



Multimodal Deep Learning Framework for Book Recommendations: Harnessing Image Processing with VGG16 and Textual Analysis via LSTM-Enhanced Word2Vec

Yanling Li^{*}, Xin Li, Qian Zhao

College of Artificial Intelligence, Baoding University, Baoding 071000, China

Corresponding Author Email: liyanling@bdu.edu.cn

<https://doi.org/10.18280/ts.400406>

ABSTRACT

Received: 18 April 2023

Revised: 25 June 2023

Accepted: 2 July 2023

Available online: 31 August 2023

Keywords:

book recommendation systems, deep learning, multimodal information processing, VGG16, Word2Vec, LSTM, CBAM attention mechanism

In the contemporary digital age, an intensified emphasis has been placed on the research of book recommendation systems. Historically, these systems predominantly focused on readers' past preferences, overlooking the inherent characteristics of the book's content and design. To address this gap, a novel algorithm, leveraging both multimodal image processing and deep learning, was designed. Features from book cover images were extracted using the VGG16 model, while textual attributes were discerned through a combination of the Word2Vec model and LSTM neural networks. The integration of the CBAM attention mechanism culminated in the creation of a modality-weighted feature fusion module, facilitating the dynamic allocation of feature weights. Furthermore, an objective function for this recommendation model was formulated, ensuring the enhancement of its performance during the training phase. This study not only presents a groundbreaking methodology to amplify the efficacy and resilience of book recommendation systems but also broadens understanding in the realm of multimodal information processing within deep learning-based recommendation platforms.

1. INTRODUCTION

In the era of digital transformation, profound shifts in the methods of knowledge and information acquisition have been observed. Despite such changes, books, as repositories of knowledge, have consistently retained their paramount significance [1, 2]. However, with the information surge, the pressing challenge that has been posed is the identification of books resonating with individual preferences from an expansive collection. Book recommendation systems, underpinned by big data and algorithmic constructs, have been developed to mitigate this issue [3-6]. Yet, a predominant focus on readers' historical behaviors and preferences has been noted, often sidelining vital attributes such as textual content and cover aesthetics [7-10]. To address this shortcoming, an algorithm incorporating multimodal image processing and deep learning techniques has been introduced.

Books, by their intricate nature, are recognized not solely for their textual content but also their visual elements. Textual components, including titles, authors, and synopses, have been shown to inform potential readers about the book's thematic core. In contrast, visual elements primarily encompass cover designs and hues, playing a pivotal role in shaping a reader's sensory perception of a book's tone and style [11-14]. An amalgamation of these elements offers a holistic understanding of a book, invariably influencing readers' selections. Therefore, an in-depth comprehension of these multimodal facets has been deemed imperative for refining recommendation systems' accuracy and enriching user experiences. The potency of deep learning techniques in feature discernment accentuates their prospective utility in this domain [15, 16].

However, a significant portion of existing book

recommendation systems, despite adopting deep learning approaches, have been observed to remain tethered to a single modality, emphasizing either text or visuals to the detriment of the other [17-19]. Such an inclination has been found to overlook the synergistic potential of textual and visual information. Moreover, in instances where amalgamation attempts have been made, rudimentary methods like averaging or linear weighting dominate, often neglecting the nuanced importance of varying features [20-22].

In light of the issues delineated, the primary objective delineated in this study centered on the design of a book recommendation algorithm, synergizing multimodal image processing and deep learning. Features from book cover images were extracted via the VGG16 model, while textual attributes were discerned employing a fusion of the Word2Vec model and LSTM neural networks. Further refinement was achieved through the integration of the CBAM attention mechanism, facilitating the development of a modality-weighted feature fusion module, adept at dynamically adjusting feature weights, thus bolstering recommendation accuracy. This study not only broadens the horizon for book recommendation system analyses but also underscores the expansive applicability of multimodal information processing within recommendation system architectures.

2. EXTRACTION OF MULTIMODAL FEATURES FROM BOOK IMAGES AND TEXT

In the realm of book recommendation, the integration of both image and text modalities from books has been recognized as pivotal. The significance of this integration is twofold. On one hand, features and attributes embedded within

book images, predominantly relayed through cover designs and color schemes, are believed to capture a reader's visual attention, potentially influence their reading interest, and, to some degree, mirror the book's thematic essence. On the other, textual elements, spanning titles, authors, publishers, and synopses, are understood to convey the fundamental content and thematic proposition of a book. Therefore, the act of synergistically merging these two distinct modalities is perceived as enhancing the portrayal of a book's multi-dimensional facets and better pinpointing readers' literary inclinations, thereby bolstering recommendation precision and impact.

In the ensuing fusion of image and textual information, steps undertaken for the extraction of features from these sources are regarded as critical. It has been acknowledged that the precision and efficiency with which these features are extracted bear significant implications for the recommendation algorithm's effectiveness. Deep learning models are often utilized for the extraction of image features, culminating in abstract representations of visual components like cover aesthetics and color blends. Concurrently, textual features are typically derived via word vector models combined with deep learning methodologies, yielding vectorized depictions of text elements such as titles and synopses. By employing these methods, conversion of unstructured image and text data into structured feature vectors is achieved, streamlining subsequent computational tasks.

However, a salient challenge presented in multimodal book recommendation is the adept fusion of both image and text modalities. Due to the inherently diverse origins and characteristics of book image and text data, epitomizing their synergistic potential and achieving a deep-rooted fusion to

amplify recommendation accuracy becomes paramount in the design of recommendation systems. A feature-level modality fusion approach has been suggested in this research context. This approach dictates that feature extraction for image and text data be conducted in isolation, with the subsequent convergence of this information at the feature tier. Such a strategy is formulated with the intent of achieving a profound integration of diverse modalities, encapsulating the unique traits of each modality while also accounting for their inter-modal correlations.

VGG16, a prominent deep convolutional neural network, is frequently employed for the extraction of image features. However, when utilized exclusively as a feature extractor for book images, certain limitations have been observed. Central to the architecture of VGG16, the fully connected layers, particularly those positioned at the latter stages, have been identified as critical. Yet, these layers have been noted to display a relative insensitivity to the positional data of the input, often bypassing crucial spatial structural information inherent in images. In the context of book images, while the background might be deemed inconsequential to the book itself, spatial attributes such as layout, design, and color schemes of book covers have been found indicative of both content and thematic elements. Thus, the application of these fully connected layers might inadvertently lead to a diminution of pivotal spatial information.

Given an input delineated by feature vectors $z_1, z_2, z_3, \dots, z_b$ and the formation of a layer encompassing three neurons within the fully connected layers, the resultant output is defined as:

$$p = q_1 \times z_1 + q_2 \times z_2 + q_3 \times z_3 + \dots + q_b \times z_b \quad (1)$$

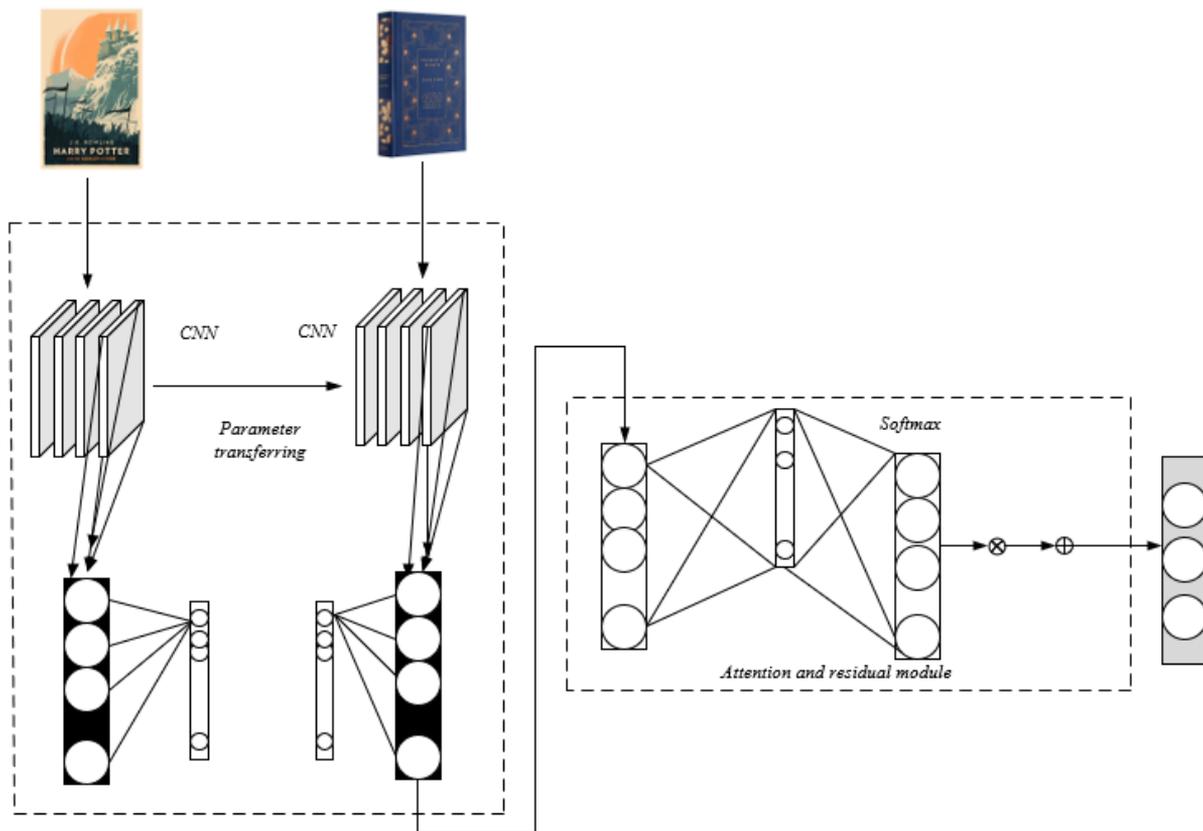


Figure 1. Application of attention mechanism in the constructed model

For subsequent input vectors $z_1, z_2, z_3, \dots, z_b$, the output is articulated as $p=q_1 \times z_2 + q_2 \times z_1 + q_3 \times z_3 + \dots + q_b \times z_b$, inadvertently accentuating feature z_2 . Through the introduction of function d , defined as $q_1=d(z_1), q_2=d(z_2), q_3=d(z_3), \dots, q_b=d(z_b)$, the final output can be re-expressed as $p=d(z_1) \times z_2 + d(z_2) \times z_1 + d(z_3) \times z_3 + \dots + d(z_b) \times z_b$. Given the pretraining status of VGG16 with its immutable weights, its adaptability to specificities inherent to book images might be questioned. While its efficacy in broader image processing contexts is well-documented, potential shortcomings in extracting niche features specific to domains such as book images have been suggested. As elucidated in the provided steps, the incorporation of function d appears to rectify complications arising from immutable weights within the fully connected layers. The assimilation of the attention mechanism within this model's framework can be visualized in Figure 1.

Additionally, the myriad parameters contained within the fully connected layers of VGG16 have been postulated to escalate the model's intricacy, potentially increasing susceptibility to overfitting, especially in data-scarce scenarios. With book images, which might not exhibit inherent complexity, the application of VGG16 in data-limited environments could compromise the model's ability to generalize. In efforts to counter these issues, the pre-weighted feature vectors were observed to be mapped identically to their post-weighted counterparts. Assuming VGG16's capability to encode the book image into a unidimensional vector θ_z, θ_y , and with $q_z, q_y \in \mathcal{R}^f$ represented by $d_{IM}(z) = v_z \in \mathcal{E}^f$, the feature dimension has been denoted as $f = 512$. The model-derived weights, symbolized as $q_z, q_y \in \mathcal{R}^f$, are subsequently described as:

$$\theta_z = v_z + (q_z \otimes v_z) \tag{2}$$

$$\theta_y = v_y + (q_y \otimes v_y) \tag{3}$$

Within the scope of this investigation, the extraction of textual features from books was methodically executed through three distinct phases:

(1) Preprocessing of Book Text Data:

Preprocessing is universally acknowledged as an indispensable precursor in the realm of natural language processing. Its primary aim was identified as the purging of extraneous noise from the dataset, coupled with the selective extraction of salient information. During this phase, several processes were typically incorporated, including normalization, cleansing, tokenization, stop-word removal, lemmatization, and stemming.

(2) Conversion of Text to Vector Representations via the Word2Vec Model:

Word2Vec, renowned for its aptitude in distilling word vector representations, has been demonstrated to transform each term into a continuous vector. Through such transformations, semantic affiliations between terms were effectively captured. Upon completion of preprocessing, the training of the Word2Vec model was undertaken. The resultant vector representations were discerned to encompass facets such as lexical semantics, syntactic order, and nuanced linguistic attributes, proffering a semantic depth seldom achieved by traditional models, like the bag-of-words approach.

(3) Gleaning of Textual Features via LSTM Neural Networks:

With the vector representations meticulously derived from the Word2Vec model, the next course of action entailed the employment of deep learning algorithms, notably LSTM, to excavate superior text features. LSTM, an acronym for Long Short-Term Memory networks, belongs to a specialized cadre of Recurrent Neural Networks (RNN) and is adept at processing sequential datasets, inclusive of text. These architectures have been lauded for their proficiency in managing elongated sequences and skirting the long-term dependency quandaries endemic to rudimentary RNNs. The vector sequences, once procured from Word2Vec, were subsequently introduced into the LSTM architecture, culminating in the extraction of refined features from elongated text segments.

3. DESIGN OF THE BOOK MULTI-MODALITY WEIGHTED FEATURE FUSION MODULE

Within the realm of multimodal information processing, distinct modalities are often observed to harbor specific advantages and characteristics. While visual cues are generally extracted from image data, descriptive and contextual insights are primarily sourced from text data. The disparities inherently present between such modalities could pose challenges to efficient book category classification unless aptly fused. In response to this, a modality-weighted feature fusion module was meticulously designed to amalgamate information across different modalities seamlessly. At the heart of this module, a novel modality-weighted fusion layer was introduced. Features from diverse modalities could be dynamically weighted by this layer, thus assigning varying degrees of prominence to each. The determination of such weights was conjectured to be influenced by factors such as the qualitative richness of information from each modality and its corresponding impact on book classification. Moreover, the integration of the CBAM attention mechanism, an attention model revered for its acuity in discerning and accentuating salient features, was seen to bolster the discriminatory prowess of the model. Such strategic integrations within the module were postulated to further refine feature selection and fusion processes, potentially heightening the precision of book category classifications.

Given the inherent discrepancies between book image and text modalities, a specialized modality-weighted fusion layer was meticulously devised. This layer's central tenet was the judicious allocation of discrete weights to both book image and text modalities, ensuring an efficient synthesis of features from these divergent sources. This synthesis was anticipated to yield a composite feature map, emblematic of both image and textual data. A preliminary step involved the dimensionality reduction of feature maps from both modalities, an endeavor successfully achieved via the Network in Network (NIN) layer. Characterized primarily for its dimensional transformation capabilities, the NIN layer played an instrumental role in compressing the dimensions of these feature maps. Upon undergoing this compression, uniformity in feature map sizes was achieved. The modality-weighted fusion layer then engaged in the act of attributing weights to these unified feature maps. This act of weighting signified the automated delineation of importance to features sourced from the dual modalities; naturally, a modality with augmented weight held greater sway in influencing the resultant combined feature map. Post this modality-centric weighting, the two

weighted modal feature maps were seamlessly concatenated, forging a composite feature map that harmoniously integrated features from both modalities (as shown in Figure 2).

Such architectural intricacies ensured that, even as differential weights were allocated to varied modalities, the channel count of the fused feature map invariably mirrored that of a singular modality feature map. This mirrored structure suggested that parameters within subsequent object detection modules could be redeployed without necessitating architectural adjustments, thereby enhancing the model's adaptability and potentially augmenting its performance metrics.

Within the multi-modal feature processing domain, the multi-scale feature maps of the book image were defined as $\{C_1, C_2, C_3\}$, whereas those corresponding to the book text were denoted as $\{U_1, U_2, U_3\}$. Post the application of the Network in Network (NIN) layer, the compressed feature maps for both the book image and book text modalities were respectively represented as C_M and U_M . In this context, the NIN function was symbolized by d^{bub} , while the concat fusion function was articulated as D^{CA} .

Emerging from the modality-weighted fusion layer, feature descriptors for the book image and text modalities were respectively denoted as S_c and S_u . Their collective summation was represented as S_l . Across the three distinct scales, the resultant weighted fusion feature maps were characterized as $\{L_1, L_2, L_3\}$. For each scale, the mathematical representation of the weighted fusion feature map is provided in Eq. (4).

$$L_u = d^{CA} \left(d^{bub} (C_u) \times S_c / (S_l), d^{bub} (U_u) \times (S_u / (S_l)) \right) \quad (4)$$

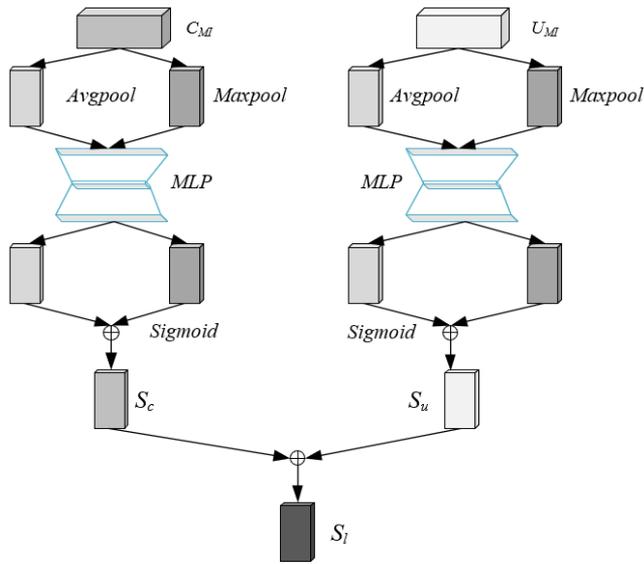


Figure 2. Intricately delineates the architecture of the modality-weighted fusion layer

In a bid to further refine the modality-weighted fusion layer and bolster both the precision and efficiency of feature extraction, the CBAM (Convolutional Block Attention Module) attention mechanism was subsequently integrated. This mechanism's strategic configuration is vividly showcased in Figure 3. By selectively navigating through the channels and spatial dimensions of the feature maps, attention was efficiently directed towards pivotal features, thus enabling the

model to accentuate them, while concurrently diminishing the impact of redundant or less significant features.

Inputs directed into the CBAM layer were expressed as $\{L_1, L_2, L_3\}$. Under the assumption that the enhanced multi-scale weighted fusion attention feature maps, influenced by this attention mechanism, were denoted by $\{L_{S1}, L_{S2}, L_{S3}\}$, with the spatial attention mechanism represented as d^{KJ} and the channel attention mechanism illustrated as d^{TD} , the mathematical expression corresponding to the feature map for each respective scale is detailed in Eq. (5).

$$L_{Su} = d^{KJ} \left(d^{TD} (L_u) \right) \quad (5)$$

Within the realm of the channel attention mechanism, it was posited that the channel count of the input multi-modal fusion feature map, termed as D , was denoted by V , while its height and width were represented as G and Q respectively. Thus, the mathematical representation of D was formalized as $D \in R^{V \times G \times Q}$. Further, the Sigmoid function, commonly encountered in these computations, was expressed as SM , with the average pooling and max pooling operations symbolized by AP and MP respectively. Through the integration of these operations, a transformed feature map, D' , was derived, as detailed in Eq. (6).

$$D' = D \times SM \left(\Gamma \left(AP(D) \right) + \Gamma \left(MP(D) \right) \right) \quad (6)$$

Transitioning to the spatial attention mechanism, $D' \in E^{V \times G \times Q}$ was accepted as the input, yielding D'' as the subsequent output feature map. To clarify further terminologies in this context, it was inferred that the book image feature map was signified by C , the book text feature map by U , the fusion feature map by L , and the attention mechanism-augmented fusion feature map was distinguished as L_S . A comprehensive computational representation encompassing these terminologies is provided in Eq. (7).

$$D'' = D' \times \left(SM \left(d^{7 \times 7} \left(d^{TD} \left(MP(D), AP(D) \right) \right) \right) \right) \quad (7)$$

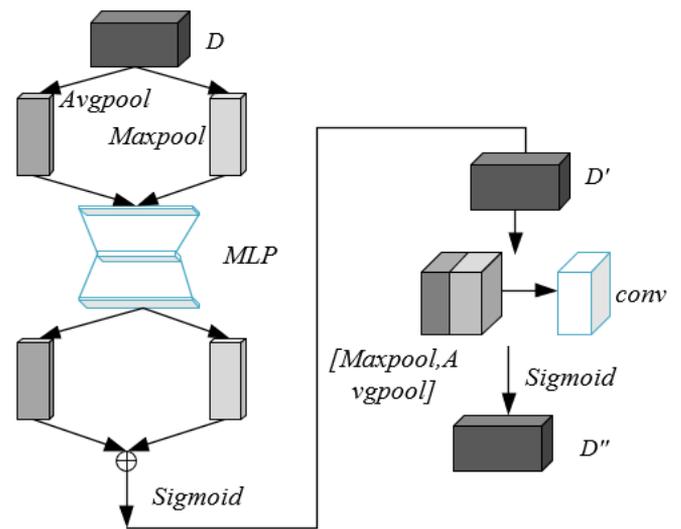


Figure 3. Meticulously outlines the structure of the CBAM attention mechanism

4. FORMULATION OF THE OBJECTIVE FUNCTION FOR MULTI-MODAL PERSONALIZED RECOMMENDATION

The formulation of a multi-modal book image personalized recommendation model, integrating both image and text data from books, has been discerned as a classification task. A premise of high-quality personalized recommendations is the accurate classification of samples. In this endeavor, the objective function designed is characterized by a seamless melding of cross-entropy and triplet loss functions. Such synthesis is postulated to ponder deeply upon the spatial distribution of book features whilst simultaneously regulating the distances between samples within similar and dissimilar book categories. The aim of this approach is to augment the classification accuracy of the model, thus potentially enriching the caliber of personalized book recommendations.

Cross-entropy loss function: Commonly adopted for classification tasks, the main pursuit of this function is to minimize disparities between the model's projected probability distribution and the authentic probability distribution associated with the true label. By incorporating cross-entropy, a nuanced comprehension of the global feature space distribution of books is facilitated. The computation for the same is delineated in the subsequent equation:

$$\begin{aligned}
 F(S||N) &= \sum_y O_A(z_u) \log\left(\frac{O_S(z_u)}{O_N(z_u)}\right) \\
 &= \sum_y P_S(z_u) \log(O_S(z_u)) - \sum_y O_S(z_u) \log(O_N(z_u)) \quad (8) \\
 &= G(O_S(z)) + \left[-\sum_u O_S(z_u) \log(O_N(z_u))\right]
 \end{aligned}$$

KL divergence between probability distributions S and N is denoted as $F(S||N)$. Distinct probability distributions are portrayed as O_S and O_N . Notably, in contexts encompassing deep learning, the entropy of S , denoted by $G(O_S(z))$, remains invariant, given that O signifies the recognized distribution of training data for both book image and text. Considering a category count denoted by b , the label symbolized by u , the genuine probability distribution as $O_S(z_u)$, and the anticipated distribution probability as $O_N(z_u)$, the extraction of KL divergence $F(S||N)$ between the distributions S and N becomes analogous to the cross-entropy $F(S||N)$, with the operative term being M_{VR} :

$$M_{VR} = G(S, N) = -\sum_{u=1}^b O_S(z_u) \ln(O_N(z_u)) \quad (9)$$

The utility of the triplet loss function in learning sample similarities is well-established. It endeavors to minimize distances between books within a category whilst magnifying differences between books from distinct categories, thereby enhancing the delineation between analogous and non-analogous books. Representative models of this application are elucidated in Figure 4. As a consequence, it is observed that models become proficient in capitalizing on the distance metrics between distinct book features. Given that book features are denoted by S , O , and B , with their corresponding encodings being $f(S)$, $f(O)$, and $f(B)$, the computational formula is articulated as:

$$M_Y(S, O, B) = MAX\left(\|d(S) - d(O)\|^2 - \|d(S) - d(B)\|^2 + \beta, 0\right) \quad (10)$$

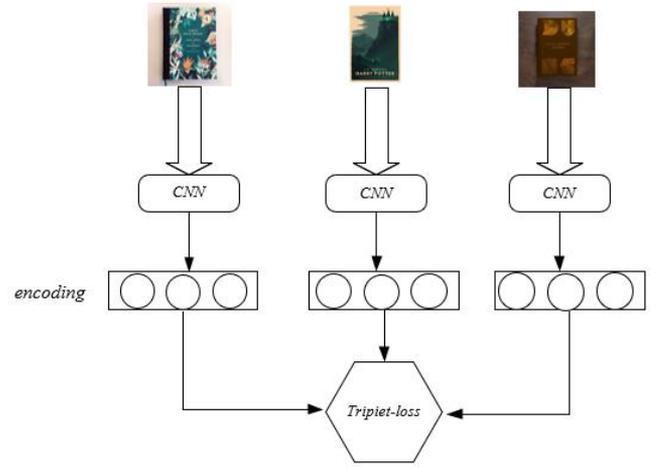


Figure 4. Elucidates the application framework for the triplet-loss model

Interrelationships between these book features are encapsulated in the equations:

$$f(S, O) = \|d(S) - d(O)\|_2^2 \quad (11)$$

$$f(S, B) = \|d(S) - d(B)\|_2^2 \quad (12)$$

The integration of the L2 norm is recognized as a mechanism to regulate book feature magnitudes, potentially curbing overfitting tendencies during the training phase. This strategy is posited to amplify the model's capability to generalize. The culminating representation of the triplet loss function is hence:

$$M_Y(S, O, B) = \log\left(1 - \exp(f(S, O) - f(S, B))\right) \quad (13)$$

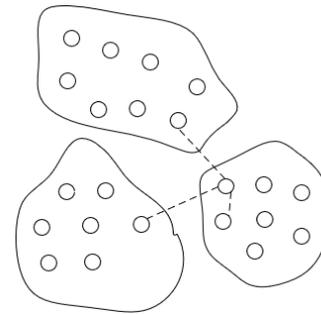


Figure 5. Visual representation of the distribution of book samples during the recommendation phase

In the realm of personalized recommendation systems, the ubiquity of the cross-entropy loss function for classifier training is well-documented. Nonetheless, a standalone dependence on this function in classification tasks appears to accentuate holistic category predictions, while possibly sidelining the intricate spatial dynamics among samples. Such dynamics imply that analogous books ought to be spatially adjacent in the feature space, with contrasting books being considerably distant. While an elevated classification accuracy might be achieved without these spatial constraints, it is posited that subtleties in user inclinations might remain unaddressed during personalized recommendations. A

visualization of the book sample distribution throughout the recommendation process is portrayed in Figure 5. To surmount this observed lacuna, the incorporation of the triplet loss function is proposed.

Incorporating a triplet-based paradigm, the triplet loss function is formulated to ensure that distances between anchor samples and positive samples are consistently smaller than those between anchor samples and negative counterparts. Through this methodology, spatial relationships within the feature space are systematically constrained. When the cross-entropy and triplet loss functions are combined, it is observed that the resultant objective function not only upholds classification accuracy but also amplifies the distinction between book feature distances. Consequently, analogous books exhibit proximity in the feature space, whereas contrasting books demonstrate noticeable spatial separation. Such a framework facilitates recommendation systems to extrapolate global feature distributions while also capitalizing on localized similarities, ultimately refining the caliber of personalized book recommendations.

Given the weights of cross-entropy and triplet loss functions are denoted by β and α respectively, and the category count and batch size are represented as G and N , relationships between samples emerge. For instance, samples analogous to X_u are indicated as X_u^+ , while those contrasting X_u are marked by X_k . The relationships can be articulated as:

$$M = \beta M_Y + \alpha M_{VR}$$

$$= \frac{1}{N} \left(\beta \sum_{u,k \neq 1, u \neq k}^N \log \left\{ 1 + \exp \left\{ \begin{array}{l} f(X_u, X_u^+) \\ -f(X_k, X_k^+) \end{array} \right\} \right\} \right. \\ \left. - \alpha \sum_{u=1}^N \sum_{g=1}^G o(X_{ug}^*) \log w(X_{ug}^*) \right) \quad (14)$$

At the classification stratum, once fused features undergo mapping through a softmax function, a distinct probability distribution emerges, represented as X_u^* . The probability distribution discerned subsequent to the softmax operation is encapsulated by $w(X_{ug}^*)$. Thus, the culmination of the integration of book image and text features can be defined as:

$$X_u = d_{CB}(z_u, y_u)^{QU} \quad (15)$$

The true probability distribution is demarcated as $o(X_{ug}^*)$. From this delineation, it is discerned that the fused feature conforms to:

$$X_u \rightarrow DV(X_u) \rightarrow w(X_{ug}^*) = \frac{\exp(X_{ug}^*)}{\sum_{g=1}^G \exp(X_{ug}^*)} \quad (16)$$

With the overarching aim of catering to personalized recommendation requisites, and to accentuate the variances between analogous and diverse book categories, the squared norm was systematically employed. This approach constrained distances between book features within the spatial domain. When the matrix dimension of book features within the fully connected layer is represented by F , the subsequent relationship is delineated:

$$f(X_u, X_k) = \sum_{j=1}^F \|X_u(j) - X_k(j)\|_2^2 \quad (17)$$

5. EXPERIMENTAL OUTCOMES AND INTERPRETATION

From Table 1, variations in the efficacy of diverse feature extraction methodologies can be observed, gauged against metrics such as sensitivity, specificity, positive predictive value (PPV), intersection over union (IoU), and the Dice coefficient. Among the explored techniques, the highest sensitivity, recorded at 0.9818, was observed for the approach presented in the present study. This result implies that 98.18% of true positive instances were accurately identified. In parallel, a remarkable specificity of 0.9947 was demonstrated by the same method, suggesting that 99.47% of genuine negative instances were correctly isolated. A PPV of 0.9032 was also documented, inferring that of its outcomes, 90.32% were verifiable positives. An apex IoU value of 0.9339 emphasizes that the predictions of the model coincided with 93.39% of actual outcomes, thus reinforcing its reliability. Regarding the Dice coefficient, a value of 0.9864 was achieved by the method under discussion, further validating the close correspondence between its projections and actual observations.

Table 1. Comparative efficacy of diverse book image feature extraction techniques

Model	Sensitivity	Specificity	PPV	IoU	Dice
<i>SIFT</i>	0.8049	0.9038	0.7173	0.6103	0.8883
<i>SURF</i>	0.9037	0.9083	0.8397	0.8038	0.8937
<i>Gabor filtering</i>	0.9247	0.9259	0.8628	0.8468	0.8284
<i>HOG</i>	0.9284	0.9537	0.8498	0.8193	0.8478
<i>LBP</i>	0.9109	0.9748	0.8049	0.8038	0.8478
<i>Canny</i>	0.9108	0.9004	0.8560	0.8630	0.8208
<i>ResNet</i>	0.9249	0.9294	0.8987	0.8398	0.9398
<i>PCA</i>	0.9349	0.9048	0.8203	0.8345	0.9309
<i>BRIEF</i>	0.9229	0.9249	0.8734	0.8849	0.9298
<i>Ours</i>	0.9818	0.9947	0.9032	0.9339	0.9864

Upon analysis of Figure 6, it becomes evident that the visualization outcomes for book images and text, when mapped onto a two-dimensional plane after feature extraction employing the discussed technique, are commendable. Samples that appeared initially as disparate entities were observed to have been transformed into a structured and interrelated assembly post-processing. It was noted that congruent feature vectors were situated adjacently within the defined space, suggesting the efficiency of the employed feature extraction technique in amalgamating and accentuating the salient traits of both book imagery and accompanying text. A critical element that likely contributed to this efficacy was the incorporation of the modality-weighted feature fusion module during the fusion phase. By affording variable weights to distinct modalities, an integration of features from disparate sources was effectively realized. This modality ensured optimal utilization of both image and textual information, potentially augmenting recommendation precision. In subsequent optimization stages, both the CBAM attention mechanism and a novel objective function—integrating cross-entropy and triplet loss—were introduced. It is postulated that the CBAM mechanism enhances the feature fusion, endowing the model with the capability to pinpoint essential features autonomously. Meanwhile, the innovative objective function facilitates the model in concurrently acknowledging the overarching spatial arrangement of book attributes and the

relative distances between them, likely leading to augmented recommendation accuracy.

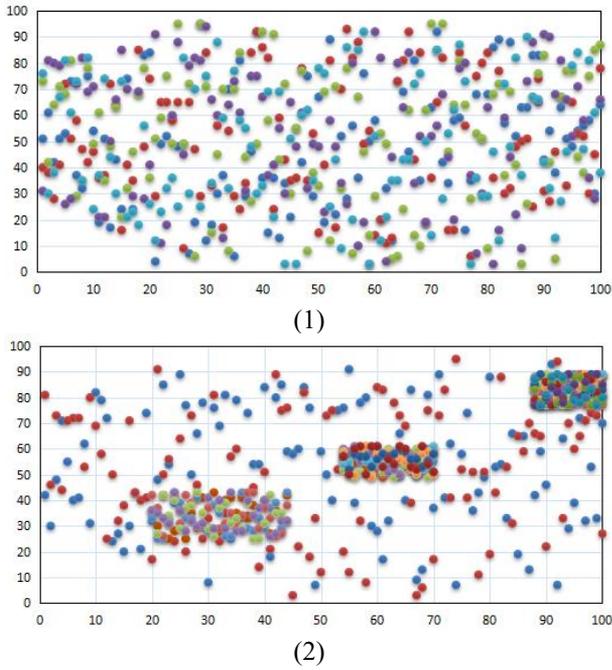


Figure 6. Synergy and visualization of book sample attributes

An examination of Figure 7 reveals the pronounced influence of varied fusion weights on the loss function during the model's training trajectory. The most pronounced descent in loss is identified in single-modal instances. As training advances, such a decrement is seen to decelerate. It can be posited that this subdued decline might stem from feature extraction restricted to a solitary modality, confining the information's representation in terms of its dimensionality and depth. In the context of $\beta=0.2$, a loss diminution rate is discerned that markedly overshadows that of the single-modal instance. Even in advanced training phases, a diminished loss is persistently registered. This observation may suggest that at such fusion weights, the model's capacity for learning and optimization is heightened. During the $\beta=0.3$ phase, a particularly swift loss decrement is recorded, arriving at an insignificant value in the incipient phases of training and sustaining that plateau. Such brisk convergence might allude to the onset of overfitting—a situation that might necessitate

rigorous regularization techniques or calibration of learning rates. In the $\beta=0.5$ phase, trends reminiscent of $\beta=0.2$ are detected, albeit with a more gradual decrement and less efficient loss mitigation.

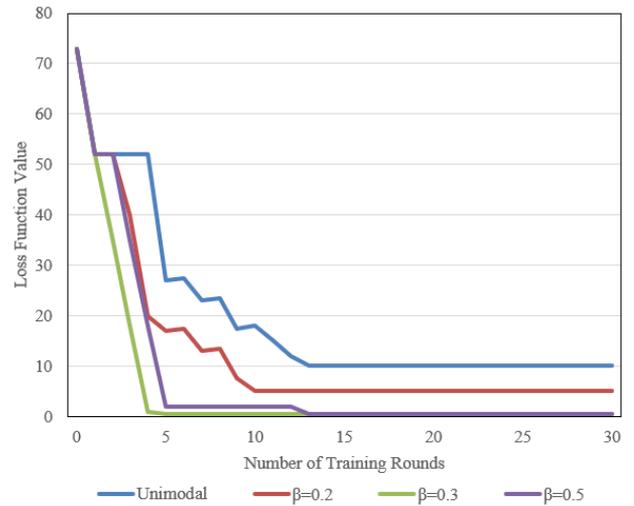


Figure 7. Dynamic alterations in model loss function under distinct fusion weights

Table 2 provides a comparative analysis of diverse 'Image + Text' amalgamated recommendation models, benchmarked against various performance metrics. These evaluation parameters span $R@1$, $R@5$, and $R@10$ for both text and image recommendations, alongside the overarching mR metric. It is noted that the proffered methodology exhibits a superior standing across all evaluative dimensions. In particular, a pronounced lead in the text recommendation's $R@5$ metric relative to competing models is evident. Parallely, for image recommendation's $R@1$ and $R@5$, superior outcomes are documented for the proposed approach. The aggregate mR metric emerges as the most elevated, underscoring the method's distinguished efficacy in addressing multi-modal book recommendation quandaries. Alternative methodologies such as CVLE, MFAE, MDBN, JRL, MVAE, CFA, BAE, MRBM, and MDBM are consistently observed to trail the suggested approach across diverse metrics. Of note are the MFAE in the realm of image recommendation's $R@1$ and BAE in text recommendation's $R@1$, both of which manifest considerably inferior outcomes compared to other methodologies.

Table 2. Performance metrics of 'Image + Text' converged recommendation models

Model	Text Recommendation			Image Recommendation			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
CVLE	43.4	64.9	83.5	31.4	53.2	79.2	57.2
MFAE	32.4	63.9	79.3	22.5	54.6	68.2	54.2
MDBN	43.2	74.7	86.3	33.9	63.2	77.5	57.9
JRL	41.3	72.1	82.7	32.5	62.5	73.5	59.3
MVAE	42.9	70.2	86.2	32.6	68.3	73.6	57.3
CFA	42.8	78.3	85.7	30.5	63.2	73.6	62.6
BAE	38.4	69.3	72.2	21.4	66.3	64.1	63.6
MRBM	41.7	71.4	84.6	33.5	61.3	74.5	65.4
MDBM	41.1	84.3	83.5	39.5	65.2	73.5	66.4
The Proposed Method	52.3	93.5	88.4	42.6	72.4	80.3	67.1

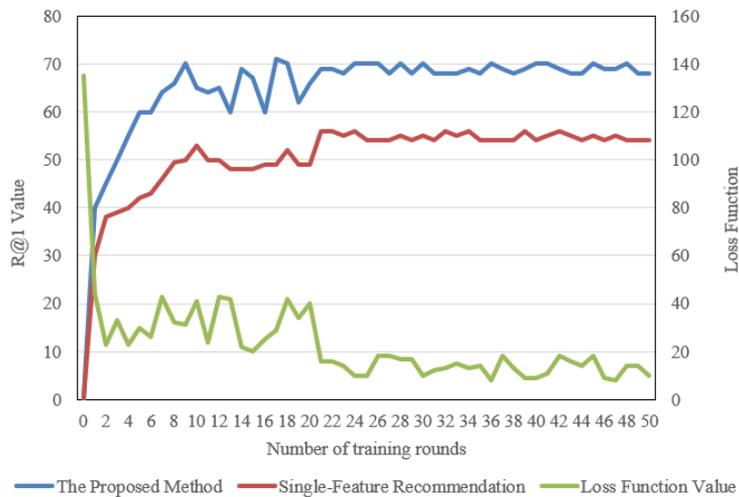


Figure 8. Trajectories of performance across multi-modal and single-modal recommendation models

Table 3. Comparative recommendation efficacy across diverse book categories

Requirements	Sensitivity	Specificity	PPV	IoU	Dice
Education and Learning	0.9937	0.9937	0.8489	0.8493	0.9239
Professional and Career Development	0.9742	0.9927	0.8994	0.8039	0.9292
Science and Technology	0.9937	0.9739	0.8773	0.8038	0.9193
Health and Lifestyle	0.9028	0.9082	0.8937	0.8304	0.9019
Literature and Art	0.9926	0.9048	0.8739	0.8830	0.9914
History and Culture	0.9074	0.9894	0.8094	0.8843	0.9187
Entertainment and Leisure	0.9729	0.9783	0.8289	0.8370	0.9018
Personal Growth and Psychology	0.9047	0.9093	0.8439	0.8034	0.9198

From Figure 8, it can be gleaned that as the training rounds multiply, performance enhancements of the multi-modal recommendation model considerably eclipse that of its single-modal counterpart. Initial phases register brisk performance upticks for both models. Yet, as training rounds proliferate, a more pronounced ascension in the performance curve of the multi-modal model is evidenced, thereby overshadowing the single-modal model across most training intervals. This observed trend aligns with the diminishing loss function values for the multi-modal model during its training, alluding to systematic error diminution and model refinement. On the other side, the single-modal model's performance uptrend appears relatively muted. Such constraints might be rooted in its singular feature reliance for recommendations, thereby failing to leverage the depth of multi-modal data and potentially limiting its performance horizon. Hence, it can be deduced that when commissioned with book recommendations, the multi-modal model showcases heightened efficacy. This superior performance is surmised to stem from its intrinsic ability to amalgamate both textual and visual information from books, enabling a more intricate and precise feature extraction, which consequently amplifies recommendation precision.

Book requirements in this investigation were demarcated into eight distinct clusters: 1) Education and Learning, embracing textbooks, reference materials, pedagogical resources, and language acquisition literature; 2) Professional and Career Development, spanning career manuals, skill enhancement literature, sectoral reports, and leadership primers; 3) Science and Technology, encapsulating popular science editions, academic research, technical manuals, and literature in programming and engineering; 4) Health and

Lifestyle, covering wellness guides, culinary literature, personal well-being, and domestic management; 5) Literature and Art, inclusive of fiction, poetry, theatrical scripts, literary assessments, and artistic discourse; 6) History and Culture, comprising historical editions, cultural analyses, humanities literature, geographical resources, and travel narratives; 7) Entertainment and Leisure, featuring illustrated literature, graphic novels, cerebral games, and recreational guides; 8) Personal Growth and Psychology, with titles on self-enhancement, psychological well-being, life philosophies, and emotional equilibrium.

An appraisal of Table 3 indicates that across disparate book requirements, the recommendation strategy delineated in this study rendered consistently elevated levels on pivotal performance metrics—Sensitivity, Specificity, Positive Predictive Value (PPV), Intersection over Union (IoU), and Dice Coefficient. In aggregate terms, this recommendation framework evidenced a commendable efficacy across varied book categories, with pronounced excellence observed within domains like Education and Learning, Professional and Career Development, and Literature and Art. Such sterling performances can be associated with the model's adeptness at accentuating salient book features within these classifications, underpinning precise recognition and recommendation. Conversely, in categories like Health and Lifestyle and Entertainment and Leisure, a modicum of inefficacy in recommendations was discerned, potentially attributable to the nuanced complexity or feature ambiguity within these domains. In summation, the advanced method signals distinct advantages in addressing a wide spectrum of book recommendation requisites, emphasizing its versatility and widespread relevance.

6. CONCLUSION

From the synthesized data, it can be discerned that the multi-modal recommendation model delineated in this investigation manifested superior performance, effectively establishing its merit across an array of application landscapes and evaluative metrics, most notably in the nuanced realm of book recommendations.

Initially, a methodology was elucidated that amalgamated product visual cues with textual descriptors, seamlessly transmuting disorganized and disparate data instances into environments wherein the preponderance of analogous feature vectors were observed to be closely aligned. Such a transformation was found not only to fortify the model's delineative capacity and resilience but also to validate, albeit indirectly, the effectiveness of the feature extraction techniques proposed herein.

Moreover, through judicious modulation of the fusion weight, denoted as β , it was revealed that the model's loss function trajectory could be optimally regulated, setting the stage for improved training ramifications. The calibration of β was perceived to necessitate nuanced tailoring, contingent upon the specific operational context and anticipated model outcomes.

When juxtaposed with other integrated "image+text" recommendation algorithms, the model delineated here consistently showcased preeminent performance metrics, both in terms of recommendation precision and diversity. Such outcomes suggest an inherent superiority in harmonizing visual and textual data streams. Across varied book demand spectrums, the recommendation efficacy of this strategy was notably salient, particularly in segments characterized by pronounced attributes such as "Education and Learning", "Professional and Career Development", and "Literature and Art". High recommendation fidelity and dependability were consistently observed. Concomitantly, a degree of resilience was also evident in this model's engagement with domains characterized by intricate and nebulous attributes.

In aggregate, both from a theoretical standpoint and practical execution, the multi-modal recommendation model elucidated in this research was evidenced to register stellar performance, emphasizing its potential and relevance, especially in the domain of book recommendations.

ACKNOWLEDGMENT

This paper was funded by Science and Technology Project of Hebei Education Department (Grant No.: ZD2021415).

REFERENCES

- [1] Samih, A., Ghadi, A., Fennan, A. (2022). Deep graph embeddings for content based-book recommendations. In *International Conference on Big Data and Internet of Things*, Cham: Springer International Publishing, pp. 105-115. https://doi.org/10.1007/978-3-031-28387-1_10
- [2] Moore, D., Petrovic, A., Bailey, C., Bodily, P. (2022). Composition of short stories using book recommendations. In *2022 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA, IEEE, pp. 1-5. <https://doi.org/10.1109/IETC54973.2022.9796781>
- [3] Bogaards, N., Schut, F. (2021). Content-based book recommendations: Personalised and explainable recommendations without the cold-start problem. In *Proceedings of the 15th ACM Conference on Recommender Systems*, Amsterdam Netherlands, pp. 545-547. <https://doi.org/10.1145/3460231.3474603>
- [4] Gao, S., Ng, Y.K. (2021). Analyzing the preferences and personal needs of teenage readers to make book recommendations. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Melbourne VIC Australia, pp. 463-468. <https://doi.org/10.1145/3486622.3493972>
- [5] Milton, A., Green, M., Keener, A., Ames, J., Ekstrand, M.D., Pera, M.S. (2019). StoryTime: Eliciting preferences from children for book recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, Copenhagen Denmark, pp. 544-545. <https://doi.org/10.1145/3298689.3347048>
- [6] Abrams, M., Gessler, L., Marge, M. (2019). Rex: A dialogue agent for book recommendations. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, pp. 418-421. <https://doi.org/10.18653/v1/W19-5948>
- [7] Tashkandi, A., Wiese, L., Baum, M. (2017). Comparative evaluation for recommender systems for book recommendations. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband. Lecture Notes in Informatics (LNI), Proceedings-Series of the Gesellschaft für Informatik (GI)*, Bonn, pp. 291-300.
- [8] Alharthi, H., Inkpen, D., Szpakowicz, S. (2018). Authorship identification for literary book recommendations. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 390-400.
- [9] Aggarwal, N., Asooja, K., Jha, J., Buitelaar, P. (2015). Top-N books recommendation using Wikipedia. In *ISWC (Posters & Demos)*.
- [10] Pera, M.S., Ng, Y.K. (2014). Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, Foster City, Silicon Valley California USA, pp. 9-16. <https://doi.org/10.1145/2645710.2645721>
- [11] Yang, W., Shi, X. (2022). Deep multi-mode learning for book spine recognition. In *International Conference on Web Information Systems and Applications*, Dalian, China, Cham: Springer International Publishing, pp. 416-423. https://doi.org/10.1007/978-3-031-20309-1_36
- [12] Rasheed, A., Umar, A.I., Shirazi, S.H., Khan, Z., Shahzad, M. (2023). Cover-based multiple book genre recognition using an improved multimodal network. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(1): 65-88. <https://doi.org/10.1007/s10032-022-00413-8>
- [13] Lee, S., Kim, J., Park, E. (2023). Can book covers help predict bestsellers using machine learning approaches? *Telematics and Informatics*, 78: 101948. <https://doi.org/10.1016/j.tele.2023.101948>
- [14] Lemos, J., Finn, E. (2019). Babel VR: Multimodal virtual reality environment for shelf browsing and book discovery. In *HCI International 2019-Late Breaking Posters: 21st HCI International Conference, HCII 2019*, Orlando, FL, USA, Springer International Publishing, pp.

- 30-38. https://doi.org/10.1007/978-3-030-30712-7_5
- [15] Oi, M., Yamada, M., Okubo, F., Shimada, A., Ogata, H. (2017). Finding traces of high and low achievers by analyzing undergraduates' e-book logs. *CEUR Workshop Proceedings*, 1828: 15-22.
- [16] Mashfufah, A., Nurkamto, J., Novenda, I.L. (2019). Conceptual: Digital book in the era of digital learning approaches (DLA). In *IOP Conference Series: Earth and Environmental Science*, 243(1): 012107. <https://doi.org/10.1088/1755-1315/243/1/012107>
- [17] Xiong, J., Yin, H., Pan, M. (2023). Book recommendation and purchase of intelligent image recognition technology under the background of 5G environment. *Journal of Computational Methods in Sciences and Engineering*, (Preprint), 23(2): 995-1005. <https://doi.org/10.3233/JCM226469>
- [18] Fu, Q., Fu, J., Wang, D. (2022). Deep learning and data mining for book recommendation: retrospect and expectation. In *2022 14th International Conference on Computer Research and Development (ICCRD)*, Shenzhen, China, IEEE, pp. 60-64. <https://doi.org/10.1109/ICCRD54409.2022.9730317>
- [19] Yang, C.C., Akçapinar, G., Flanagan, B., Ogata, H. (2019). Developing e-book page ranking model for pre-class reading recommendation. In *Proceeding of 27th International Conference on Computer in Education (ICCE 2019)*, Taiwan, pp. 360-362.
- [20] Paul, A., Wu, Z., Liu, K., Gong, S. (2022). Personalized recommendation: From clothing to academic. *Multimedia Tools and Applications*, 81(10): 14573-14588. <https://doi.org/10.1007/s11042-022-12259-7>
- [21] Oleksiv, N., Veres, O., Vasyliuk, A., Rishnyak, I., Chyrun, L. (2022). Recommendation system for monitoring the energy value of consumer food products based on machine learning. *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems*, Gliwice, Poland, pp. 1321-1350.
- [22] Yang, X., Shyu, M.L., Yu, H.Q., Sun, S.M., Yin, N.S., Chen, W. (2018). Integrating image and textual information in human-robot interactions for children with autism spectrum disorder. *IEEE Transactions on Multimedia*, 21(3): 746-759. <https://doi.org/10.1109/TMM.2018.2865828>