Product Data* containing sampled reviews varying from 492,970 to 1 million records across 22 different domains. Three state-of-the-art recommendation models were used as the baselines for performance comparisons. Averagely, MSVA reduced the mean squared error (MSE) of predicted ratings by 6.00%, 3.14% and 3.25% as opposed to the three baselines. It was demonstrated that combining semantic and visual analyses enhanced MSVA's performance across a wide variety of products, and the multiscale scheme used in both the review and visual modules of MSVA made significant contributions to the rating prediction.

## 1. Introduction

With the rapid rise of the digital economy, personalized recommendation has become a widely used tool for e-commerce providers (e.g., *Amazon, Yelp*) to discover potential customers and improve user loyalty. In most scenarios, the ratings on purchased items can intuitively reflect users' satisfaction with the shopping experience and the products, and therefore can be used to train a model to deliver personalized recommendations. The overall process of a recommendation system is first to learn the unique representations of users and items from existing data (e.g., reviews, ratings), to estimate a user's rating for unrated items by calculating the similarity between user-item pairs, and finally to recommend items that have high predicted ratings to specific users. For example, in collaborative filtering, a rating matrix is constructed with user indices, item indices and corresponding ratings to show interactions between

* Corresponding author at: 1155 Union Circle #311100, Denton, TX 76203-5017, United States.
*E-mail address:* bugao.xu@unt.edu (B. Xu).

https://doi.org/10.1016/j.ipm.2022.103166
Received 26 May 2022; Received in revised form 13 October 2022; Accepted 8 November 2022
Available online 16 November 2022

users and items. The ratings of non-interacted cells (i.e., null data) in the matrix are then estimated with a well-known algorithm, Matrix Factorization (MF). Finally, the item recommendations are made in the order of the estimated ratings to targeted users. However, MF may suffer from low interpretability and cold start problems due to insufficient rating data in the matrix (Koren et al., 2009).

To tackle these limitations, recent studies attempted to incorporate various forms of side information, such as semantic information of user reviews and visual features of product images, into the MF model. Unlike rating numbers, which provide overall impressions on purchased items, user reviews contain rich sentiment information expressing more fine-grained opinions on products in perspectives. As a semantic review method, the aspect-sentiment approach with attention mechanism was widely adopted in such models as *Adaptive Aspect attention-based Neural Collaborative Filtering* ($A^3$NCF) (Cheng et al., 2018) and *Aspect-based Neural Recommender* (ANR) (Chin et al., 2018). Deep neural recommenders, e.g., *Deep Cooperative Neural Networks* (DeepCoNN) (L. Zheng et al., 2017) and *Dynamic Review-based Recommenders* (DRR) (Cvejoski et al., 2021), demonstrated more powerful representation abilities for extracting aspect-sentiment features.

On the other hand, user preferences can be reflected by the visual appearance of purchased products. Thus, many recommendation models, such as *Visual Bayesian Personalized Ranking* (VBPR) (He and McAuley, 2016b) and *P*ersonalized *C*ompatibility *M*odeling (GP-BPR) (Song et al., 2019), were developed to utilize visual features to improve recommendation performance. Tang et al. (2020) introduced an adversarial training procedure to the recommender to enhance its robustness. Anelli et al. (2021) designed a visual adversarial recommender (VAR) with the adversarial training strategy to improve rating predictions. But these models separately learned the latent factors of users (without items) and items (without users) from user-item interactions.

The aforementioned studies do not take advantage of side information from both the semantic analysis of user reviews and the visual analysis of item images in the rating prediction because semantic and visual features are in different dimensions that complicate the end-to-end joint training. To improve the relevance of recommendations that reflect purchase experience and personal preferences, in this paper, we present a model that can perform multiscale semantic and visual analyses (MSVA) with attention mechanisms on both user reviews and item images for more precise rating predictions. The MSVA model extracts the semantic and visual representations of user-item pairs in two parallel paths (or modules) with the inputs of review texts and product images, and then combines the two representations to infer a rating for a personalized recommendation. The datasets used to train, validate and test MSVA are the *Amazon Product Data* in 22 different domains ranging from automotive, phone, …, to toy (Amazon, 2021). The key contributions of this study can be summarized as follows:

- An end-to-end neural recommender (MSVA) is constructed to model latent factors of users and items with review texts and product images jointly to reinforce the predicted ratings that are more relevant to users' previous experience and personal preference.
- Semantic and visual feature representations are extracted at multiple scales with attention mechanisms so that fine-grained implicit sentiments can be mined.
- The MSVA model is trained and evaluated with the real-world datasets, *Amazon Product Data* in 22 various domains, and against three state-of-the-art baselines, ANR (Chin et al., 2018), DRR (Cvejoski et al., 2021) and VAR (Anelli et al., 2021). MSVA outperforms the baseline models.
- The impacts of MSVA's unique hyperparameters and variants on the rating prediction are analyzed to optimize the model.

## 2. Related Work

In recent years, many recommendation methods have been developed based on either semantic reviews or product images. Four key techniques related to our work are briefly reviewed below.

### 2.1. Review-based Recommendation

A review-based recommender extracts sentiment information about user experience and item characteristics from review comments. It is often trained with reviews and corresponding ratings to improve performance and interpretability (Chambua and Niu, 2021; Liu et al., 2020; Xie et al., 2021). For example, $A^3$NCF (Cheng et al., 2018) utilized topic models to extract more fine-grained information with aspect-sentiment analysis. SULM (Sentiment Utility Logistic Model) (Bauman et al., 2017) and ANR (Chin et al., 2018) extracted aspect-based features with an estimated importance for each aspect. However, some of these methods relied on external sentiment analysis tools and therefore are not self-contained. Recently, deep learning was introduced to review analysis with tremendous success. For instance, Zheng et al. (2017) and Catherine and Cohen (2017) generated feature maps from a review via multiple channels by using convolutional operations of two deep nets, DeepCoNN and TransNet, respectively, and highlighted the key features with a max-pooling layer. Since reviews at different times reveal possible changes in a user's sentiment, Cvejoski et al. (2021) implemented a dynamic review-based recommender (DRR) with two recurrent neural networks (RNNs) to capture the evolution of user and item preferences. In this paper, we will extract users' fine-grained sentiments from reviews in various aspects and on multiple scales.

### 2.2. Visual-based Recommendation

CNN is a deep net frequently used to extract high-dimensional visual features for product recommendation (Anelli et al., 2021). He et al. (2016) utilized the visual appearance of items and the feedback of users to build a recommender to endorse clothing and

accessory items. In their later work, visual-aware personalized ranking was introduced to transform an MF model into the VBPR model which significantly enhanced recommendation performance (He and McAuley, 2016b). Song et al. (2019) fused general compatibility factors and personal preference factors in clothing images to match clothes using GP-BPR. Recent developments in visual-based recommendation have been largely focused on exploring adversarial machine learning to improve recommendation accuracy (He et al., 2018). Tang et al. (2020) proposed an adversarial multimedia recommender to enhance the efficiency of adversarial learning by mitigating perturbations in product images. Anelli et al. (2021) studied the defensive side of a visual-based recommender against adversarial perturbation of images. In this paper, we will try the transfer learning to extract visual representations from product images with a pre-trained CNN called ResNet-152 (He et al., 2016).

### 2.3. Multiscale Scheme

Extracting multiscale features is one of the common schemes used in the fields of natural language processing (NLP) and image processing because a multiscale scheme can enrich the granularity of information and capture the recurrent relations between different local features. He et al. (2019) utilized local, extractive, and multiscale convolution-based features as recurrent gating functions for item recommendation. Song et al. (2019) presented a memory model with multiscale features for session-based recommendations. More recently, Zhang et al. (2022) proposed an attention-based frequency-aware multiscale network (AFMN) for sequential recommendations. In light of this, we will apply the multiscale scheme simultaneously to the semantic and visual analyses on review texts and product images by changing the size of the convolutional kernels.

### 2.4. Attention Mechanism

Attention mechanism is another major technique used in NLP for performing self-learning with the *softmax* function and assigning adaptive weights to features (Zheng et al., 2020). The weights reflect the importance levels of corresponding features. Guan et al. (2019) captured user attentions in specific aspects to identify personalized interests among features for explainable recommendations. Li and Xu (2020) detected both local and global aspect features simultaneously with attention layers in two parallel paths. More recently, Liu et al. (2021) devised a three-tier attention framework to capture hierarchical representations of reviews, aspects, and users/items, respectively. Based on these methods, we will construct a word-aware attention layer in MSVA to seek aspect and sentiment words of different importance weights in semantic reviews, and a scale-aware attention layer to select informative *n*-grams from review texts and item images, respectively.

## 3. Multiscale Semantic-Visual Representations

Multiscale semantic-visual representations refer to a set of representative features in user reviews and item images to be extracted through natural-language-processing (semantic analysis) and image-processing (visual analysis) techniques for rating predictions.

### 3.1. Architecture and Intuition

Let $\mathscr{D}$ be a textual review corpus from a set of users ($\mathscr{U}$) for a set of items ($\mathscr{I}$), and $\mathscr{P}$ be a collection of images (or pictures) of items in $\mathscr{I}$ retrieved from the metadata of an e-commerce website. Each user-item interaction is denoted as a tuple, $(u,i,d_{u,i},p_i,r_{u,i})$, where $u$

**Table 1**
Notations and Definitions.

| Notation | Definition |
|---|---|
| $(u, i, d_{u,i}, p_i, r_{u,i})$ | User-item interaction tuple |
| $r_{u,i}$ | Rating for user $u$ on item $i$ |
| $\mathbf{D}_u$ | User document, i.e., a collection of all reviews from user $u$ |
| $\mathbf{D}_i$ | Item document, i.e., a collection of all reviews for item $i$ |
| $\mathscr{A}$ | Set of aspects |
| $\mathscr{K}$ | Set of kernel sizes |
| $\mathbf{W}_a$, $\mathbf{b}_a$ | Word transformation matrix and bias for aspect $a \in \mathscr{A}$ |
| $\mathbf{w}_{a,k}^{word}$ | Word-aware embedding vector for aspect $a \in \mathscr{A}$ at kernel size of $k \in \mathscr{K}$ |
| $\mathbf{w}_a^{scale}$ | Scale-aware embedding vector for aspect $a \in \mathscr{A}$ |
| $\mathbf{p}_{u,a,k}$ | Implicit review vector of user $u$ for aspect $a \in \mathscr{A}$ at kernel size of $k \in \mathscr{K}$ |
| $\mathbf{d}_{u,a}$ | Review representation of user $u$ for aspect $a \in \mathscr{A}$ |
| $\mathbf{P}_u$ | Product image set from user $u$ |
| $\mathbf{P}_i$ | Product image set for item $i$ |
| $\mathscr{B}$ | Set of block size |
| $\mathbf{W}_b$ | Visual transformation matrix for block size $b \in \mathscr{B}$ |
| $\mathbf{w}_b^{block}$ | Block-aware embedding vector for block size $b \in \mathscr{B}$ |
| $\mathbf{w}^{visual}$ | Visual embedding vector |
| $\mathbf{q}_{u,b}$ | Implicit visual vector of user $u$ for block size $b \in \mathscr{B}$ |
| $\mathbf{v}_u$ | Visual representation of user $u$ |

and $i$ are the indices of user and item, $d_{u,i} \in \mathscr{D}$ is a review text from user $u$ on item $i$, $p_i \in \mathscr{P}$ is the image set for item $i$, and $r_{u,i}$ is a rating on the overall satisfaction of user $u$ towards item $i$. The ultimate task of the MSVA model is to predict a rating, $\hat{r}_{u,i}$, for any non-interacted user-item pair, i.e., a case where user $u$ has not reviewed or rated item $i$. In this paper, we denote scalars with italic lowercases (e.g., $x$), vectors with bold lowercases (e.g., $\mathbf{x}$), and matrices or high dimensional tensors with bold uppercases (e.g., $\mathbf{P}$). Besides, we use the Python-like array indexing, e.g., $\mathbf{x}[i]$ to denote the $i$-th element of vector $\mathbf{x}$, and $\mathbf{W}[i,:]$ to denote all the elements on the $i$-th row of matrix $\mathbf{W}$. To facilitate the following discussion, all the key notations are listed in Table 1.

The architecture of our proposed MSVA model is illustrated in Fig. 1. The model is composed of two modules, namely the review and visual representation modules. In the review representation module, we feed two inputs: (1) the user document, $\mathbf{D}_u$, which is a set of tokenized reviews written by user $u$ for all the items she/he reviewed, and (2) the item document, $\mathbf{D}_i$, which is a set of tokenized reviews written for item $i$ by all users who reviewed it. Similarly, we feed the product image set of all items reviewed by user $u$, $\mathbf{P}_u$, and the product image set of item $i$ reviewed by all users, $\mathbf{P}_i$, to the visual representation module. Within each module, the network structures for processing $\mathbf{D}_u$ and $\mathbf{D}_i$ (or $\mathbf{P}_u$ and $\mathbf{P}_i$), are completely consistent as indicated in Fig. 1. Thus, we will only elaborate the processes for $\mathbf{D}_u$ and $\mathbf{P}_u$ of user $u$, and forgo those for $\mathbf{D}_i$ and $\mathbf{P}_i$ of item $i$.

Before elaborating on the modules, we introduce a few intuitions that provide basic guidance for originating some important components or operations in MSVA, some of which were explained in previous studies (Chin et al., 2018). Examples of actual review sentences used to illustrate these intuitions can be found in Fig. 2, where a blue color represents noun (or aspect) words or phrases and a red color represents adjective (or sentiment) words.

**Intuition 1**: The same word may have various polarities in meaning when used in different contexts (Da'u et al., 2020; Li and Xu, 2020). For example, the word '*fast*' expresses a positive sentiment in a sentence like "*the express delivery is very fast.*" On the other hand, in the sentence "*the battery power drops fast in use,*" the same word carries a negative sentiment.
**Intuition 2**: Contextual information is crucial in the field of NLP (Zheng et al., 2019). Aspect-related words (e.g., storage, price, size, color) and sentiment-related words (e.g., sufficient, reasonable, available, favorite) often appear in relatively close positions in a sentence. For example, "*this is a sufficient storage*" and "*the storage is sufficient.*" Processing a word in a review text should consider not only the word itself but also its context, i.e., the words around it.
**Intuition 3**: A user may comment on a target item by using different sentiment words to describe different aspects. For example, "*It's a great phone with a fantastic price but the storage is a bit small.*" Word '*fantastic*' should have more attention on the aspect of '*price*,' while the word '*small*' should have more weight on the aspect of '*storage*.' Thus, different weights should be assigned to different words with respect to specific aspects. This is so-called the word attention mechanism.

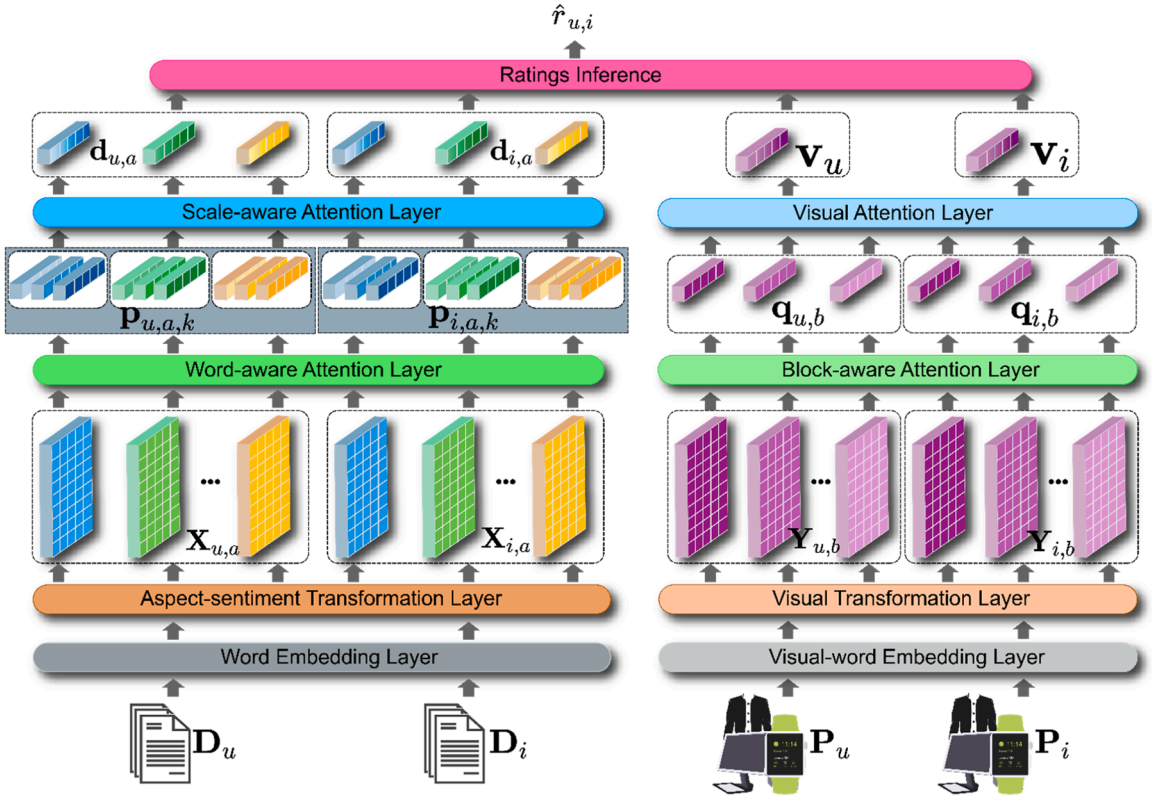


**Fig. 1.** Architecture of Multiscale Semantic-Visual Analysis (MSVA) Model.

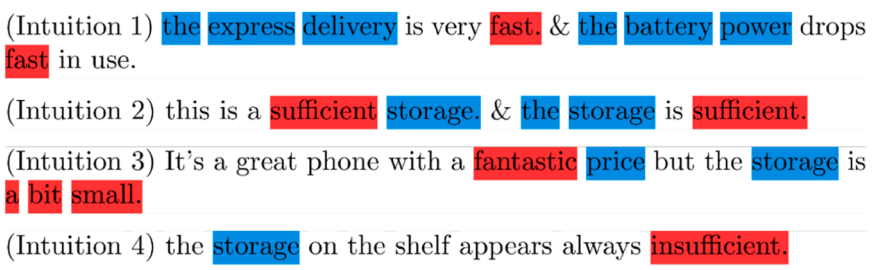


**Fig. 2.** Sentence examples with color highlights on aspect (blue) and sentiment (red) words for intuitions used in MSVA.

**Intuition 4**: As stated in Intuition 2, aspect words and sentiment words in a sentence are usually close to each other, but exceptions are not uncommon. For example, in a review sentence, "*the storage on the shelf appears always insufficient,*" the sentiment word, '*insufficient,*' is not adjacent to the aspect word, '*storage.*' To address such an issue, latent features should be extracted at multiple scales when processing a review document, and a scale-aware attention mechanism should be performed to weigh the multiscale features for selecting the most informative features.

### 3.2. Review Representation Module

#### 3.2.1. Word Embedding Layer

As an input, user review document $\mathbf{D}_u$ is fed to an embedding layer $f: \mathscr{V} \rightarrow \mathbb{R}^{d_w}$. The task of the embedding layer can be simply regarded as a lookup function that finds a word within a vocabulary set, $\mathscr{V}$, and then assigns it with a high-dimension (-dim) vector. To this end, the output of the embedding layer is a matrix, $\mathbf{X}_u \in \mathbb{R}^{m \times d_w}$, where $m$ is the number of words in $\mathbf{D}_u$ and $d_w$ is the dimension of $\mathscr{V}$. The embedding weights are initialized by word vectors generated by a pre-trained model, *word2vec* (Demeester et al., 2016). Compared with the traditional bag-of-words method, this word embedding layer has the advantage of preserving the contextual information of words.

#### 3.2.2. Aspect-Sentiment Transformation Layer

Inspired by **Intuition 1**, an aspect-sentiment transformation layer is used to capture associated sentimental polarities with different aspects in a review document. A previous study (Li and Xu, 2020) has shown that an aspect-based recommendation system can detect fine-grained implicit information behind a text.

Since each row vector in matrix $\mathbf{X}_u$ represents one word and all words share the $d_w$-dim space across aspects, a specific aspect-sentiment transformation matrix, $\mathbf{W}_a \in \mathbb{R}^{d_w \times h_1}$ and a specific bias vector $\mathbf{b}_a \in \mathbb{R}^m$, can be defined to indicate the fine-grained feedback for a given aspect $a \in \mathscr{A}$, where $h_1$ is the hidden dimension of aspect set $\mathscr{A} \in \mathbb{R}^{h1}$. In order to ensure the transformation to be data-driven without any external supervision, $\mathbf{W}_a$ is initialized randomly with a uniform distribution $\mathscr{U}(-0.01,\ 0.01)$, while $\mathbf{b}_a$ is initialized simply with zero. $|\mathscr{A}|$ is referred to as the number of aspects. We have:

$$\mathbf{X}_{u,a} = f(\mathbf{X}_u \mathbf{W}_a + \mathbf{b}_a),\ \forall a \in \mathscr{A} \tag{1}$$

here, $f(\cdot)$ denotes a nonlinear activation function (e.g., the ReLU function), and $\mathbf{X}_{u,a} \in \mathbb{R}^{m \times h_1}$ is a specific aspect-sentiment transformation of $\mathbf{X}_u$. Through supervised training with the ratings ($r_{u,i}$), $\mathbf{W}_a$ can be subsequently optimized with weights related to the importance levels of all the words. Note that $\mathbf{W}_a$ is shared by words in both $\mathbf{D}_u$ and $\mathbf{D}_i$ for the purpose to reduce the parameters that need to be learned. As a result, the output of this layer is a three-dimensional tensor in $\mathbb{R}^{|\mathscr{A}| \times m \times h_1}$ obtained by concatenating all $|\mathscr{A}|$ aspect-sentiment transformations.

#### 3.2.3. Word-Aware Attention Layer

This layer takes the output of the upstream layer (the aspect-sentiment transformation), $\mathbf{X}_{u,a}$, as the input. Based on **Intuition 2** and **Intuition 4**, a convolutional-like operation is performed to extract $n$-gram features from word sequences for mining multiscale context information (Wang et al., 2018). For each $\mathbf{X}_{u,a}$, variable $n$-gram features are produced with multiple kernel sizes, and an attention weight is calculated for each word, i.e., each row of the matrix (**Intuition 3**). Specifically, we estimate the attention weight of the $i$th word with a variable kernel size $k \in \mathscr{K}$ based on the word itself as well as the $(k-1)/2$ words before and after it if $k$ is odd, or the $k/2$ words before and after it if $k$ is even. The zero padding is applied to both ends of the input matrix to ensure the resulting implicit features have the same size among $|\mathscr{K}|$ kernels. In more detail, we pad $0 \in \mathbb{R}^{1 \times h_1}$ after $\mathbf{X}_{u,a}[m,:]$ when $k$ equals 2, pad $\mathbf{0}$ before $\mathbf{X}_{u,a}[1,:]$ and after $\mathbf{X}_{u,a}[m,:]$ when $k$ equals 3, and so on. Then, we concatenate the $k$ rows of matrix $\mathbf{X}_{u,a}$ to form a long row vector $\mathbf{x}_{u,a,k}[i] \in \mathbb{R}^{1 \times (k \times h_1)}$ ($i$=1, …, $k \times h_1$) for an inner product operation with a word-aware embedding vector, $\mathbf{w}_{a,k}^{word} \in \mathbb{R}^{(k \times h_1)}$, as shown in Eq. (2):

$$\mathbf{x}_{u,a,k}[i] = \mathbf{X}_{u,a}\left[i - \frac{k-1}{2},:\right] \oplus \cdots \oplus \mathbf{X}_{u,a}[i,:] \oplus \cdots \oplus \mathbf{X}_{u,a}\left[i + \frac{k-1}{2},:\right],\ \forall k \in \mathscr{K}, a \in \mathscr{A} \tag{2}$$

where $\oplus$ is the concatenating operation. $\mathbf{w}_{a,k}^{word}$ is initialized randomly with a uniform distribution $\mathscr{U}(-0.01,\ 0.01)$ before the training. Moreover, the word attention weight of the $i$th word, $\mathbf{attn}_{u,a,k}^{word}[i]$, is calculated with the *softmax* function:

$$\mathbf{attn}_{u,a,k}^{word}[i] = softmax\left(\mathbf{x}_{u,a,k}[i]\mathbf{w}_{a,k}^{word}\right),\ \forall k \in \mathscr{K}, a \in \mathscr{A} \tag{3}$$

where $softmax(\omega_i) = exp(\omega_i)/\sum_i exp(\omega_i)$.

In terms of semantic review with word-aware attention, the implicit vector $\mathbf{p}_{u,a,k} \in \mathbb{R}^{1\times h_1}$ can be derived from the following weighted sum:

$$\mathbf{p}_{u,a,k} = \sum_{i=1}^{m}\left(\mathbf{attn}_{u,a,k}^{word}[i]\mathbf{X}_{u,a}[i,:]\right),\ \forall k \in \mathscr{K}, a \in \mathscr{A} \tag{4}$$

### 3.2.4. Scale-Aware Attention Layer

This layer is designed for re-weighting the multiscale implicit feature vector $\mathbf{p}_{u,a,k}, k \in \mathscr{K}$ based on **Intuition 4**. For each aspect $a \in \mathscr{A}$, we concatenate $|\mathscr{K}|$ row vectors $\mathbf{p}_{u,a,k}$ in the column direction to form matrix $\mathbf{P}_{u,a} \in \mathbb{R}^{|\mathscr{K}|\times h_1}$ as follows:

$$\mathbf{P}_{u,a} = \left(\mathbf{p}_{u,a,k_1}^T \oplus \mathbf{p}_{u,a,k_2}^T \oplus \cdots \oplus \mathbf{p}_{u,a,k_{|\mathscr{K}|}}^T\right)^T,\ \forall a \in \mathscr{A} \tag{5}$$

Similar to the word-aware attention weights, the scale-aware attention weights, $\mathbf{attn}_{u,a}^{scale} \in \mathbb{R}^{|\mathscr{K}|\times 1}$, can be calculated with the *softmax* function on the inner product of $\mathbf{P}_{u,a}$ and a scale-aware embedding vector, $\mathbf{w}_a^{scale} \in \mathbb{R}^{h_1}$, as follows:

$$\mathbf{attn}_{u,a}^{scale} = softmax\left(\mathbf{P}_{u,a}\mathbf{w}_a^{scale}\right),\ \forall a \in \mathscr{A} \tag{6}$$

$\mathbf{w}_a^{scale}$ is randomly initialized as done for $\mathbf{w}_{a,k}^{word}$ above. Subsequently, the semantic review representation of aspect $a$, $\mathbf{d}_{u,a} \in \mathbb{R}^{1\times h_1}$, which is a $h_1$-long row vector, can be inferred with the following formula:

$$\mathbf{d}_{u,a} = \sum_{k=1}^{|\mathscr{K}|}\left(\mathbf{attn}_{u,a}^{scale}[k]\mathbf{p}_{u,a,k}\right),\ \forall a \in \mathscr{A} \tag{7}$$

where $\mathbf{attn}_{u,a}^{scale}[k]$ denotes the scale-aware attention weight of $\mathbf{p}_{u,a,k}$ at the scale size of $k$.

## 3.3. Visual Representation Module

### 3.3.1. Visual-Word Embedding Layer

Similar to a sentence that is composed of words, an image can be regarded as a sequence of blocks or so-called visual words. In the image-processing pipeline of the visual representation module in Fig. 1, the visual-word embedding layer performs five tasks: (1) dividing each product image from $\mathscr{P}$ into a number of blocks with a fixed block size of $b \in \mathscr{B}$ (height $\times$ width); (2) feeding the blocks into a pre-trained network, ResNet-152, which is a 152-layer neural network to perform an identity mapping to extract 2048-dim visual features (He et al., 2016); (3) building a visual vocabulary with the block size ($b$) through a clustering algorithm (e.g., K-means clustering) so that each visual word corresponds to a cluster center; (4) dividing each image in the product image set of a user, $\mathbf{P}_u$, into blocks with the same size of $b$ so that the image is treated as a sequence of concatenated blocks; and (5) replacing the block features in the embedding with their corresponding nearest cluster center among visual vocabularies. To this end, the embedding output for the block size of $b$ is $\mathbf{V}_{u,b} \in \mathbb{R}^{n\times d_v}$, where $n$ is the number of the blocks.

### 3.3.2. Visual Transformation Layer

Originally, the dimension of a visual-word output used in MSVA is 2,048, which is a large value causing a high computational cost. Therefore, we introduce a transformation matrix, $\mathbf{W}_b \in \mathbb{R}^{d_v\times h_2}$, where $h_2$ is the hidden dimension of the visual representation, to reduce dimension as follows:

$$\mathbf{Y}_{u,b} = \mathbf{V}_{u,b}\mathbf{W}_b,\ \forall b \in \mathscr{B} \tag{8}$$

here, $\mathbf{Y}_{u,b} \in \mathbb{R}^{n\times h_2}$ is the visual transformation output.

### 3.3.3. Block-Aware Visual Attention Layer

In the review representation model, the contextual information is captured when learning the word-aware attention weights. However, it is difficult to assess the associations between adjacent blocks of a product image when concatenating the blocks into a visual-word sequence. Therefore, a block-aware attention weight, $\mathbf{attn}_{u,b}^{block}$, is learned only based on the block itself:

$$\mathbf{attn}_{u,b}^{block} = softmax\left(\mathbf{Y}_{u,b}[j,:]\mathbf{w}_b^{block}\right),\ \forall b \in \mathscr{B} \tag{9}$$

$$\mathbf{q}_{u,b} = \sum_{j=1}^{n} \left( \mathbf{attn}_{u,b}^{block}[j] \mathbf{Y}_{u,b}[j,:] \right) \tag{10}$$

where $\mathbf{w}_b^{block} \in \mathbb{R}^{h_2}$ is a block-aware embedding vector that is initialized with a uniform distribution $\mathscr{U}(-0.01,\ 0.01)$, and $\mathbf{q}_{u,b} \in \mathbb{R}^{1\times h_2}$ is an implicit visual vector with respect to the block size of $b$.

### 3.3.4. Visual Attention Layer

In terms of visual similarity, multiscale information hidden at various block sizes are useful as well. Similar to the review representation module, a multiscale visual attention mechanism is applied to the implicit visual feature vector, $\mathbf{q}_{u,b} \in \mathbb{R}^{1\times h_2}$. This is done by learning a vector $\mathbf{w}^{visual} \in \mathbb{R}^{h_2}$ when calculating multiscale visual attention weights, $\mathbf{attn}_u^{visual}$, as follows:

$$\mathbf{Q}_u = \left( \mathbf{q}_{u,b_1}^T \oplus \mathbf{q}_{u,b_2}^T \oplus \cdots \oplus \mathbf{q}_{u,b_{|\mathscr{B}|}}^T \right)^T \tag{11}$$

$$\mathbf{attn}_u^{visual} = softmax\left( \mathbf{Q}_u \mathbf{w}^{visual} \right) \tag{12}$$

Then, the visual representation, $\mathbf{v}_u \in \mathbb{R}^{1\times h_2}$, can be derived with the following formula:

$$\mathbf{v}_u = \sum_{b=1}^{|\mathscr{B}|} \left( \mathbf{attn}_u^{visual}[b] \mathbf{q}_{u,b} \right) \tag{13}$$

## 3.4. Rating Inference

Given the review and visual representations for user $u$ and item $i$, the predicted rating, $\widehat{r}_{u,i}$, can be inferred as follows:

$$\widehat{r}_{u,i} = \sum_{a\in\mathscr{A}} \left( \mathbf{d}_{u,a} \left( \mathbf{d}_{i,a} \right)^T \right) + \mathbf{v}_u (\mathbf{v}_i)^T + b_u + b_i + b_0 \tag{14}$$

where $b_u, b_i$, and $b_0$ are the user, item, and global biases, respectively (Chin et al., 2018).

In order to optimize $\widehat{r}_{u,i}$ with the backpropagation and stochastic gradient descent (SGD) techniques (Bottou, 2012), the following loss function is designed based on the standard mean squared error (MSE) form:

$$\mathscr{L} = \frac{1}{|\mathscr{U}| \cdot |\mathscr{I}|} \sum_{u\in\mathscr{U}, i\in\mathscr{I}} + \frac{\lambda_\Theta}{2} \| \Theta \|_2^2 \tag{15}$$

$$\Theta = \left\{ \mathbf{W}_a,\ \mathbf{b}_a, \mathbf{w}_{a,k}^{word}, \mathbf{w}_a^{scale}, \mathbf{W}_b, \mathbf{w}_b^{block}, \mathbf{w}^{visual} \middle| a \in \mathscr{A}, k \in \mathscr{K}, b \in \mathscr{B} \right\} \tag{16}$$

**Table 2**
Numbers of users, items and ratings, and data sparsity of Amazon Product Data.

| Domain | *Users* | *Items* | *Ratings* | *Sparsity* (%) |
|---|---|---|---|---|
| Automotive | 668,449 | 265,988 | 1,000,000 | 99.9994 |
| Baby | 525,592 | 63,557 | 902,474 | 99.9973 |
| Beauty | 703,928 | 180,482 | 1,000,000 | 99.9992 |
| Books | 754,748 | 454,839 | 1,000,000 | 99.9997 |
| CDs andVinyl | 583,868 | 267,719 | 1,000,000 | 99.9994 |
| Cell Phones and Accessories | 835,716 | 168,211 | 1,000,000 | 99.9993 |
| Clothing, Shoes and Jewelry | 812,177 | 390,133 | 1,000,000 | 99.9997 |
| Digital Music | 470,808 | 261,826 | 820,822 | 99.9993 |
| Electronics | 834,983 | 189,367 | 1,000,000 | 99.9994 |
| Grocery and Gourmet Food | 629,302 | 148,114 | 1,000,000 | 99.9989 |
| Health and Personal Care | 774,368 | 155,982 | 1,000,000 | 99.9992 |
| Homeand Kitchen | 802,487 | 200,101 | 1,000,000 | 99.9994 |
| Kindle Store | 583,374 | 259,979 | 1,000,000 | 99.9993 |
| Movies and TV | 650,755 | 116,516 | 1,000,000 | 99.9987 |
| Musical Instruments | 335,081 | 82,137 | 492,970 | 99.9982 |
| Office Products | 758,539 | 117,002 | 1,000,000 | 99.9989 |
| Patio, Lawn and Garden | 700,037 | 104,250 | 969,757 | 99.9987 |
| Pet Supplies | 633,705 | 94,202 | 1,000,000 | 99.9983 |
| Sports and Outdoors | 779,542 | 252,866 | 1,000,000 | 99.9995 |
| Tools and Home Improvement | 725,234 | 188,107 | 1,000,000 | 99.9993 |
| Toys and Games | 718,115 | 221,737 | 1,000,000 | 99.9994 |
| Video Games | 643,985 | 47,166 | 1,000,000 | 99.9967 |
| **Average** | **678,400** | **192,286** | **963,001** | **99.9993** |

where $\|\cdot\|_2$ denotes the $l_2$ norm regularization for preventing model overfitting, and $\Theta$ is the set of model parameters to be updated. Besides, the dropout technique is adopted to improve generalization performance.

## 4. Experiment

To compare the performance of the proposed MSVA model with available state-of-the-art baselines, we implemented the algorithms in the PyTorch framework (Paszke et al., 2019) on a machine that has a GPU of NVIDIA Geforce GTX 1080Ti. All the experiments were performed on publicly available Amazon datasets.

### 4.1. Dataset

The datasets used to train the models are the *Amazon Product Data* (He and McAuley, 2016a), which contain over 142.8 million real-world reviews and 9.4 million metadata (e.g., product descriptions, prices, and image URLs) in 22 various domains for an 18-year period from 1996 to 2014 (Amazon, 2021). Originally, the datasets possessed 24 domains, but two of them, *Apps for Android* and *Amazon Instant Video*, did not provide image URLs in their metadata, and thus they were excluded from the experiment. The product images were automatically downloaded by a crawler with the feeds of the URLs in the metadata. The users who did not write reviews and the items that were not reviewed by any user or did not have any product image were screened out of each domain dataset. For a domain dataset whose user-item interactions exceed 1,000,000 after the screening, only 1,000,000 records were randomly sampled from the dataset. Each of these 22 datasets was divided into a training, validation and testing set in a ratio of 80:10:10. Table 2 shows the detailed information, such as the numbers of users, items and ratings, and the data sparsity of *Amazon Product Data* in the 22 domains. Here, $Sparsity = 1 - Ratings/(Users \times Items)$. The average numbers of ratings per user and per item are 1.44 and 6.65, respectively.

### 4.2. Baseline Model

Three baseline models, **ANR** (Chin et al., 2018), **DRR** (Cvejoski et al., 2021) and **VAR** (Anelli et al., 2021), were used to evaluate the performance of the MSVA model. The following is a brief introduction to the baselines and their hyperparameter settings.

(a) **ANR** is a recommender that utilizes both review texts and ratings to extract the fine-grained implicit feedback of users and items by aspect-aware analysis. Furthermore, ANR performs neural attention and co-attention, as well as aspect attention mechanisms for improving its performance. We set the parameters similar to those in (Chin et al., 2018), such as the number of aspects $|\mathscr{A}| = 5$, the context window size $c = 3$, latent factors size $h_1 = 10, h_2 = 50$, and the dropout rate $\rho = 0.5$. As for the word embedding, ANR uses 300-*dim* word vectors (*word2vec*) trained on Google News.

**Table 3**
MSE comparisons between MSVA and baselines (ANR, DRR and VAR).

| | Dataset | ANR | DRR | VAR | MSVA | Improvement (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) | (d) vs. (a) | (d) vs. (b) | (d) vs. (c) |
| 1 | *Automotive* | 1.249 | 1.187 | 1.175 | **1.163** | 6.89 | 2.02 | 1.02 |
| 2 | *Baby* | 1.315 | 1.306 | 1.286 | **1.241** | 5.63 | 4.98 | 3.50 |
| 3 | *Beauty* | 1.407 | 1.392 | 1.385 | **1.358** | 3.48 | 2.44 | 1.95 |
| 4 | *Books* | 0.967 | 0.955 | 0.966 | **0.952** | 1.55 | 0.31 | 1.45 |
| 5 | *CDs and Vinyl* | 0.954 | **0.946** | 0.972 | 0.959 | -0.52 | -1.37 | 1.34 |
| 6 | *Cell Phones and Accessories* | 1.847 | 1.637 | 1.654 | **1.608** | 12.94 | 1.77 | 2.78 |
| 7 | *Clothing, Shoes and Jewelry* | 1.316 | 1.267 | 1.256 | **1.208** | 8.21 | 4.66 | 3.82 |
| 8 | *Digital Music* | 0.695 | **0.693** | 1.698 | 0.696 | -0.14 | -0.43 | 0.29 |
| 9 | *Electronics* | 1.565 | 1.468 | 1.499 | **1.406** | 10.16 | 4.22 | 6.20 |
| 10 | *Grocery and Gourmet Food* | 1.235 | 1.239 | 1.248 | **1.196** | 3.16 | 3.47 | 4.17 |
| 11 | *Health and Personal Care* | 1.437 | 1.395 | 1.385 | **1.320** | 8.14 | 5.38 | 4.69 |
| 12 | *Home and Kitchen* | 1.435 | 1.374 | 1.369 | **1.296** | 9.69 | 5.68 | 5.33 |
| 13 | *Kindle Store* | 0.807 | 0.804 | 0.802 | **0.799** | 0.99 | 0.62 | 0.37 |
| 14 | *Movies and TV* | 1.152 | 1.141 | 1.144 | **1.132** | 1.74 | 0.79 | 1.05 |
| 15 | *Musical Instruments* | 1.075 | 1.062 | 1.071 | **1.023** | 4.84 | 3.67 | 4.48 |
| 16 | *Office Products* | 1.404 | 1.421 | 1.439 | **1.318** | 6.13 | 7.25 | 8.41 |
| 17 | *Patio, Lawn and Garden* | 1.456 | 1.406 | 1.406 | **1.383** | 5.01 | 1.64 | 1.64 |
| 18 | *Pet Supplies* | 1.437 | 1.394 | 1.387 | **1.362** | 5.22 | 2.30 | 1.80 |
| 19 | *Sports and Outdoors* | 1.207 | 1.125 | 1.114 | **1.080** | 10.52 | 4.00 | 3.05 |
| 20 | *Tools and Home* | 1.293 | 1.256 | 1.243 | **1.199** | 7.27 | 4.54 | 3.54 |
| 21 | *Toys and Games* | 1.107 | 1.071 | 1.076 | **1.038** | 6.23 | 3.08 | 3.53 |
| 22 | *Video Games* | 1.304 | 1.307 | 1.302 | **1.266** | 2.91 | 3.14 | 2.76 |
| | **Average** | 1.257 | 1.220 | 1.222 | **1.182** | 6.00 | 3.14 | 3.25 |

(Note: The best performance of each dataset is highlighted in bold.)

(b) **DRR** models the latent factors of user and item with two independent RNNs using review texts in a chronicle order for rating prediction. The overall rating is predicted by an inner product of the user and item latent factors. Following the methods in (Cvejoski et al., 2021), we utilized the pre-trained 300-*dim* word vectors from *GloVe* for the word embedding, set the hidden dimension of the temporal representation as $\boldsymbol{h}_t^e = 32$, the embedding dimension as $E = 100$, and the attention dimension as $A = 64$.

(c) **VAR** extracts the visual features of item images with an adversarial training procedure. It includes a suite of adversarial attacks against deep CNNs and defensive strategies to counteract them. Considering the complexity of this model, we kept the parameter settings the same as reported by Anelli et al. (2021).

Our proposed recommender (MSVA) attempts to model the review and visual representations from review documents, product images as well as ratings through semantic and image analyses. We embedded each review word with a 300-*dim* vector using *word2vec* as done in ANR. For fair comparisons with the baselines, we also set the number of aspects as $|\mathscr{A}| = 5$, and the dropout rate as $\rho = 0.5$. Additional parameters in MSVA, such as the set of kernel size $\mathscr{K}$, the set of block size $\mathscr{B}$, the hidden dimensions, $h_1$ and $h_2$, of the aspect-sentiment and visual transformations were selected as $\mathscr{K} = \{2, 3\}$, $\mathscr{B} = \{10, 20\}$, $h_1 = 50$ and $h_2 = 50$, respectively. Same as in (Anelli et al., 2021; Chin et al., 2018), we set the sizes of the word and visual-word vocabularies of each dataset to be 50,000 and 4,096, respectively. The baselines were trained with the Adam optimizer, a learning rate of 2e-3, a batch size of 128 and the MSE loss function. Additionally, the experiments were repeated five times with random seeds, and the average MSEs of the five tests were used in the following reports.

### 4.3. Model Comparison

Table 3 presents the performance comparisons between MSVA and the three baselines in terms of MSE, where the best result of each dataset is highlighted in bold. It is clear that MSVA outperforms ANR, DRR and VAR with the lowest MSEs on almost all the datasets, except the *CDs and Vinyl* and *Digital Music* datasets. Averagely, MSVA lowers the MSEs by 6.00%, 3.14% and 3.25% as opposed to ANR, DRR and VAR, respectively. Of the 22 sets of Amazon Product Data, some datasets, such as *CDs and Vinyl* and *Digital Music*, are content-based, which means that the visual features (e.g., shape, color, or style) of the products in the datasets are not as important as products' contents (e.g., music, movie, etc.). Some datasets, such as *Automotive, Clothing, Shoes and Jewelry and Home and Kitchen,* are appearance/visual-based, meaning visual features are critical to users' perceptions of products. For the visual-based datasets (1, 2, 3, 7, 11, 12, 19, and 20 in Table 3), VAR, which utilizes product images, performs better than the two review-based models, ANR and DRR. The review-based models are less effective when rating visual-based products if no visual features are commented on the reviews. On the other hand, ANR and DRR are superior to VAR for content-based databases (5, 8, 10, 14, and 16 in Table 3). The rest of the datasets require both content and visual features, and thus there is no certain trend in the performance of the three baselines. However, MSVA is able to attain low MSEs consistently across the 22 datasets because it includes both user reviews and item images in the rating

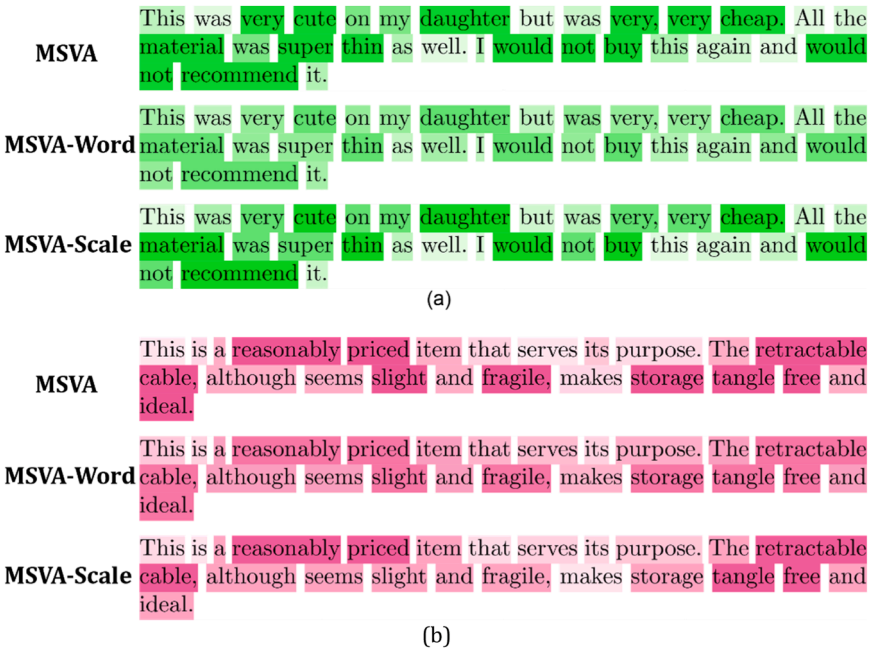


**Fig. 3.** Heatmap examples of review sentences by MSVA and two variants (MSVA-Word and MSVA-Scale) based on attention weights, (a) *Sports and Outdoors* datasets, (b) *Cell Phones and Accessories.*

prediction. Only in the two content-based datasets, *CDs and Vinyl* and *Digital Music*, MSVA generates the second lowest MSEs, slightly higher than DRR. It is also worth noting that ANR, DRR and MSVA have significantly lower MSEs (<1.000) on four datasets (4, 5, 8 and 13 in Table 3) than on the other datasets, so does VAR on three of them (4, 5 and 13). This means that recommendations made by these models for such products have better chances for users' acceptance.

## 5. Model Analysis

This section investigates the model ablation and optimal hyperparameters.

### 5.1. Model Ablation

In order to verify the effectiveness of the visual representation module and to investigate the contribution of each attention layer in MSVA, an ablation study was performed on MSVA and its five variants:

(1) MSVA: it is the complete model as described in Section 3 with the parameter settings stated in Section 4.2.
(2) MSVA-Word, MSVA-Scale, MSVA-Block and MSVA-Visual: they are four variants when MSVA excludes the word-aware, scale-aware, block-aware or visual attention layer. Removing an attention layer from the model means that its attention weights are set to be uniform for all entries.
(3) MSVA-R: it is the 5$^{th}$ variant when only the review representation module is taken into account, that is, the visual representation module is omitted.

Fig. 3 presents the examples for the heatmaps of review sentences made by MSVA, MSVA-Word, and MSVA-Scale with different color shades that highlight words in terms of their attention weights, $\mathbf{attn}^{word}_{u,a,k}$. The sentences were taken from two datasets, *Sports and Outdoors* (a) and *Cell Phones and Accessories* (b). Compared to MSVA, the MSVA-Word variant lacks the word importance information because it deletes the word-aware attention mechanism, and therefore the distinctions of the color shades in its heatmaps are lower than in the other heatmaps. As to the heatmaps of MSVA-Scale, it can be readily noticed that the shades of sentiment words that are at different distances to an aspect word (e.g., '*very' and 'cute'* to "*daughter*" in *Sports and Outdoor* datasets; '*slight' and 'fragile*' to "*cable*" in *Cell Phones and Accessories*) are not so differentiable because MSVA-Scale eliminates the scale-aware attention mechanism. The above comparisons provide an evidence for the fact that the performances of the two variants are degraded from MSVA.

Fig. 4 provides the MSE comparisons between MSVA and its five variants on the two example datasets: *Cell Phones and Accessories* and *Sports and Outdoors*. It is noticeable that the MSVA-R variant, which contains only the review representation module, generates the highest MSEs for both datasets, justifying the necessity of integrating visual features with semantic features to reduce prediction errors. In the review representation module, the performance of the MSVA-Scale variant (without the scale-aware attention layer) is lower than that of the MSVA-Word variant (without the word-aware attention layer) on both datasets. This means that the multiscale information generated by the scale-aware attention layer makes a significant contribution to the rating inference. The same phenomenon can be seen with the visual representation module. The MSVA-Visual variant (without the visual attention layer) has a worse performance than the MSVA-Block variant (without the block-aware attention layer), indicating the information generated with multi-block sizes in the visual attention layer is critical to the rating inference. Compared to the five variants, MSVA outputs the lowest MSEs on the two datasets, proving that the word-aware, scale-aware, block-aware and visual attention layers are all necessary components that contribute to the rating accuracy.

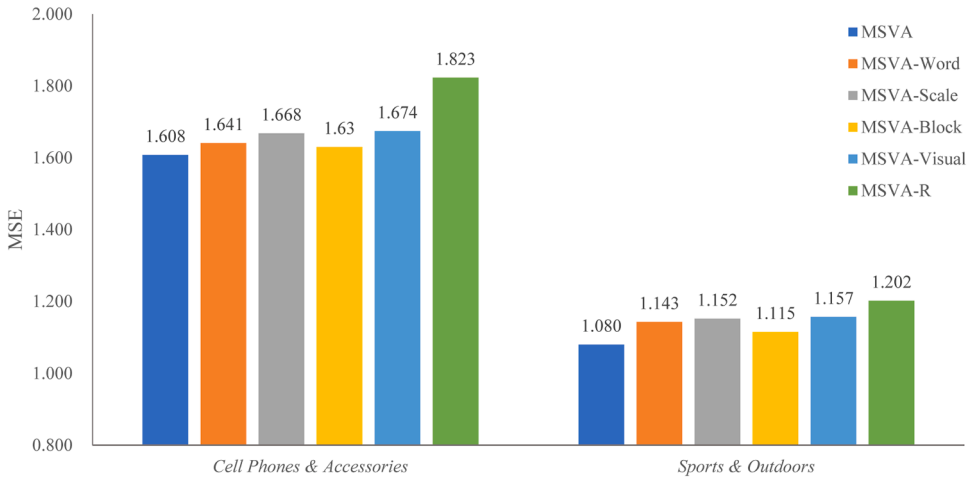


**Fig. 4.** MSEs of MSVA variants on *Cell Phones and Accessories* and *Sports and Outdoors* datasets.

### 5.2. Hyperparameter Sensitivity

Several hyperparameters of MSVA need to be optimized. Some of them, such as the number of aspects, $|\mathscr{A}|$, and the hidden dimensions of the review and visual transformations, $h_1$ and $h_2$, were explored in other studies before. Thus, we only focus on unique hyperparameters, i.e., the number of kernel size ($\boldsymbol{k}$) and the number of block size ($\boldsymbol{b}$). They represent multiscale ranges of review features and visual features, respectively. Let the set of kernel sizes ($\mathscr{K}$) be {2, 3, 4, 5, 6} and the set of block sizes ($\mathscr{B}$) be{10, 20, 30, 40, 50} for the testing. Due to the diversity in product categories, the needed scales may vary across the datasets. The first $\boldsymbol{k}$ elements in set $\mathscr{K}$ and the first $\boldsymbol{b}$ elements in set $\mathscr{B}$ are selected when $\boldsymbol{k}$ varies from 1 to $|\mathscr{K}| = 5$ and $\boldsymbol{b}$ from 1 to $|\mathscr{B}| = 5$. For example, the first kernel size (i.e., 2) will be used when $\boldsymbol{k}$ is 1; the first and second sizes (i.e., 2 and 3) will be used when $\boldsymbol{k}$ is 2; and so on. The 3D bar charts in Fig. 5 show the MSEs of four different datasets for $\boldsymbol{k}$ and $\boldsymbol{b}$ in the range of [1, 5]. It can be seen that MSE changes more drastically on *Cell Phones and Accessories* (Fig. 5a) and *Electronics* (Fig. 5c), compared with those on *Clothing, Shoes and Jewelry* (Fig. 5b) and *Sports and Outdoors* (Fig. 5d). Additionally, the lowest MSE is reached on the *Cell Phones and Accessories* and *Electronics* datasets when both $\boldsymbol{k}$ and $\boldsymbol{b}$ equal 2. On the *Clothing, Shoes and Jewelry* and *Sports and Outdoors* datasets (Figs. 5b and 5c), the MSE appears to be the second best/lowest when $\boldsymbol{k} = \boldsymbol{b} = 2$, or negligibly higher than the best.

## 6. Discussion

The model ablation analysis proves that MSVA outperforms its five variants. The MSVA-R variant, a review-based model, has the lowest performance among the five MSVA variants on the two visual-based datasets, *Cell Phones and Accessories* and *Sports and*

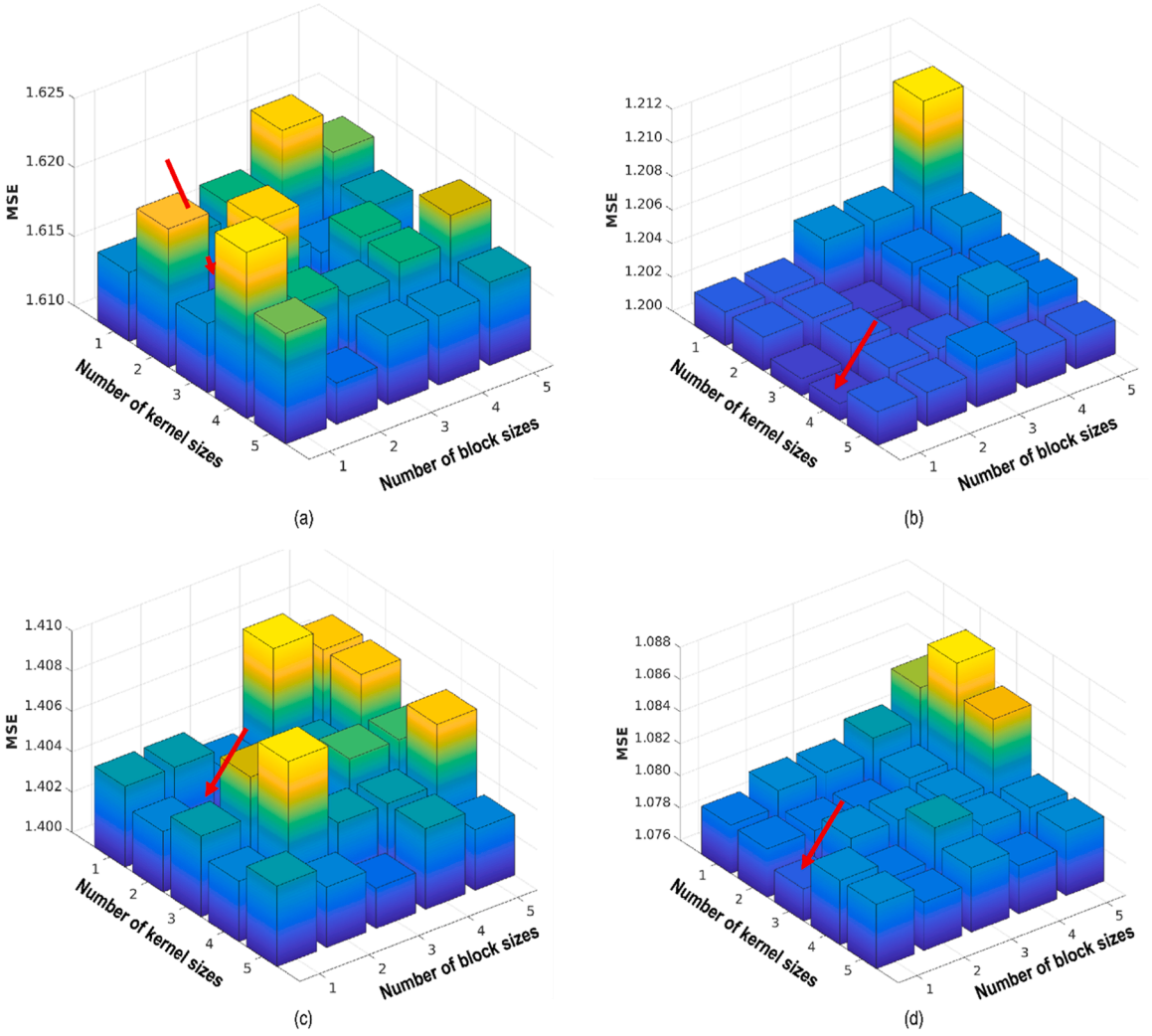


**Fig. 5.** Effects of the number of kernel size and block size on MSE for four datasets. (a) *Cell Phones and Accessories* datasets, (b) *Clothing, Shoes and Jewelry* datasets, (c) *Electronics* datasets, (d) *Sports and Outdoors* datasets.

*Outdoors*. Generally speaking, when visual features are important to a product, its images, which show product style and appearance, can be greatly useful for showing what a user likes or dislikes, and thus a visual representation module is a significant addition to the prediction model. Both MSVA-Scale and MSVA-Visual variants show lower performances than their peer variants, MSVA-Word and MSVA-Block. This is because the two variants skip mining latent factors of a review text in different kernel sizes and blocks of different sizes in a product image. The information captured by adjusting kernel sizes in the review representation module or block sizes in the visual representation module is needed for inferring ratings. In other words, multiscale information in both review and visual representations plays a big role in improving the accuracy of rating predictions.

In the hyperparameter analysis, the numbers of both the kernel sizes and the block sizes, $\boldsymbol{k}$ and $\boldsymbol{b}$, can influence the performance significantly on the four sampled datasets, *Cell Phones and Accessories, Clothing, Shoes and Jewelry, Electronics,* and *Sports and Outdoors* datasets (Fig. 5). Overall, when both $\boldsymbol{k}$ and $\boldsymbol{b}$ increase from 1 to 5, the MSEs of MSVA on the four datasets seem to change in different patterns. For *Cell Phones and Accessories* (Fig. 5a) and *Electronics* (Fig. 5c), an optimal selection for $\boldsymbol{k}$ and $\boldsymbol{b}$ seems to be 2. This means that only two kernel sizes {2, 3} and two block sizes {10, 20} should be used in the multiscale analyses. For *Clothing, Shoes and Jewelry* (Fig. 5b), the optimal ($\boldsymbol{k}$, $\boldsymbol{b}$) are (4, 1), that is, the set of kernel sizes ($\mathscr{K}$) and the set of block sizes ($\mathscr{B}$) should be {2, 3, 4, 5} and {10}, respectively. As shown in Fig. 5b, another good choice for ($\boldsymbol{k}$, $\boldsymbol{b}$) can be (3, 3) where the MSE is slightly higher than at (4, 1). Similarly, for *Sports and Outdoors* (Fig. 5d), the optimal ($\boldsymbol{k}$, $\boldsymbol{b}$) are (3, 1), i.e., $\mathscr{K} = \{2,\ 3,\ 4\}$ and $\mathscr{B} = \{10\}$. Based on the analysis of these four datasets, it can be seen that the optimal numbers of kernel sizes and block sizes vary with the product domains or categories of the datasets. This is because the importance of product images (or visual features) relative to review comments (or semantic features) in the rating prediction relies on the nature of product categories. But overall, $\boldsymbol{k}$ and $\boldsymbol{b}$ should be limited between 1 and 4. For an unknown dataset, a default value of 2 is recommended for both $\boldsymbol{k}$ and $\boldsymbol{b}$.

MSVA includes both user reviews and product images as its input, and thus is able to achieve the best performances on 20 datasets and the second best on two other datasets among the 22 Amazon datasets when compared to the three baselines. MSVA is a hybrid model that detects the latent factors of users and items from review texts and product images jointly to make the rating prediction more relevant to user's experience and preference. The multiscale scheme and the attention mechanisms introduced to both semantic and visual modules facilitate the mining of fine-grained implicit factors and the end-to-end training. MSVA represents a new big-data approach proven to be effective for datasets containing diverse product categories. However, there are a few improvements needed in the future study:

(1) MSVA does not consider the timestamp of user reviews, instead, it merges all the reviews by the same user or for the same item into an input document. However, the time of the reviews reflects the trend of user's preference, and thus can be a useful factor.
(2) MSVA uses a pretrained word2vec model for its word-embedding, but the words in the model may not fit well to a specific domain. Retraining an adaptive word2vec model or using context-preserving model like BERT (Devlin, et al. 2019) will be considered in the future.
(3) For the extraction of visual features, MSVA currently uses only product images, not images attached to user reviews. Since the attached images are more indicative of personal preference, we will make efforts to incorporate them into MSVA in the future.

## 7. Conclusion

This paper presented a novel end-to-end personalized recommendation model based on multiscale semantic-visual representation analysis (MSVA) for latent factors in user reviews and item images. MSVA uses two parallel modules to process review documents and product images simultaneously. In the review module, the multiscale semantic representations are extracted from review texts in various aspects with word-aware and scale-aware attention mechanisms. In the visual module, the multiscale visual representations are learned from item images with the block-aware and visual-aware attention mechanisms at multiple block sizes. By combining the semantic and visual representations, MSVA can improve the generalization of latent factors to strengthen rating inference based on user's sentiment and preference over purchased items. After the training with over 21 million reviews and over 9 million product images from the 22 real-world Amazon product datasets MSVA achieved the best performance on 20 datasets and the second best on two other datasets against the three state-of-the-art baseline models. More specifically, MSVA reduced the average MSE of predicted ratings by 6.00%, 3.14% and 3.25% compared to ANR, DRR and VAR, respectively. The experiment also proved the visual representation module and the multiscale scheme were necessary additions that made MSVA's rating predictions more robust and accurate when dealing with datasets in a wide range of product domains.

In the future study, the timestamp and the images attached to user reviews should be included in MSVA's input, and the word2vec model should be retrained with words more relevant to the domains of the target datasets.

## Declaration of Competing Interest

The authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

In this study, Zhu Zhan was sponsored by the Outstanding Doctoral Student Visiting Program from Donghua University, China, and a fund from University of North Texas, USA.

## References

Amazon, Amazon Product Data. Accessed: August 1, 2021. http://jmcauley.ucsd.edu/data/amazon.

Anelli, V. W., Deldjoo, Y., Di Noia, T., Malitesta, D., & Merra, F. A. (2021). A Study of Defensive Methods to Protect Visual Recommendation against Adversarial Manipulation of Images. In *, 1. SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* Association for Computing Machinery.

Bauman, K., Liu, B., & Tuzhilin, A. (2017). Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F129685* (pp. 717–725).

Bottou, L. (2012). Stochastic gradient descent tricks. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7700 LECTURE NO(1)* (pp. 421–436).

Catherine, R., & Cohen, W. (2017). TransNets: Learning to transform for recommendation. In *RecSys 2017 - Proceedings of the 11th ACM Conference on Recommender Systems* (pp. 288–296).

Chambua, J., & Niu, Z. (2021). Review text based rating prediction approaches: preference knowledge learning, representation and utilization. *Artificial Intelligence Review, 54*(2), 1171–1200.

Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., & Kankanhalli, M. (2018). A3NCF: An adaptive aspect attention model for rating prediction. In *IJCAI International Joint Conference on Artificial Intelligence, 2018* (pp. 3748–3754).

Chin, J. Y., Joty, S., Zhao, K., & Cong, G. (2018). ANR: Aspect-based Neural Recommender. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 147–156).

Cvejoski, K., Sanchez, R. J., Bauckhage, C., & Ojeda, C. (2021). Dynamic Review-based Recommenders. In *International Data Science Conference 2021 (iDSC21)*. arXiv: 2110.14747.

Da'u, A., Salim, N., Rabiu, I., & Osman, A. (2020). Recommendation system exploiting aspect-based opinion mining with deep learning method. *Information Sciences, 512*, 1279–1292.

Demeester, T., Rocktäschel, T., & Riedel, S. (2016). Lifted rule injection for relation embeddings. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (pp. 1389–1399).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

Guan, X., Cheng, Z., He, X., Zhang, Y., Zhu, Z., Peng, Q., & Chua, T. S. (2019). Attentive aspect modeling for review-aware recommendation. *ACM Transactions on Information Systems, 37*(3), 1–27.

He, C., Liu, Y., Guo, Q., and Miao, C. (2019). Multi-Scale Quasi-RNN for next item recommendation. arXiv preprint arXiv:1902.09849.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December* (pp. 770–778).

He, R., Lin, C., & McAuley, J. (2016). Fashionista: A fashion-aware graphical system for exploring visually similar items. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 199–202).

He, R., & McAuley, J. (2016a). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *25th International World Wide Web Conference, WWW 2016* (pp. 507–517).

He, R., & McAuley, J. (2016b). VBPR: Visual Bayesian personalized ranking from implicit feedback. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016* (pp. 144–150).

He, X., He, Z., Du, X., & Chua, T. S. (2018). Adversarial personalized ranking for recommendation. In *, 10. 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018* (pp. 355–364).

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

Li, W., & Xu, B. (2020). Aspect-Based Fashion Recommendation with Attention Mechanism. *IEEE Access, 8*, 141814–141823.

Liu, H., Wang, W., Chen, H., Zhang, W., Peng, Q., Pan, L., & Jiao, P. (2020). Hierarchical Multi-view Attention for Neural Review-Based Recommendation. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 267–278).

Liu, H., Wang, W., Peng, Q., Wu, N., Wu, F., & Jiao, P. (2021). Toward Comprehensive User and Item Representations via Three-tier Attention Network. *ACM Transactions on Information Systems, 39*(3), 25.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*. NeurIPS.

Song, B., Zhang, W., Cao, Y., & Xu, C. (2019). Session-based recommendation with hierarchical memory networks. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 2181–2184).

Song, X., Chen, J., Han, X., Xu, X. S., Li, Y., & Nie, L. (2019). GP-BPR: Personalized compatibility modeling for clothing matching. In *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (pp. 320–328).

Tang, J., Du, X., He, X., Yuan, F., Tian, Q., & Chua, T. S. (2020). Adversarial Training towards Robust Multimedia Recommender System. *IEEE Transactions on Knowledge and Data Engineering, 32*(5), 855–867.

Wang, S., Huang, M., & Deng, Z. (2018). Densely connected CNN with multi-scale feature attention for text classification. In *IJCAI International Joint Conference on Artificial Intelligence, 2018-July* (pp. 4468–4474).

Xie, J., Zhu, F., Li, X., Huang, S., & Liu, S. (2021). Attentive preference personalized recommendation with sentence-level explanations. *Neurocomputing, 426*, 235–247.

Zhang, Y., Yin, G., Dong, H., & Zhang, L. (2022). Attention-based frequency-aware multi-scale network for sequential recommendation. *Applied Soft Computing, 127*, Article 109349.

Zheng, L., Noroozi, V., & Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining* (pp. 425–433).

Zheng, W., Zheng, Z., Wan, H., & Chen, C. (2019). Dynamically route hierarchical structure representation to attentive capsule for text classification. In *IJCAI International Joint Conference on Artificial Intelligence, 2019-August* (pp. 5464–5470).

Zheng, Y., Wang, H., Zhang, Y., Gao, X., & Min, X. (2020). Poly(A)-DG: a deep-learning-based domain generalization method to identify cross-species Ploy(A) signal without prior knowledge from target species. *PLOS Computational Biology*.