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

Guoli Xu1,2*, Cora Un In Wong1

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

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

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

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

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 pave 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) recommendation models 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.

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 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. This 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.

(a) Feature extraction across the entire image sequence 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.

(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

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 \( b \). 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 vector be represented by \( u^t \), the forget gate by \( d^t \), the output gate by \( p^t \), the memory state by \( z^t \), and the information updating the memory state by \( w^t \). The sigmoid activation function is denoted by \( \delta \), the element-wise multiplication by \( \odot \), the hyperbolic tangent function by \( \tanh \), the learnable parameters by \( Q_w \), 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:

\[
\begin{align*}
\quad u^t &= \delta \left( Q_w u^t + Q_n c^{t-1} + Q_n c^t + n_s \right)
\quad d^t &= \delta \left( Q_w u^t + Q_n c^{t-1} + Q_n c^t + n_s \right)
\quad p^t &= \delta \left( Q_w u^t + Q_n c^{t-1} + Q_n c^t + n_s \right)
\quad w^t &= \tanh \left( Q_w u^t + Q_n c^{t-1} + Q_n c^t + n_s \right)
\quad z^t &= d^t \odot \Phi z^{t-1} + u^t \odot \Phi w^t
\quad g^s &= p^t \odot \Phi \tanh \left( z^t \right)
\end{align*}
\]

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

\[
\begin{align*}
\quad g^s, z^s &= \text{LSTM} \left( a^t, c^t \right) g^{t-1}, z^{t-1} \right)
\end{align*}
\]

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 \( g^0 \), the learnable parameters by \( Q_w \), the visual features at time \( s \) by \( a^s \), the averaging function by \( \Theta \), and the feature of the \( s \)-th image extracted from the entire educational image content description image features \( a \) by \( Q_w \). The current sentence vector is represented by \( t^s \), the current visual feature by \( a^t \), the previous moment's hidden layer state by \( g^{t-1} \), and the corresponding memory state by \( z^{t-1} \), with the current moment's hidden layer state denoted by \( g^t \), and the corresponding memory state by \( z^t \). The expression for S_LSTM is given as:

\[
\begin{align*}
\quad g^t_0 &= Q^0_w \theta(a)
\quad a^t_\ast &= \Omega(a)
\quad g^t, z^t &= \text{LSTM}_{SE} \left( t^s, a^t_\ast \right) g^{t-1}, z^{t-1} \right)
\end{align*}
\]

The features inputted into the S_LSTM layer consist of contextual semantic features, which are composed of \( t \) and \( d^s \). Thus, by feeding contextual semantic information along with \( g^t_\ast \) and \( z^t_\ast \), into the S_LSTM layer, the outputs \( g^s_\ast \) and \( z^s_\ast \) 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 \( a^s \), the length of the generated sentence by \( j \), and a fully connected function by \( \Theta \). 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 \( x^j_k \), and the word-level attention feature by \( z^j \). The expression for the W_LSTM layer is as follows:

\[
\begin{align*}
\quad g^s &= Q^s_w \theta(a^t_\ast)
\quad x^j_k &= \text{softmax} \left( \Theta \left( g^j_{s-1}, a^t_\ast \right) \right)
\quad c^j &= \sum_{k=1}^{M} (x^j_k, a^s_k)
\quad g^j, z^j &= \text{LSTM}_{W_L} \left( g^j_{s-1}, z^j_{s-1} \right) g^j_{s-1}, z^j_{s-1} \right)
\quad g^j, z^j &= \text{LSTM}_{W_L} \left( g^j_{s-1}, z^j_{s-1} \right) g^{j-1}, z^{j-1} \right) \left( j = 1 \right)
\quad g^j, z^j &= \text{LSTM}_{W_L} \left( g^{j-1}, z^{j-1} \right) g^{j-1}, z^{j-1} \right) \left( j > 1 \right)
\end{align*}
\]

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

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:

\[
\begin{align*}
\quad O(b|a,\phi) &= \prod_{j=1}^{M} O(q_{u,j}|a, q_{u,1}, q_{u,2}, \ldots, q_{u,j-1})
\quad O(b|a,\phi) &= \sum_{a=0}^{V} (b(a|a,\phi))
\end{align*}
\]

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

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, \( j \), 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 models the corresponding topics, \( e \), across different 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, \( D^e \), across different topics, i.e., the feature's mixture components, \( \phi_j \).

(b) Image topic mixture selection. For each image, \( f_i \), within the educational image collection, the model selects a topic mixture, \( \psi^e_{f_i} \), consisting of \( J \) 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, \( f_i \), in the training library, the model generates the visual features, \( d_{ij} \). This process involves two steps: initially, based on the obtained topic mixture distribution, the model employs a multinomial distribution to generate a topic, \( c^{e_{f_i}}, \text{Multinomial}(\psi^e_{f_i}) \); subsequently, based on the generated topic, \( c^{e_{f_i}} \), the model further utilizes a multinomial distribution to generate the visual vocabulary related to that topic, \( d_{f_i} \), i.e., image features, \( d_{f_i} \text{Multinomial}(\theta^{c_{f_i}}) \).

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 \( \beta \) and \( \alpha \), the formula for calculating the joint distribution of hidden topics and vocabularies is provided as follows:

\[
O(q, c | \beta, \alpha) = O(c | \beta)O(q | c, \alpha)
\]  

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

\[
O(c | \beta) = \frac{\Gamma(J\beta)}{\Gamma(\beta)^J} \prod_j \frac{\Gamma(v_{ij} + \beta)}{\Gamma(V_j + J\beta)}
\]  

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

\[
O(q | c, \alpha) = \prod_c \left( \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \right)^J \prod_j \frac{\Gamma(v_{je} + \alpha)}{\Gamma(V_j + N\alpha)}
\]

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, \( \phi_j \) and \( \psi_{ij} \). 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 \( D_{ij} \), the current state’s topic by \( c_{ij} \), the current state’s topic by \( z_{je} \), other topic assignments besides the current state by \( e_{c_{ij}} \), the number of vocabularies \( s \) assigned to topic \( j \) excluding the current state by \( v^{ish}_{ij} \), and the number of vocabularies in document \( l \) assigned to topic \( j \) excluding the current state by \( v^{ish}_{lj} \). Given all states except \( c_{ij} \), the probability that the \( v \)-th feature in region \( e \) of educational image \( f \) is related to topic \( j \) can be calculated as follows:

\[
M = -\sum_{n=1}^{V} \sum_{k=1}^{M} \log O(q_{n,k} | n, q_{n,1}, q_{n,2}, \ldots, q_{n,k-1})
\]
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 $\phi_{j}$ and $\psi_{j}$ are as follows:

\[
\phi_{j}^{\mu} = \frac{\nu_{j}^{\mu} + \beta_{j}^{\mu}}{\sum_{j=1}^{J} \nu_{j}^{\mu} + \beta_{j}^{\mu}}
\]

\[
\psi_{j}^{\nu,e} = \frac{\nu_{j}^{\nu,e} + \alpha_{j}^{\nu,e}}{\sum_{j=1}^{J} \nu_{j}^{\nu,e} + \alpha_{j}^{\nu,e}}
\]

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

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

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 lack of statistical 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 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST Acc@4</th>
<th>MNIST Acc@10</th>
<th>CIFAR-10 Acc@4</th>
<th>CIFAR-10 Acc@10</th>
<th>MERLOT II Acc@4</th>
<th>MERLOT II Acc@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM</td>
<td>0.5326</td>
<td>0.3358</td>
<td>0.5214</td>
<td>0.3215</td>
<td>0.5248</td>
<td>0.3124</td>
</tr>
<tr>
<td>DMTL</td>
<td>0.5124</td>
<td>0.2789</td>
<td>0.4895</td>
<td>0.2789</td>
<td>0.4895</td>
<td>0.2587</td>
</tr>
<tr>
<td>DCRS</td>
<td>0.5269</td>
<td>0.3215</td>
<td>0.5123</td>
<td>0.3216</td>
<td>0.4756</td>
<td>0.2659</td>
</tr>
<tr>
<td>NTN-R</td>
<td>0.5268</td>
<td>0.3154</td>
<td>0.4789</td>
<td>0.2659</td>
<td>0.4895</td>
<td>0.2641</td>
</tr>
<tr>
<td>MBCF</td>
<td>0.4569</td>
<td>0.2369</td>
<td>0.5136</td>
<td>0.2845</td>
<td>0.4126</td>
<td>0.2123</td>
</tr>
<tr>
<td>FHN</td>
<td>0.6123</td>
<td>0.3789</td>
<td>0.6147</td>
<td>0.3784</td>
<td>0.5326</td>
<td>0.3159</td>
</tr>
<tr>
<td>Single-feature LDA recommendation model</td>
<td>0.6589</td>
<td>0.4251</td>
<td>0.6239</td>
<td>0.4126</td>
<td>0.5895</td>
<td>0.3895</td>
</tr>
<tr>
<td>Multi-feature LDA recommendation model</td>
<td>0.6589</td>
<td>0.4589</td>
<td>0.6258</td>
<td>0.4236</td>
<td>0.6123</td>
<td>0.4259</td>
</tr>
</tbody>
</table>
Table 3. Performance of various algorithms in visual feature topic modeling

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ImageNet</th>
<th>COCO</th>
<th>PASCAL VOC</th>
<th>Flickr30k</th>
<th>CelebA</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM</td>
<td>0.5685</td>
<td>0.4158</td>
<td>0.4258</td>
<td>0.5148</td>
<td>0.4897</td>
<td>0.3514</td>
</tr>
<tr>
<td>CSN</td>
<td>0.5412</td>
<td>0.4869</td>
<td>0.5126</td>
<td>0.5126</td>
<td>0.5216</td>
<td>0.3789</td>
</tr>
<tr>
<td>SCE-Net</td>
<td>0.5123</td>
<td>0.5124</td>
<td>0.5268</td>
<td>0.5269</td>
<td>0.5124</td>
<td>0.4216</td>
</tr>
<tr>
<td>Multi-feature LDA recommendation model</td>
<td>0.6523</td>
<td>0.5169</td>
<td>0.5789</td>
<td>0.6895</td>
<td>0.6784</td>
<td>0.5698</td>
</tr>
</tbody>
</table>

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 (DMLT), 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 was 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.

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


