IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 1, February, 2024, pp. 459-467
Journal homepage: http://iieta.org/journals/ts

Deep Learning-Based Educational Image Content Understanding and Personalized Learning Path Recommendation

Guoli Xu[1,2*], Cora Un In Wong[1]

[1] Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macao 999078, China
[2] Education College, Fuzhou University of International Studies and Trade, Fujian 350202, China

Corresponding Author Email: p2209717@mpu.edu.mo

https://doi.org/10.18280/ts.410140

Received: 12 September 2023
Revised: 26 December 2023
Accepted: 10 January 2024
Available online: 29 February 2024

Keywords:
*deep learning, educational image content understanding, personalized learning pathways, Bidirectional Encoder Representations from Transformer (BERT) models, attention mechanisms, hierarchical Long Short-Term Memory (LSTM) models, multi-*feature *Latent Dirichlet Allocation (LDA) recommendation models*

ABSTRACT

With the breakthroughs in deep learning technology in image processing and language models, its potential application in the educational domain is gradually being unlocked. Particularly, in the understanding and analysis of educational image content, deep learning paves a new path for recommending personalized learning trajectories. This study aims to construct a system that interprets educational image content using deep learning technology and recommends personalized learning paths based on this content. Initially, an end-to-end visual narrative framework that integrates the Bidirectional Encoder Representations from Transformer (BERT) model, attention mechanisms, and hierarchical Long Short-Term Memory (LSTM) models is proposed to enhance the depth of understanding of educational image content. Subsequently, a recommendation model based on multi-feature Latent Dirichlet Allocation (LDA) is developed, facilitating the learning of correspondences among various features across different educational images, thereby promoting accurate recommendations of personalized learning paths. Existing research commonly overlooks the comprehensive consideration of semantic layers of images and educational backgrounds; this method is designed to bridge that gap. Results indicate that the system is capable of effectively understanding educational image content and providing precise learning path recommendations based on learner characteristics, promising to significantly improve learning efficiency and quality.

1. INTRODUCTION

With the rapid development of artificial intelligence (AI) technologies, particularly the notable achievements in deep learning within the realms of image recognition and natural language processing, an increasing number of studies have been dedicated to exploring these technologies' applications in the educational sector [1-4]. Educational image content, as a significant learning resource, has been recognized for its potential in providing new possibilities for the recommendation of personalized learning paths [5, 6]. End-to-end deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been demonstrated to possess substantial potential in understanding visual content and generating descriptive narratives [7]. This lays the technical foundation for establishing a system capable of accurately interpreting educational image content and, based on this, recommending personalized learning paths.

Personalized learning, one of the core objectives of educational technology development, aims to customize appropriate teaching strategies and content for each learner [8, 9]. The extraction of information from educational images using deep learning models not only deepens the understanding of the content of learning materials but also allows for the recommendation of the most suitable learning paths based on the specific needs and learning statuses of learners [10-12]. The implementation of this approach is expected to realize truly personalized teaching, assisting learners in studying in the manner most suited to them, thereby improving learning efficiency and effectiveness.

However, existing research often exhibits limitations in the depth of image understanding and the degree of personalization of recommendation systems. Traditional models tend to overlook the semantic layers and educational backgrounds of complex educational images, leading to imprecise understanding and recommendations [13-15]. Moreover, existing recommendation models still lack detail in capturing learner characteristics and learning preferences, making it challenging to provide genuinely personalized learning paths [16-18].

The main contribution of this study lies in the development of a novel system for the understanding of educational image content and the recommendation of personalized learning paths based on deep learning. Firstly, an end-to-end visual narrative framework, incorporating the BERT model, attention mechanisms, and hierarchical LSTM models, is proposed to enhance the depth and accuracy of understanding educational image content. Secondly, by constructing a multi-feature LDA recommendation model based on the content of educational

images and introducing several improved LDA models, it is possible to learn and understand the relationships among various features across different educational images, offering learners more accurate and personalized learning paths. These studies not only advance the development of educational technology but also provide new perspectives and practical experiences for the application of deep learning in education.

2. UNDERSTANDING AND GENERATION OF EDUCATIONAL IMAGE CONTENT BASED ON END-TO-END VISUAL NARRATION

In the research conducted, a novel end-to-end visual narrative approach for the understanding and generation of educational image content is introduced. This methodology encompasses a framework consisting of three components: CNN, BERT, and hierarchical LSTM (hLSTM) networks, with the specific architecture depicted in Figure 1. Through multiple layers of convolution and pooling operations, the CNN module is capable of abstracting higher-level semantic features from basic pixel-level characteristics. These detailed visual features are crucial for comprehending the educational information inherent in images, serving as key bridges linking image content with educational contexts. The BERT model is employed to process and understand descriptive texts associated with image sequences, such as explanatory text or chart data adjacent to images. The incorporation of BERT allows the system not only to focus on the surface meaning of individual words but also to capture the deep semantic connections between words within sentences, thereby providing accurate semantic understanding for subsequent text generation and personalized recommendation. hLSTMs are responsible for integrating the visual and linguistic features extracted by the previous two parts, learning their mapping relationships, and generating coherent, meaningful descriptions of educational image content. These descriptions not only enrich the understanding of the educational content of images but also provide important contextual information for establishing effective personalized learning paths.

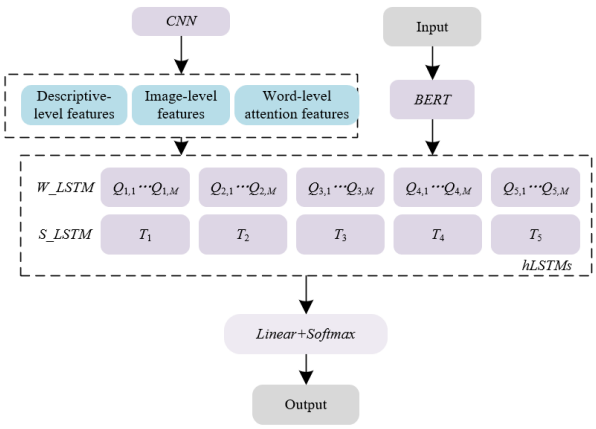

Figure 1. Structure of the proposed method

2.1 Extraction of image features

In this study, the extraction of educational image features is facilitated through the deployment of the Visual Geometry Group 16 (VGG16) network within the CNN component.

(a) Feature extraction across the entire image sequence is achieved by utilizing the final convolutional layer of VGG16. Each image is encoded into a collection of high-dimensional feature vectors. Such representation captures crucial information within images, including textures, shapes, and edges, providing detailed visual context for every frame within the narrative of educational image content descriptions. By incorporating an additional convolutional layer and adjusting the output dimensions to align with the size of the LSTM hidden layers, a smoother transition of visual information is ensured. This facilitates the initialization of subsequent S_LSTM and W_LSTM layers, ensuring seamless integration of image features with the text generation module.

(b) The extraction of features from individual images involves the utilization of feature vectors generated by VGG16 to drive the corresponding sentence-level S_LSTM. This enables the model to generate a text description for each specific image. This step ensures that the generation of descriptions is image-driven, with the unique features of each image reflected in the generated narrative, thus making the image-to-text conversion more accurate and personalized. This is critical for the understanding of educational image content, as it allows the educational content of each image to be explicitly recognized and articulated.

(c) The extraction of word attention visual features is accomplished through an attention mechanism, which allows the model to dynamically focus on specific parts of a single image feature when generating the description of each word. This mechanism enables the W_LSTM to generate more detailed and relevant descriptions, as it can selectively focus on specific areas of the image based on the semantic requirements of the current word. In the context of understanding educational image content and recommending learning paths, this means that the textual descriptions can more precisely reflect the key educational information of the images.

2.2 Text encoding

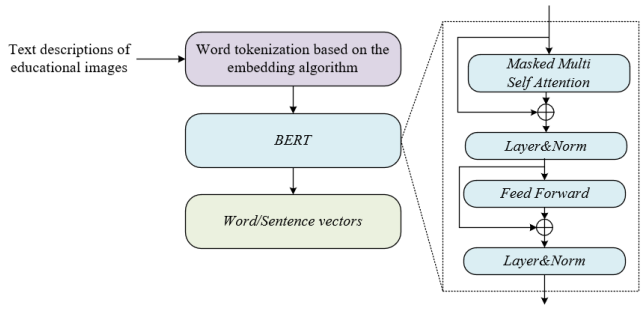

Figure 2. Text encoding process flowchart

In this study, the application of the BERT model is employed to obtain high-quality linguistic representations, thereby enhancing the text component of educational image content understanding and personalized learning path recommendations. Specifically, by constructing a dictionary encompassing all sentences within the dataset and a lexicon of all words, initial high-dimensional vector representations for each sentence and word are generated using the pre-trained BERT model. The educational image content description generated includes V images and their corresponding sentences $b=\{b_1,b_2,\ldots,b_V\}$. The number of words in each sentence is denoted by b_u. The sentence vectors for the generated educational image content description are denoted

by t, and the word vectors are represented by q. These vectors capture the semantic information and contextual relationships of words and sentences, providing rich features for the text. Based on this, the sentence and word vectors are fine-tuned through subsequent model training to ensure better alignment with specific educational content and learning paths. The fine-tuned vectors are then fed into the hierarchical LSTM module, with the dimension of the LSTM hidden layers denoted by F. Figure 2 presents the text encoding process flowchart.

2.3 Generation of descriptions for educational image content

Within the hLSTMs architecture proposed in this study, the S_LSTM layer functions to understand and generate sentence-level semantic information. The input to this layer comprises sentence vectors extracted by the BERT model, encoding the deep semantic context of text related to educational images. Utilizing these sentence vectors, the S_LSTM generates a sequence of sentence descriptions, each corresponding to a set of images or a single image's instructional content. This LSTM layer is designed to capture the sequential and intrinsic logical relationships between sentences, enabling the generated text to coherently describe the image sequence, thereby providing a structured and semantically rich foundation for subsequent word-level descriptions. Let the input gate be represented by u^s, the forget gate by d^s, the output gate by p^s, the memory state by z^s, and the information updating the memory state z^s by w^s. The sigmoid activation function is denoted by δ, the element-wise multiplication by $\odot$, the hyperbolic tangent function by *tanh*, the learnable training parameters by $Q_{(*)}$ and $y_{(*)}$, and the hidden layer state at time s by g^s. The expression for the LSTM structure with an attention mechanism is given as follows:

$$\begin{aligned} u^s &= \delta\left(Q_{au}a^s + Q_{gu}g^{s-1} + Q_{cu}c^s + n_u\right) \\ d^s &= \delta\left(Q_{ad}a^s + Q_{gd}g^{s-1} + Q_{cd}c^s + n_d\right) \\ p^s &= \delta\left(Q_{ap}a^s + Q_{gp}g^{s-1} + Q_{cw}c^s + n_w\right) \\ w^s &= \tanh\left(Q_{aw}a^s + Q_{gw}g^{s-1} + Q_{cw}c^s + n_w\right) \\ z^s &= d^s \Phi z^{s-1} + u^s \Phi w^s \\ g^s &= p^s \Phi \tanh\left(z^s\right) \end{aligned} \tag{1}$$

g^s is then input into the LSTM unit of the next moment. Let the current textual feature be denoted by a^s, and the current dynamic image feature by c^s, resulting in the expression:

$$g^s, z^s = \mathrm{LSTM}\left(\left[a^s, c^s\right], g^{s-1}, z^{s-1}\right) \tag{2}$$

Let the memory state at time s be denoted by z^s, the image features of the entire educational image content description by a, with a after passing through a fully connected network to yield the feature initializing the hidden layer state represented by $g^0{}_t$, the learnable parameters by $Q^0{}_t$, the visual features at time s by $a^s{}_t$, the averaging function by Θ, and the feature of the s-th image extracted from the entire educational image content description image features a by Ω. The current sentence vector is represented by t^s, the current visual feature by $a^s{}_t$, the previous moment's hidden layer state by $g^{s-1}{}_t$, and the corresponding memory state by $z^{s-1}{}_t$, with the current moment's hidden layer state denoted by $g^s{}_t$, and the corresponding memory state by $z^s{}_t$. The expression for S_LSTM is given as:

$$\begin{aligned} g_t^0 &= Q_t^0 \theta(a) \\ a_t^s &= \Omega(a) \\ g_t^s, z_t^s &= \mathrm{LSTM}_{SE}\left(\left[t^s, a_t^s\right], g_t^{s-1}, z_t^{s-1}\right) \end{aligned} \tag{3}$$

The features inputted into the S_LSTM layer consist of contextual semantic features, which are composed of t^s and $a^s{}_t$. Thus, by feeding contextual semantic information along with $g^{s-1}{}_t$ and $z^{s-1}{}_t$ into the S_LSTM layer, the outputs $g^s{}_t$ and $z^s{}_t$ are obtained.

The W_LSTM layer is tasked with further refining and generating word-level semantic information by delving deeper into the sentence-level descriptions outputted by the S_LSTM layer. At each timestep, the W_LSTM layer receives the previously generated word, the current sentence-level semantic features, and the visual features of the image, all dynamically integrated through an attention mechanism. The aim of this layer is to generate accurate and detailed words, forming an educational description closely corresponding to the image content. Suppose the feature of the s-th image in the entire educational image content description is denoted by $a^s{}_t$, the length of the generated sentence by j, and a fully connected function by Θ. When predicting the j-th word of the s-th sentence, the attention weight of the k-th region of the s-th image is represented by $x^k{}_{s,\,k}$, and the word-level attention feature by $z^j{}_s$. The expression for the W_LSTM layer is as follows:

$$\begin{aligned} g_s^0 &= Q_s^0 \theta\left(a_t^s\right) \\ x_{s,k}^j &= \mathrm{softmax}\left(\Theta\left(g_s^{j-1}, a_t^s\right)\right) \\ c_s^j &= \sum_{k=1}^{M}\left(x_{s,k}^j, a_{s,k}\right) \\ g_s^j, z_s^j &= \mathrm{LSTM}_{WO}\left(\left[gt_s^s, z_s^j\right], g_s^{j-1}, z_s^{j-1}\right)(j=1) \\ g_s^j, z_s^j &= \mathrm{LSTM}_{WO}\left(\left[gt_s^{j-1}, z_s^j\right], g_s^{j-1}, z_s^{j-1}\right)(j>1) \end{aligned} \tag{4}$$

Assume the contextual semantic information at the j-th moment of the s-th sentence is denoted by $n^j{}_s$. Within the W_LSTM layer, $n^j{}_s$ maintains a dimension size consistent with the LSTM hidden layer state. When j=1, $n^j{}_s$ is the $g^s{}_t$ outputted by the S_LSTM; for j>1, $n^j{}_s$ is the word feature $q^{j-1}{}_s$.

The ultimate goal of this method is to automatically generate coherent, logical, and narrative descriptions from a given set of educational images. By maximizing the generation probability of the entire narrative text, the model sequentially generates a sentence containing M words for each image, where the generation probability of each sentence is the sum of the joint probabilities of its internal words. Suppose the j-th word of the u-th sentence generated by the model is represented by $q_{u,j}$, and the length of the sentence by M. The probability calculation formula is as follows:

$$\begin{aligned} O\left(b_u \mid a, \phi\right) &= \prod_{j=1}^{M} O\left(q_{u,j} \mid a, q_{u,1}, q_{u,2}, \cdots q_{u,j-1}\right) \\ O\left(b \mid a, \phi\right) &= \sum\nolimits_{u=1}^{V}\left(b_u \mid a, \phi\right) \end{aligned} \tag{5}$$

The model is trained using cross-entropy loss, resulting in the expression:

$$M = -\sum_{u=1}^{V}\sum_{k=1}^{M}\log O\left(q_{u,k} \mid n, q_{u,1}, q_{u,2}, \cdots, q_{u,k-1}\right) \quad (6)$$

3. CONSTRUCTION OF A MULTI-FEATURE LDA RECOMMENDATION MODEL BASED ON EDUCATIONAL IMAGE CONTENT

In the context of educational image content understanding and personalized learning path recommendation, a pivotal challenge is the accurate extraction of instructional content from images, transforming it into structured knowledge conducive to learner comprehension and absorption. Text descriptions generated are required to reflect key information within images, closely aligning with educational objectives. Furthermore, the recommendation of the most suitable learning content based on the individual needs and learning history of learners is considered. To address these challenges, a multi-feature LDA recommendation model is proposed for personalized learning path recommendations, with the model architecture illustrated in Figure 3. This model integrates educational image text descriptions with other relevant features (e.g., image features, user interaction data) to uncover latent topic distributions. Such an approach not only facilitates a deeper understanding of educational content but also reveals hidden relationships between learning materials, providing a more accurate basis for personalized recommendations.

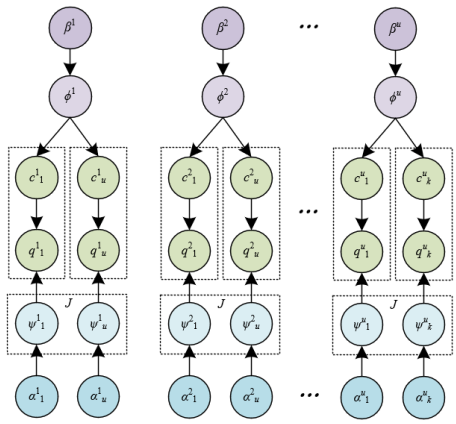

Figure 3. Architecture of the multi-feature fusion LDA recommendation model

The generation process of the proposed multi-feature LDA recommendation model based on educational image content is realized through the following steps:

(a) Visual feature topic modeling. Initially, the model models the corresponding topics, j^u, for each visual feature. These visual features may include intrinsic image attributes such as color, texture, shape, etc. The topic areas, e, for each visual feature are governed by a Dirichlet distribution, which describes the probability distribution across multiple categories (i.e., topics). In this phase, for each visual feature, the model samples from a Dirichlet distribution to determine the distribution proportion of each feature, D^u, across different topics, i.e., the feature's mixture components, ϕ_{fl}.

(b) Image topic mixture selection. For each image, f_l, within the educational image collection, the model selects a topic mixture, $\psi^{e,u}{}_{cj}$, consisting of J_u topics based on a Dirichlet distribution. This step establishes which topics are associated with the image collection and their respective association strengths.

(c) Visual feature generation. For each image, f_l, in the training library, the model generates the visual features, d_u. This process involves two steps: initially, based on the obtained topic mixture distribution, the model employs a multinomial distribution to generate a topic, $c^u{}_{l,v} \sim Multinomial(\phi^u{}_l)$; subsequently, based on the generated topic, $c^{ue}{}_{l,v}$, the model further utilizes a multinomial distribution to generate the visual vocabulary related to that topic, $d^{ue}{}_{l,v}$, i.e., image features, $d^{ue}{}_{l,v} \sim Multinomial(\psi^{e,u} c^{ue}{}_{l,v})$.

The proposed multi-feature LDA model for educational image content employs the Dirichlet distribution as a prior to simulate the distribution of latent topics within an image collection containing educational content. A multinomial distribution is utilized to generate specific visual feature vocabularies, thereby calculating the joint distribution of topics and vocabularies. Given the Dirichlet distribution hyperparameters β and α, the formula for calculating the joint distribution of hidden topics and vocabularies is provided as follows:

$$O(q,c \mid \beta,\alpha) = O(c \mid \beta) O(q \mid c,\alpha) \quad (7)$$

Assuming the number of vocabularies assigned to topic j by f_l in the image collection is represented by $v_{l,j}$, the calculation formula for $O(q,c|\beta,\alpha)$ is as follows:

$$O(c \mid \beta) = \frac{\Gamma(J\beta)}{\Gamma(\beta)^j} \prod_l \frac{\prod_j \Gamma(v_{l,j} + \beta)}{\Gamma(V_l + J\beta)} \quad (8)$$

Similarly, assuming the number of vocabularies assigned to topic j by vocabularies q_n is represented by v_{jn}, the calculation formula for $O(q|c,\alpha)$ is:

$$O(q \mid c,\alpha) = \prod_e \left(\frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \right)^J \prod_j \frac{\prod_q \Gamma(v_{jn} + \alpha)}{\Gamma(V_j + N\alpha)} \quad (9)$$

To ensure feasibility in high-dimensional space sampling and to guarantee the accuracy and stability of the final parameter estimation, the Gibbs sampling method is adopted for estimating the parameters of the multi-feature LDA model, ϕ_{fl} and ψ_{cj}. Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method, gradually approximates the true parameters of complex distributions through an iterative process. This is particularly beneficial for handling the high-dimensional features in images, as it effectively draws samples from the joint distribution for estimation. Moreover, Gibbs sampling does not require complex mathematical derivations to directly compute posterior distributions, making it more flexible and efficient in practical applications.

According to the learning process of the multi-feature LDA recommendation model, given an educational image training set, the latent topic inference for each image can be obtained through Gibbs sampling. Assuming each visual feature is represented by D_u, the current state by $c^u{}_k$, the current state's topic by $z_i c_u$, other topic assignments besides the current state by $e^e{}_{c\text{-}u}$, the number of vocabularies s assigned to topic j excluding the current state by $v^{(s)}{}_{j,\text{-}u}$, and the number of vocabularies in document l assigned to topic j excluding the current state by $v^{(s)}{}_{l,\text{-}u}$. Given all states except $c^u{}_k$, the probability that the v-th feature in region e of educational image f is related to topic j can be calculated as follows:

$$O\left(c_k^u = j_u \mid \overset{e}{c}_{\neg k}, \overset{e}{q}_u\right) = \frac{v_{j,\neg k}^{u,e(s)} + \alpha_s^{u,e}}{\sum_{u=1}^{N_u} v_{j,\neg u}^{u,e(s)} + \alpha_s^{u,e}} g \frac{v_{l,\neg u}^{u(j)} + \beta_j^u}{\sum_{k=1}^{J_u}\left(v_{l,\neg u}^{u(j)} + \beta_j^u\right) - 1} \tag{10}$$

In the methodology of implementing a multi-feature LDA model for educational image recommendation via Gibbs sampling, the process unfolds as follows:

(a) During the initialization phase, features within educational images are randomly assigned an initial topic through Gibbs sampling. This step is crucial for assigning a starting category label to each visual feature, albeit these labels are randomly distributed at the outset.

(b) In the iteration phase, Gibbs sampling is utilized to compute the posterior probability of each visual feature vocabulary based on current parameter estimates. Subsequently, topics are reassigned to each feature vocabulary based on this probability distribution. This equates to updating the topic assignment of the current feature vocabulary, conditional upon the topic assignments of all other feature vocabularies.

(c) Upon convergence, the Gibbs sampling process yields stable topic distributions. At this juncture, the topic distribution results of all visual feature vocabularies across the document set are aggregated to estimate model parameters, namely, the mixture components of features and the topic mixtures of documents. The specific formulas for estimating model parameters ϕ_{fl} and ψ_{cj} are as follows:

$$\phi_{l,j}^u = \frac{v_l^{u(j)} + \beta_j^u}{\sum_{j=1}^{J_u} v_l^{u(j)} + \beta_j^u} \tag{11}$$

$$\psi_{j,s}^{u,e} = \frac{v_l^{u,e(s)} + \alpha_s^{u,e}}{\sum_{s=1}^{N_u} v_j^{u,e(s)} + \alpha_s^{u,e}} \tag{12}$$

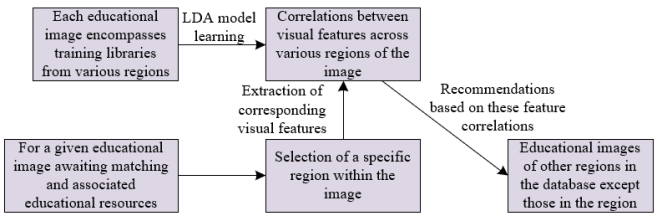

Figure 4. Flowchart for personalized learning resource recommendation based on the LDA model

The foundational principle of the proposed multi-feature LDA recommendation model lies in the extraction of features from a plethora of educational images via deep learning techniques, coupled with the identification and categorization of latent topics within image content through the LDA topic model. This approach facilitates the association of these visual features with corresponding pedagogical topics. Building upon this, the model identifies personalized needs and interests of students by analyzing their learning history, preferences, and performance. Utilizing the established correlations between visual features and pedagogical topics, the model recommends educational images that best match the individualized learning trajectories of students. Consequently, each student receives educational resources most suited to their current stage of learning and interests, thereby fostering a genuinely personalized learning experience, enhancing learning outcomes, and promoting effective knowledge absorption and application. The detailed recommendation process is depicted in Figure 4.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 5 presents a comparison of word probability test outcomes for varying beam sizes (1, 2, 3) in the description of educational images. It is observed that at a beam size of 1, the generation of descriptions appears to have higher probability peaks, as indicated by probabilities of 0.7 and 0.58 at word counts of 10 and 25, respectively. This suggests that with a beam size of 1, the model exhibits considerable confidence in the selection of specific vocabulary. When the beam size is increased to 2, the probability distribution of descriptive words becomes more dispersed than at a beam size of 1, lacking very high probability peaks, which reflects a more even distribution of probabilities. For a beam size of 3, neither exceptionally high peaks in the probability distribution are observed nor is a more dispersed trend evident compared to a beam of 2, indicating that further increasing the beam size does not significantly enhance the diversity or confidence in word selection. It can be concluded that smaller beam sizes tend to generate vocabularies with high confidence, albeit at the expense of diversity and creativity. A medium beam size offers a better balance between accuracy and diversity, ensuring that generated descriptions are both confident and varied. Larger beam sizes did not exhibit a distinct advantage in this context, as for specific educational image content, an excess of alternatives did not assist the model in identifying better descriptions. These observations validate the effectiveness of the proposed method for understanding and generating educational image content through an end-to-end visual narrative approach. The integration of BERT, attention mechanisms, and hierarchical LSTM allows for the maintenance and enhancement of the naturalness and diversity of descriptions without sacrificing accuracy. Particularly at a medium beam size, the model is capable of finding an appropriate balance, generating descriptions that are both precise and rich, thereby affirming the practicality and innovativeness of the proposed method.

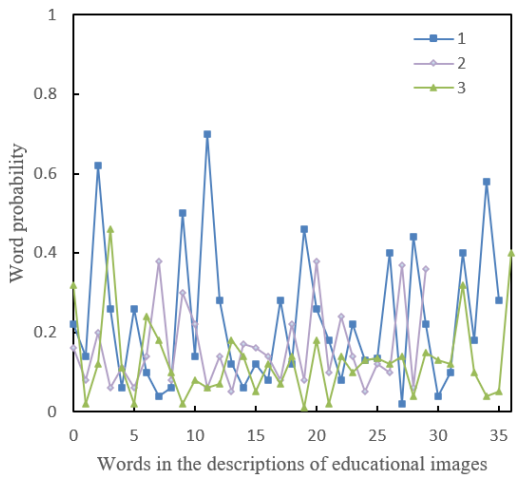

Figure 5. Comparative results of word probability tests for different beam sizes

Table 1. Performance comparison of different methods for understanding and generating educational image content

	Relevance	Coherence	Expressiveness
C-VAEs	26.3%	23.8%	18.9%
The proposed method	62.3%	65.8%	71.4%
Fleiss kappa coefficient	9.4%	8.8%	8.1%
P-value	0.0022	0.0002	0.0012
SAGAN	42.3%	41.2%	39.5%
The proposed method	45.3%	45.7%	51.2%
Fleiss kappa coefficient	11.2%	12.8%	11.6%
P-value	0.36	0.12	0.024

Table 1 presents a comparative analysis of different methods in the understanding and generation of educational image content, focusing on relevance, coherence, and expressiveness as the evaluation metrics. These metrics are utilized to assess the consistency of the generated textual descriptions with the educational image content, their logical coherence, and the richness and creativity of expression. It is observed that Conditional Variational Autoencoders (C-VAEs), as a generative model, are outperformed by the method described herein across all three metrics. The method achieves scores of 62.3%, 65.8%, and 71.4% in relevance, coherence, and expressiveness, respectively, compared to 26.3%, 23.8%, and 18.9% achieved by C-VAEs. The Fleiss kappa coefficient, a statistical measure for assessing the consistency among multiple raters, is relatively lower in the comparison with C-VAEs, indicating a less unanimous agreement among evaluators on these metrics. Nevertheless, significant p-value (0.0022, 0.0002, 0.0012) demonstrates a statistically significant difference, favoring the described method over C-VAEs with superior performance. In comparison with Self-Attention Generative Adversarial Networks (SAGAN), another generative model, the method shows improved performance across relevance, coherence, and expressiveness by 3%, 4.5%, and 11.7%, respectively. Higher Fleiss kappa coefficient (11.2%, 12.8%, 11.6%) in this comparison indicates better consistency among evaluators. Although p-value for relevance and coherence exceeds 0.05, indicating a lack of statistical significance, a p-value of 0.024 for expressiveness signifies a statistically significant difference, highlighting at least in terms of expressiveness, a significant improvement over SAGAN by the proposed method.

In summary, the method based on end-to-end visual storytelling demonstrated here significantly surpasses both C-VAEs and SAGAN across the evaluation metrics of relevance, coherence, and expressiveness. The expressiveness metric, in particular, shows significant enhancement relative to C-VAEs and SAGAN, underscoring the effectiveness of the algorithm in generating descriptions that are both creative and engaging. High Fleiss kappa coefficient and low p-value statistically support the efficacy of the proposed method.

Figure 6 illustrates the comparison of accuracies between the single-feature model and the multi-feature LDA model as the number of recommended resources increases. Within the range from 0 to 30 recommended resources, an enhancement in accuracies for both models is observed, suggesting that a higher number of recommended resources correlates with increased accuracy. Analysis of the single-feature model reveals an initial accuracy of 0.14, which gradually improves with the increment of recommended resources, peaking at 0.70 when the number of resources reaches 30. This model considers only one dimension of features, such as content of images or the difficulty level of educational resources alone. In contrast, the multi-feature LDA model, starting with an accuracy of 0.14 as well, demonstrates a faster rate of improvement and a higher final accuracy compared to the single-feature model. At 30 recommended resources, the accuracy of the multi-feature LDA model reaches 0.75. It is concluded that the multi-feature LDA model outperforms the single-feature model across all levels of recommended resources. This model integrates a variety of features, including content, structure, semantic information of images, user interactions, and feedback, thereby offering a more comprehensive and in-depth analysis. By amalgamating these diverse features, the multi-feature LDA model more accurately captures users' preferences and learning needs, providing more personalized and precise recommendations.

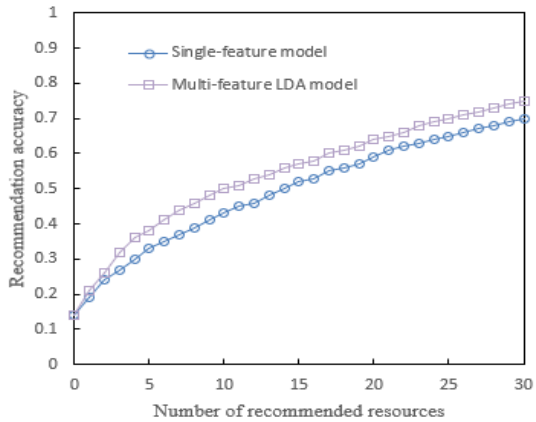

Figure 6. Comparison of recommendation accuracy between single-feature and multi-feature LDA recommendation models

Table 2. Performance of different educational image content recommendation algorithms in the personalized learning resource completion task

	MNIST		*CIFAR*-10		*MERLOT II*	
	Acc@4	*Acc*@10	*Acc*@4	*Acc*@10	*Acc*@4	*Acc*@10
Bi-LSTM	0.5326	0.3358	0.5214	0.3215	0.5248	0.3124
DMTL	0.5124	0.2789	0.4895	0.2789	0.4895	0.2587
DCRS	0.5269	0.3215	0.5123	0.3216	0.4756	0.2659
NTN-R	0.5268	0.3154	0.4789	0.2659	0.4895	0.2641
MBCF	0.4569	0.2369	0.5136	0.2845	0.4126	0.2123
FHN	0.6123	0.3789	0.6147	0.3784	0.5326	0.3159
Single-feature LDA recommendation model	0.6589	0.4251	0.6239	0.4126	0.5895	0.3895
Multi-feature LDA recommendation model	0.6589	0.4589	0.6258	0.4236	0.6123	0.4259

Table 3. Performance of various algorithms in visual feature topic modeling

	ImageNet	*COCO*	*PASCAL VOC*	*Flickr30k*	*CelebA*	*SUN*
Bi-LSTM	0.5685	0.4158	0.4258	0.5148	0.4897	0.3514
CSN	0.5412	0.4869	0.5126	0.5126	0.5216	0.3789
SCE-Net	0.5123	0.5124	0.5268	0.5269	0.5124	0.4216
Multi-feature LDA recommendation model	0.6523	0.5169	0.5789	0.6895	0.6784	0.5698

Table 2 presents the performance of various educational image content-based recommendation algorithms in the task of personalized learning resource completion, as measured by accuracies (Acc@4 and Acc@10). The algorithms evaluated include Bidirectional LSTM (Bi-LSTM), Deep Multi-Task Learning (DMTL), Distributed Collaborative Reference Services (DCRS), NTN-R, MBCF, FHN, the single-feature LDA recommendation model proposed in this study, and the multi-feature LDA recommendation model. It is observed that across the three datasets (MNIST, CIFAR-10, MERLOT II), the multi-feature LDA recommendation model either matches or exceeds the highest accuracies for both Acc@4 and Acc@10. This indicates that when considering the top four and top ten recommendations, the multi-feature LDA model consistently provides the most accurate recommendations. Consequently, it can be concluded that the multi-feature LDA recommendation model, which leverages educational image content, offers more precise and effective recommendations compared to other benchmark algorithms in the task of personalized learning resource completion, thereby enhancing the personalized learning experience.

Table 3 illustrates the performance of different algorithms in the domain of visual feature topic modeling, including Bi-LSTM, Channel Splitting Network (CSN), Similarity Condition Embedding Network (SCE-Net), and the multi-feature LDA recommendation model proposed in this study. Performance evaluations were conducted across a variety of image datasets, namely ImageNet, COCO, PASCAL VOC, Flickr30k, CelebA, and SUN. It is noted that the multi-feature LDA recommendation model achieved the highest performance metrics across all listed datasets. Particularly, on the Flickr30k and CelebA datasets, the performance advantage of the multi-feature LDA recommendation model over other methods were especially pronounced. Compared to traditional methods such as Bi-LSTM and other algorithms focused on visual features (e.g., CSN and SCE-Net), the multi-feature LDA recommendation model was found to be more effective in extracting and utilizing the features of images for visual content processing. These findings further validate the significant efficacy of the multi-feature LDA recommendation model, proposed in this study, in visual feature topic modeling. By integrating different types of visual features and applying topic modeling techniques, this model is capable of more accurately analyzing and recommending image content, thereby delivering superior performance in various image recognition and classification tasks.

Figure 7 demonstrates how the accuracy (Acc) of the multi-feature LDA recommendation model, proposed for the analysis of educational image content, varies with an increasing number of samples across different datasets, including ImageNet, COCO, PASCAL VOC, Flickr30k, CelebA, and SUN. It was observed that for all datasets, a notable enhancement in the model's accuracy was achieved as the number of samples increased. This indicates that the multi-feature LDA recommendation model is capable of effectively utilizing a larger dataset to improve its performance. In most datasets, the improvement in model performance tends to plateau after reaching a certain threshold of sample quantity. For instance, the accuracy on the ImageNet dataset increased by only 0.002 from 70 to 90 samples, suggesting that the marginal gains in performance diminish with the addition of more samples. Without the addition of extra samples (i.e., at 0 samples), the model already exhibited high accuracy across all datasets, particularly on ImageNet and COCO. This underscores the model's ability to maintain commendable performance even with limited data support. In summary, the multi-feature LDA recommendation model demonstrated enhanced performance with the addition of more samples, especially during the initial phase of sample inclusion. Although the performance gains decrease with further increases in sample quantity, the model still exhibited robust accuracy across different levels of sample volume. These results suggest that the multi-feature LDA recommendation model responds well to the addition of samples and is an effective approach, particularly suited for scenarios requiring the processing and analysis of large volumes of educational image content, capable of providing precise recommendations and optimization of teaching resources.

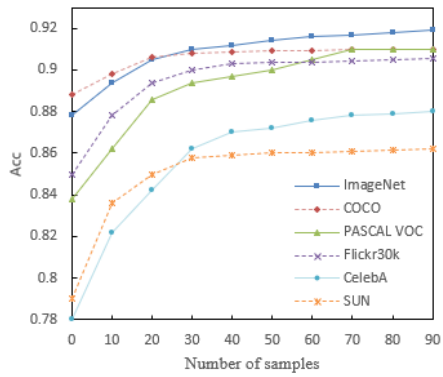

Figure 7. Performance of the proposed model across different sample sizes

5. CONCLUSION

In the research and experimental outcomes documented herein, significant advancements in the realm of educational image content comprehension and personalized learning pathway recommendation have been demonstrated. An end-to-end visual narrative framework has been developed through the successful integration of the BERT model, attention mechanisms, and hierarchical LSTM models. This framework, leveraging deep learning technologies, has enhanced the depth and accuracy of understanding educational image content, laying a solid foundation for subsequent personalized recommendations. By constructing a multi-feature LDA recommendation model based on educational image content and integrating various enhanced LDA models, an understanding and learning of multiple features among different educational images have been achieved. This

approach has revealed the complex relationships between different educational images, thereby providing learners with accurate and personalized learning pathways.

A series of experiments were conducted, comparing the probabilities of educational image descriptor words under different beam sizes, various educational image content comprehension and generation methods, as well as the performance of single-feature versus multi-feature LDA recommendation models. The results from these experiments have validated the effectiveness of the methods proposed in personalized learning resource recommendation and visual feature theme modeling tasks. An analysis on the performance of models with varying sample sizes was also conducted, indicating an enhancement in model performance with an increase in sample quantity, particularly notable during the initial stages of sample addition. This observation has underscored the sensitivity of the model to data volume and its excellent adaptability in big data environments.

In summary, through the innovative combination of advanced deep learning technologies, an efficient system for understanding educational image content and recommending personalized learning pathways has been developed. The experimental results have confirmed that this system is capable not only of a profound understanding of educational image content but also of providing personalized recommendations based on learners' needs, showcasing superior performance compared to traditional methods. Thus, the research conducted herein holds significant theoretical value for the academic community and possesses potential practical applications in educational practice, especially in designing personalized learning resources and enhancing teaching effectiveness.

REFERENCES

[1] Jiang, Y., Ma, K. (2021). Blended teaching design in higher education based on deep learning model. In 2021 2nd International Conference on Information Science and Education (ICISE-IE), Chongqing, China, pp. 673-676. https://doi.org/10.1109/ICISE-IE53922.2021.00158

[2] Zhang, H., Huang, T., Liu, S., Yin, H., Li, J., Yang, H., Xia, Y. (2020). A learning style classification approach based on deep belief network for large-scale online education. Journal of Cloud Computing, 9: 1-17. https://doi.org/10.1186/s13677-020-00165-y

[3] Liu, H., Ko, Y.C. (2021). Cross-media intelligent perception and retrieval analysis application technology based on deep learning education. International Journal of Pattern Recognition and Artificial Intelligence, 35(15): 2152023. https://doi.org/10.1142/S0218001421520236

[4] Gao, L., Qin, G. (2021). Application of deep learning in the reform of Japanese education in the information age. In Application of Intelligent Systems in Multi-modal Information Analytics: 2021 International Conference on Multi-modal Information Analytics (MMIA 2021), Huhehaote, China, pp. 604-610. https://doi.org/10.1007/978-3-030-74811-1

[5] Xiao, X., Liu, X., Chen, X., Zhang, C., Hu, Q. (2020). Hot topic detection based on image intelligent understanding and comment mining. In 2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, pp. 1191-1194. https://doi.org/10.1109/ICIBA50161.2020.9276761

[6] Siagian, M.D., Suryadi, D., Nurlaelah, E., Tamur, M., Sulastri, R. (2021). Investigating students' concept image in understanding variables. In Journal of Physics: Conference Series, 1882(1): 012058. https://doi.org/10.1088/1742-6596/1882/1/012058

[7] Liu, Z., Dong, L., Wu, C. (2020). Research on personalized recommendations for students' learning paths based on big data. International Journal of Emerging Technologies in Learning (iJET), 15(8): 40-56. https://doi.org/10.3991/ijet.v15i08.12245

[8] Shi, D., Wang, T., Xing, H., Xu, H. (2020). A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning. Knowledge-Based Systems, 195: 105618. https://doi.org/10.1016/j.knosys.2020.105618

[9] Carbone, M., Colace, F., Lombardi, M., Marongiu, F., Santaniello, D., Valentino, C. (2021). An adaptive learning path builder based on a context aware recommender system. In 2021 IEEE frontiers in education conference (FIE), Lincoln, NE, USA, pp. 1-5. https://doi.org/10.1109/FIE49875.2021.9637465

[10] Huang, X. (2021). Hybrid collaborative filtering personalized learning path algorithm based on big data. In Frontier Computing: Proceedings of FC 2020, Springer Singapore, pp. 531-538. https://doi.org/10.1007/978-981-16-0115-6_59

[11] Sun, Y., Liang, J., Niu, P. (2021). Personalized recommendation of English learning based on knowledge graph and graph convolutional network. In International Conference on Artificial Intelligence and Security, Dublin, Ireland, pp. 157-166. https://doi.org/10.1007/978-3-030-78612-0

[12] Ismail, H. (2018). WikiRec: A personalized content recommendation framework to support informal learning in wikis. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore pp. 273-276. https://doi.org/10.1145/3209219.3213594

[13] Khan, A., Javed, A., Mahmood, M. T., Khan, M.H.A., Lee, I.H. (2021). Directional magnitude local hexadecimal patterns: A novel texture feature descriptor for content-based image retrieval. IEEE Access, 9: 135608-135629. https://doi.org/10.1109/ACCESS.2021.3116225

[14] Fresnedo, Ó., Laport, F., Castro, P.M., Dapena, A. (2021). Educational graphic tool for teaching fundamentals of digital image representation. Computer Applications in Engineering Education, 29(6): 1489-1504. https://doi.org/10.1002/cae.22402

[15] Gao, L., Wu, X., Wu, J., Xie, X., Qiu, L., Sun, L. (2021). Sensitive image information recognition model of network community based on content text. In Proceedings - 2021 IEEE 6th International Conference on Data Science in Cyberspace, DSC 2021, Shenzhen, China, pp. 47-52. https://doi.org/10.1109/DSC53577.2021.00014

[16] Zheng, Y., Wang, D., Xu, Y., Mao, Z., Zhao, Y., Li, Y. (2023). A bio-inspired method for personalized learning path recommendation problem. In 31st International Conference on Computers in Education, ICCE 2023 – Proceedings, 1: 147-149.

[17] Li, H., Peng, K., Ren, H., Shang, F. (2022). Personalized learning paths Recommendation based on Learner Personas. In 2022 3rd International Conference on Intelligent Design (ICID), Xi'an, China, pp. 290-293. https://doi.org/10.1109/ICID57362.2022.9969728

[18] Wang, Y., Han, L., Qian, Q., Xia, J., Li, J. (2022). Personalized recommendation via multi-dimensional meta-paths temporal graph probabilistic spreading. Information Processing & Management, 59(1): 102787. https://doi.org/10.1016/j.ipm.2021.102787