Advancing Educational Management with the ATT-MR-WL Intelligent Question-answering Model

Ying Ba

College of Civil and Architectural Engineering, North China University of Science and Technology, Tangshan 063210, China
E-mail: by20240716@163.com; baying@ncst.edu.cn

Received 08 March 2024; Accepted 17 October 2024

Abstract

Higher education plays a critical role in cultivating talent, preserving culture, and promoting social progress. However, current challenges, such as inefficient information dissemination and low problem-solving efficiency among students, highlight the need for intelligent question-answering systems. These systems, leveraging artificial intelligence and natural language processing technologies, enable rapid and accurate responses to student queries, thereby providing intelligent support for higher education management. This study introduces the ATT-MR-WL model, a generative AI system integrating Mask R-CNN and Word2Vec+LSTM to enhance intelligent question-answering functionality. The model, customized to handle both text and visual data, is evaluated using the established VQA v2.0 dataset and a specially developed EM dataset reflecting university management scenarios. The ATT-MR-WL model demonstrates a 3% accuracy improvement over traditional methods and enhances its ability to handle multimodal queries. This research provides important insights for enhancing the efficiency and quality of higher education management and advancing the process of educational informatization.

Keywords: Intelligent question-answering, higher education management, Mask R-CNN, Word2Vec+LSTM, ATT-MR-WL.

Journal of Web Engineering, Vol. 23_7, 973–1002.
doi: 10.13052/jwe1540-9589.2373

1 Introduction

Contemporary education is no longer merely about imparting knowledge, it faces complex and evolving challenges. Education aims to cultivate students' comprehensive qualities and promote their holistic development [1]. However, at present, educational systems universally encounter a range of issues. These issues include, but are not limited to, unequal distribution of educational resources, uneven teaching quality, challenges in meeting personalized student needs, and insufficient integration of educational technologies [2]. Traditional teaching methods often struggle to meet diverse learning needs, prompting educators to urgently seek new approaches and tools to address these challenges. The swift advancements in artificial intelligence and deep learning technologies have led to their increasing adoption in the education sector to address various challenges [3, 4]. Deep learning is extensively used in educational research as a subset of artificial intelligence. Through deep learning techniques, researchers can better analyze educational data, uncover hidden patterns within them, and provide a scientific basis for educational decision-making. For example, deep learning can be used in course recommendation systems, intelligent teaching aids, and other areas, bringing new possibilities to educational practice [5]. In current educational research, intelligent question-answering technology is increasingly valued and considered one of the effective approaches for addressing educational challenges. Intelligent question-answering systems can utilize deep learning and other technologies to automatically understand users' questions and provide accurate, timely answers or suggestions. This technology not only helps students better understand knowledge but also provides personalized teaching support for educators [6, 7]. The application of intelligent question-answering systems will introduce novel approaches to educational activities, fostering continuous innovation and progress in the education sector.

There has been a series of studies in the education field aiming to solve various problems and promote innovative development using technologies such as deep learning. This article will introduce four representative studies and outline their models and existing shortcomings. In the first study, researchers developed a "knowledge graph-driven personalized learning recommendation system" [8]. This system utilizes deep learning technology to construct a knowledge graph to capture information such as students' learning interests, knowledge levels, and learning paths. Based on this knowledge graph, the system can provide personalized learning resource

recommendations for each student. However, the system still has limitations in understanding and responding to students' personalized needs, requiring more data and algorithm optimization to improve accuracy and intelligence. The second study involves the development of a "virtual teacher intelligent question-answering system" [9]. The research team used deep learning models to construct a virtual teacher capable of understanding students' questions and providing corresponding answers or explanations. By engaging in real-time interaction with students, the system offers personalized teaching assistance. However, due to limited training data for the virtual teacher model, the system performs poorly in answering complex questions and requires more real-world data for training. Another study focuses on a "deep learning-based essay evaluation system" [10]. This system utilizes deep learning technology to automatically score and evaluate students' essays. By analyzing aspects such as the language structure, logical thinking, and expressive ability of essays, the system provides corresponding evaluations and suggestions. However, due to limitations in understanding essay content and context, evaluation results may contain subjective biases, necessitating further research to enhance objectivity and accuracy. The last study is a question-answering system based on generative adversarial networks (GANs), which proposes a novel method to expand training data in the education field [11]. By generating question–answer pairs through a generative adversarial network, the system can better train intelligent question answering models and enhance its effectiveness in the educational domain. However, this method may encounter the problem of training instability in specific scenarios in the education field. In summary, although these related studies have propelled the development of the education field using technologies like deep learning, they still have some shortcomings, such as insufficient training data and limited ability to handle complex problems. Therefore, future research needs to further refine models, enhance system performance, and improve intelligence to better serve educational practices and teaching needs.

Based on existing shortcomings, we propose the ATT-MR-WL network, which integrates an image feature extraction module (Mask R-CNN), a text feature extraction module (Word2Vec+LSTM), and multiple attention units. The ATT-MR-WL network comprehensively understands and represents multimodal information, providing richer features for answer reasoning. Compared to traditional intelligent question-answering systems,

it possesses stronger expressive and reasoning capabilities, enabling more accurate responses to user queries. This opens up new possibilities for the application of intelligent question-answering systems in higher education management, offering more accurate and convenient management tools and decision support.

In our research, we propose an intelligent question-answering system that contributes to higher education management in the following three aspects:

- We have proposed a novel intelligent question-answering system model, namely the ATT-MR-WL model. This model integrates natural language processing and deep learning technologies, achieving a more accurate understanding and processing of questions in the education domain through multimodal information fusion.
- We thoroughly validated the effectiveness and performance of our proposed model through extensive experiments. Training and testing on a large amount of real education data demonstrated that our model has significant advantages in addressing intelligent question-answering problems in the education domain, exhibiting higher accuracy and stronger generalization capability.
- We successfully applied our model to practical education scenarios and achieved remarkable results. Through collaboration with educational institutions, our intelligent question-answering system provides educators and learners with more convenient and efficient teaching tools, fostering further development and application of educational informatization.

The rest of this paper is structured as follows. Section 2 reviews the related work in the field of intelligent question-answering systems and educational management. In Section 3, we describe the proposed ATT-MR-WL model, detailing its architecture and components. Section 4 presents the experimental setup and evaluation, including datasets and performance metrics. The results and comparative analysis are discussed in Section 5. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

2.1 Progress in Intelligent Assisted Teaching on Online Learning Platforms

Intelligent assisted teaching systems are gradually becoming an integral part of the education landscape on today's online learning platforms [12, 13].

These systems utilize advanced technologies such as artificial intelligence, machine learning, and data analysis to offer customized learning support and teaching assistance for students and educators. Their key features include individualized learning support, instant feedback and guidance, a variety of learning resources, and teaching aids [14]. By evaluating students' learning behaviors and performance, these systems can personalize learning paths and resources for each student, providing targeted advice and assistance. Additionally, the systems provide teachers with the ability to monitor students' learning progress and performance in real-time, helping them adjust teaching strategies and course design accordingly.

These systems have made significant contributions to enhancing the flexibility and efficiency of learning and teaching methods. By collecting and analyzing large volumes of educational data, they also contribute to decision-making processes and instructional improvements, fostering educational innovation. While these systems offer many benefits, there remains room for further development, particularly in expanding their ability to process diverse types of educational data.

In summary, the application of intelligent assisted teaching systems on online learning platforms offers students and teachers more flexible and efficient learning and teaching methods. This contributes to improving learning outcomes and teaching quality, ultimately driving progress and development in education.

2.2 Research on Personalized Education Models Based on Big Data Analysis

Research on personalized education models based on big data analysis is currently a key focus in education. These models utilize advanced data analysis techniques to collect and process large amounts of student learning data, aiming to achieve a deep understanding and precise analysis of each student's individualized learning needs [15, 16] In this model, data on students' learning behaviors, study habits, progress, interests, and hobbies are collected and recorded, and then subjected to in-depth analysis through data mining and analysis algorithms. Through the examination of this learning data, personalized education models can discover each student's learning characteristics and needs, thereby providing personalized learning paths and resources [17]. For example, for students who enjoy reading, the system may recommend more reading materials and related learning activities; for students who struggle in certain subjects, the system may provide

more targeted practice questions and problem-solving techniques [18]. Additionally, personalized education models can provide important reference information for teachers. By analyzing students' data, educators can gain a deeper understanding of each student's progress, quickly identify learning challenges, and implement appropriate teaching strategies. This approach enhances teaching effectiveness and improves students' learning experiences.

While these models have made significant progress in personalizing learning experiences through big data, there remains potential for further refinement, particularly in enhancing real-time data processing and adapting to students' changing learning needs more dynamically.

2.3 The Latest Research on Multimodal Information Fusion Techniques in Educational Intelligent Question-answering Systems

Recent research has been dedicated to integrating information from different modalities to enhance the performance of educational intelligent question-answering systems. These studies focus on designing text and image fusion models to enable systems to comprehensively understand questions [19]. These models are capable of simultaneously processing textual descriptions and related images, thereby providing more accurate and comprehensive answers. Additionally, audio and text fusion models have also been developed to better understand the semantics and tones of students' inquiries [20]. By combining audio information with textual content, systems can better comprehend students' questions and provide more suitable responses. These studies not only concentrate on model design but also explore a range of optimization methods to improve system performance [21]. For instance, the introduction of attention mechanisms allows systems to pay more attention to important information segments [22, 23]. Moreover, the application of feature fusion techniques effectively integrates features from different modalities, thereby enhancing the accuracy and efficiency of systems.

These advances represent important contributions to the development of intelligent question-answering systems in education. By leveraging multimodal fusion, these systems are better equipped to meet the demands of educational management and provide robust learning support. Future research may continue to explore optimization strategies to further improve the capabilities of multimodal systems in complex educational environments.

3 Method

3.1 Overview of Our Network

The intelligent question answering system we propose is based on the fusion of multimodal information, aiming to enhance the ability to answer and solve questions in educational scenarios. This system adopts a comprehensive framework called the ATT-MR-WL network, which integrates both image and text data and utilizes multiple attention units for deep interaction and information fusion, achieving more precise answer reasoning. Here's a brief description of the model. Image feature extraction module (Mask R-CNN): This module employs the Mask R-CNN model to extract rich visual features from input images, capturing objects and scene information crucial for understanding the context and answering questions. Text feature extraction module (Word2Vec+LSTM): This module utilizes Word2Vec to convert textual questions into word vector representations. It then employs an LSTM model to capture semantic information and contextual relationships within the text sequences, generating text features essential for understanding the meaning and intent of questions. The attention mechanism module introduces multiple attention units, including multi-head attention and guided attention, to perform weighted fusion of image and text features, facilitating interaction and information exchange between different modalities.

Our network construction process involves utilizing a pre-trained Mask R-CNN model to extract image features and convert the textual data into word vectors using Word2Vec. Next, we input the image features and text features into an LSTM model to obtain the textual feature representation of the question. Then, through multiple attention units, we perform weighted fusion of the image and text features to obtain a richer feature representation. Finally, we input the fused features into the answer reasoning module, where the model conducts reasoning and answers the questions. Figure 1 is the ATT-MR-WL model framework diagram.

Our model holds significant importance for intelligent question-answering systems in higher education by integrating image and text data, which achieves a comprehensive understanding of educational scenarios and improves the accuracy of responses. Moreover, the attention mechanism within our model effectively orchestrates interaction and information flow among different modalities, facilitating profound information fusion and endowing the intelligent question-answering system with robust reasoning capabilities. Most notably, our model stands poised to provide more effective

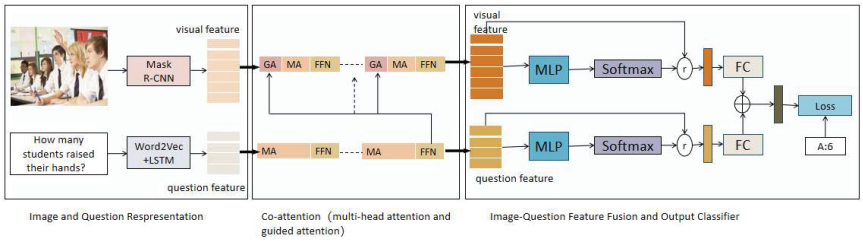

Figure 1 ATT-MR-WL multimodal network model diagram.

support for higher education management by furnishing personalized, real-time question answering, and learning guidance to both students and educators. Consequently, this advancement is expected to elevate teaching standards and learning outcomes.

3.2 Image Feature Extraction Module: Mask R-CNN

Mask R-CNN (Mask region-based convolutional neural network) is an advanced deep learning model used for object detection and image segmentation tasks. Its principle is based on Faster R-CNN, but it introduces an additional segmentation branch, allowing the model not only to accurately detect objects in images but also to generate precise pixel-level segmentation for each object [24]. In Mask R-CNN, convolutional neural networks (CNNs) are first used to extract image features. The extracted features are passed to the region proposal network (RPN) to generate candidate object bounding boxes. The RoI (region of interest) align layer then accurately crops and aligns these features for the classification and segmentation branches. The classification branch is responsible for determining the category of candidate objects, while the segmentation branch applies convolution and upsampling operations on the RoI Align outputs to create binary masks for each object, achieving pixel-level semantic segmentation [25]. The model schematic is presented in Figure 2.

The Mask R-CNN model equations are formulated as follows:

Given an input feature map F, the region proposal network (RPN) proposes candidate object bounding boxes:

$$\mathrm{RPN}(\mathrm{F}) \rightarrow \{\mathrm{B}\}. \tag{1}$$

The RoI Align process extracts a small feature map for each object proposal, aligning the features precisely with the object:

$$\text{RoI Align}(\mathrm{F}, \mathrm{B}) \rightarrow \mathrm{F}_{\mathrm{B}}. \tag{2}$$

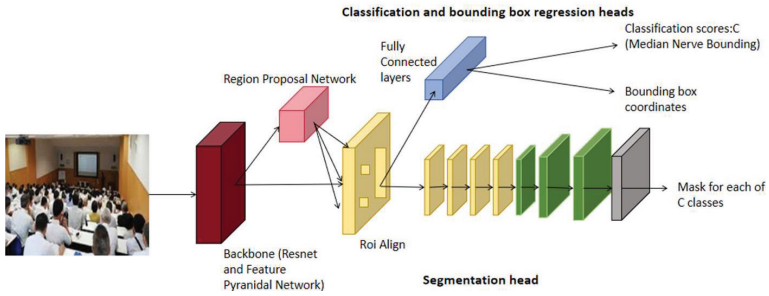

Figure 2 The architecture of Mask R-CNN.

The classification head outputs a probability distribution over C classes for each proposed box:

$$\text{Classification head}(\mathrm{F_B}) \rightarrow \mathrm{p}. \tag{3}$$

The regression head outputs bounding box refinements:

$$\text{Bounding box regression head}(\mathrm{F_B}) \rightarrow \Delta\mathrm{B}. \tag{4}$$

The segmentation head outputs a binary mask that indicates the pixel-wise position of the object within the bounding box:

$$\text{Segmentation head}(\mathrm{F_B}) \rightarrow \mathrm{M}. \tag{5}$$

The Mask R-CNN employs a loss function L, aggregating $\mathrm{L_{cds}}$ for classification, $\mathrm{L_{box}}$ for bounding box adjustments, and $\mathrm{L_m ask}$ for binary mask cross-entropy.

$$\mathrm{L} = \mathrm{L_{cls}}(\mathrm{p}, \mathrm{u}) + \mathrm{L_{box}}(\Delta\mathrm{B}, \mathrm{v}) + \mathrm{L_{mask}}(\mathrm{M}, \mathrm{w}) \tag{6}$$

where u is the ground truth class, v is the ground truth bounding box, and w is the ground truth mask.

In the field of education, the Mask R-CNN model can be utilized to identify objects and scenes within educational contexts, such as classrooms, blackboards, and books. This capability aids intelligent question-answering systems to better understand the context of questions. For instance, by recognizing objects within a classroom, the system can more accurately comprehend questions related to the classroom environment and provide more appropriate answers. Additionally, Mask R-CNN can be employed to identify students' behaviors and emotions, such as attentiveness or confusion. This functionality enables the system to personalize responses and offer tailored learning support more effectively.

The integration of Mask R-CNN in educational intelligent question-answering systems enriches the system with detailed contextual information, enhancing both its understanding capabilities and overall effectiveness in delivering personalized education.

3.3 Question Text Feature Extraction Module: Word2Vec+LSTM

The Word2Vec+LSTM model is a deep learning architecture that combines word embedding and long short-term memory (LSTM) networks for feature extraction and semantic modeling of text sequences. Word2Vec is a widely used word embedding technique that maps vocabulary into a continuous low-dimensional vector space, capturing the semantic relationships between words [26]. Conversely, LSTM is a unique form of RNN that efficiently handles extended dependencies in text sequences, thereby enhancing the understanding and modeling of textual data [27]. In this model, input text sequences are first transformed into word vector representations using Word2Vec. These word vectors are then passed into the LSTM network, which extracts features and models the semantics of the text. Through its continuous update and forget gate mechanisms, the LSTM network captures long-term dependencies in the text, generating hidden representations that contain semantic information and contextual relationships. These hidden representations offer rich semantic features for subsequent question-answering tasks. The overall structure of the Word2Vec+LSTM model is shown in Figure 3.

Word2Vec includes two main models: the continuous bag-of-words model (CBOW) and the continuous Skip-gram model [28]. CBOW predicts the current word t from the surrounding context, while Skip-gram predicts context words using the current word t. Both models feature an input layer, a projection layer, and an output layer, as shown in Figure 4.

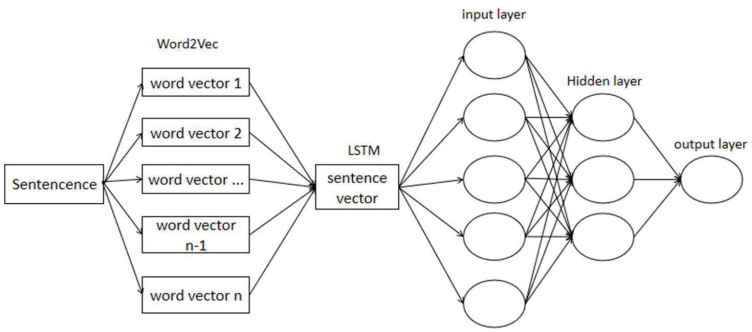

Figure 3 Word processing structure diagram.

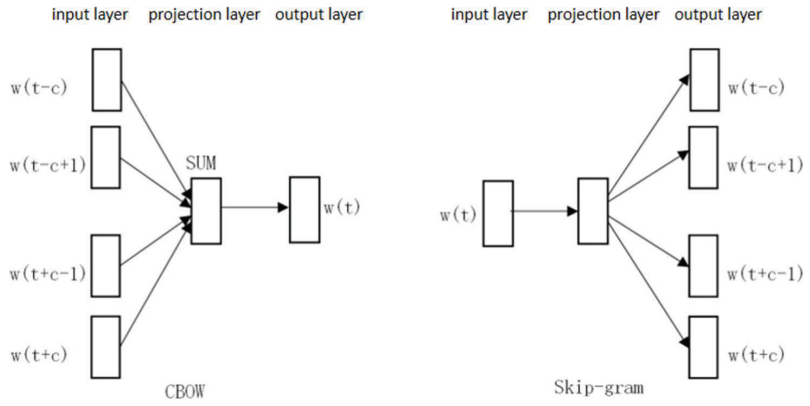

Figure 4 Word2Vec model.

In the CBOW model, context words are first transformed into vector representations, which are then averaged to predict the target word. The formula for the CBOW model is:

$$\mathrm{v_{context}} = \frac{1}{2\mathrm{c}} \sum_{-\mathrm{c}\le \mathrm{j}\le \mathrm{c}, \mathrm{j}\neq 0} (\mathrm{v_{w(t+j)}} \odot \sigma(\mathrm{W_{f1}} \mathrm{v_{w(t+j)}} + \mathrm{b_{f1}})) \tag{7}$$

$$\mathrm{P(w(t)|w(t-c),\ldots,w(t+c))} = \frac{\exp((\mathrm{v'_{w(t)}} \cdot \mathrm{W_o} + \mathrm{b_o}) \cdot \mathrm{v_{context}})}{\sum_{\mathrm{w}\in \mathrm{W}} \exp((\mathrm{v'_w} \cdot \mathrm{W_o} + \mathrm{b_o}) \cdot \mathrm{v_{context}})} \tag{8}$$

where $\mathrm{v_w(t+j)}$ is the vector representation of a context word, $\mathrm{v_w(t)'}$ is the "output" vector representation of the target word, $\mathrm{v_{context}}$ is the averaged context vector, W is the vocabulary, $\mathrm{W_f1}$ and $\mathrm{W_o}$ are learnable weight matrices, $\mathrm{b_f1}$ and $\mathrm{b_o}$ are bias terms, σ is a non-linear activation function, $\odot$ denotes element-wise multiplication, and c is the size of the context window.

For the Skip-gram model, the objective is to predict the surrounding context words based on a given target word $\mathrm{w(t)}$. The formula for the Skip-gram model is:

$$\mathrm{P(w(t-c),\ldots,w(t+c)|w(t))} = \prod_{-\mathrm{c}\le \mathrm{j}\le \mathrm{c}, \mathrm{j}\neq 0} \mathrm{P(w(t+j)|w(t))} \tag{9}$$

$$\mathrm{P(w(t+j)|w(t))} = \frac{\exp((\mathrm{v'_{w(t+j)}} \cdot \mathrm{W_1} + \mathrm{b_1}) \cdot (\mathrm{v_{w(t)}} \cdot \mathrm{W_2} + \mathrm{b_2}))}{\sum_{\mathrm{w}\in \mathrm{W}} \exp((\mathrm{v'_w} \cdot \mathrm{W_1} + \mathrm{b_1}) \cdot (\mathrm{v_{w(t)}} \cdot \mathrm{W_2} + \mathrm{b_2}))} \tag{10}$$

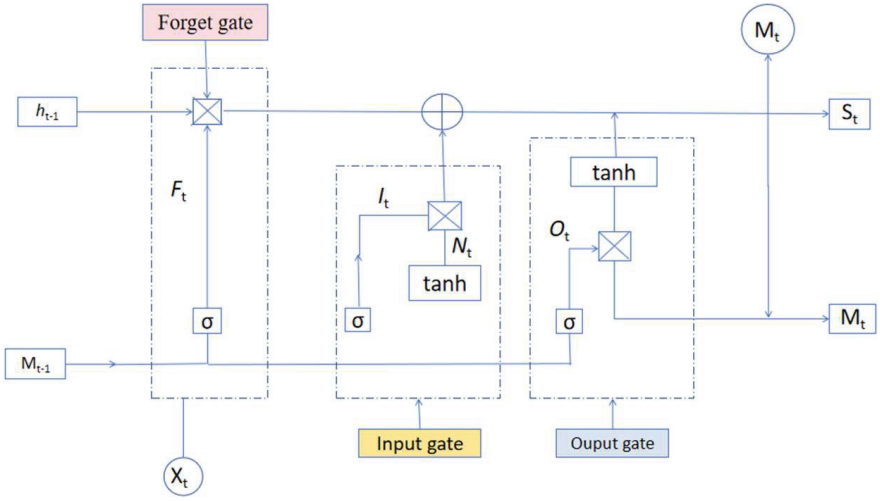

Figure 5 Word processing structure diagram.

where $\mathrm{v_w(t)}$ is the vector representation of the target word, $\mathrm{v_w(t+j)'}$ is the output vector representation of a context word, $\mathrm{W_1}$ and $\mathrm{W_2}$ are learnable weight matrices, $\mathrm{b_1}$ and $\mathrm{b_2}$ are bias terms, and c is the size of the context window.

A recurrent neural network (RNN) can add memory units to the original neural network to have a memory function for history. It can theoretically handle sequences of any length and can handle sequence-related problems better than ordinary neural networks. However, in practice, it is impossible to memorize information with a long time span. The RNN variant LSTM model can learn long-term dependencies. LSTM allows information to pass through selectively through the structure of the gate. The LSTM unit consists of several gates that control the flow of information, as depicted in Figure 5.

To provide a more detailed understanding of the training and optimization process of LSTM, we introduce the following core formulas and calculations:

Loss function: For a sequence prediction task, the cross-entropy loss L is used to measure the difference between the predicted output $\widehat{\mathrm{y}}_\iota$ and the actual target y_ι:

$$\mathrm{L} = -\sum_{\mathrm{t}=1}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{y_{t,k}} \log(\widehat{\mathrm{y}}_{\mathrm{t,k}}. \tag{11}$$

Gradient for the forget gate: The gradient for the weight matrix $\mathrm{W_f}$ of the forget gate is computed as follows:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_f}} = \sum_{\mathrm{t=1}}^{\mathrm{T}} \delta_{\mathrm{F_t}} \cdot \left[\mathrm{h_{t-1}} \bigoplus \mathrm{X_t}\right]^{\mathrm{T}}. \tag{12}$$

Error term for the output gate: The error term for the output gate O_ι is given by:

$$\delta_{\mathrm{O_t}} = \frac{\partial \mathrm{L}}{\partial \mathrm{h_t}} \odot \tanh(\mathrm{C_t}) \odot \mathrm{O_t} \odot (1 - \mathrm{O_t}). \tag{13}$$

Error term for the input gate: The error term for the input gate I_ϵ is computed as:

$$\delta_{\mathrm{I_t}} = \frac{\partial \mathrm{L}}{\partial \mathrm{C_t}} \odot \widehat{\mathrm{C_t}} \odot \mathrm{I_t} \odot (1 - \mathrm{I_t}). \tag{14}$$

Gradient for the cell state: The gradient for the cell state is calculated by:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{C_t}} = \frac{\partial \mathrm{L}}{\partial \mathrm{h_t}} \odot \mathrm{O_t} \odot (1 - \tanh^2(\mathrm{C_t})) + \frac{\partial \mathrm{L}}{\partial \mathrm{C_{t+1}}} \odot \mathrm{F_{t+1}}. \tag{15}$$

Weight update for the forget gate: The weight update for the forget gate is performed as follows:

$$\mathrm{W_f} \leftarrow \mathrm{W_f} - \eta \frac{\partial \mathrm{L}}{\partial \mathrm{W_f}}. \tag{16}$$

In these formulas, η represents the learning rate, W denotes the weight matrices. The σ function is the sigmoid activation, tanh is the hyperbolic tangent function, $\bigoplus$ signifies concatenation, and G denotes element-wise multiplication (Hadamard product).

The Word2Vec+LSTM model is vital to our intelligent question-answering system. By processing question text sequences, the model effectively extracts semantic features and converts them into continuous vector representations. These representations encapsulate the semantic information and contextual relationships of the questions, helping the system to better understand and answer even complex natural language questions by grasping both intent and context, thus improving reasoning accuracy.

In educational settings, the Word2Vec+LSTM model can analyze the semantic relationships and logical sequences of questions posed by students, allowing the system to better understand their learning needs. Additionally, the model can assess students' learning behaviors, such as identifying their

interests and difficulties. This enables educators to provide personalized teaching suggestions and learning guidance. Consequently, applying the Word2Vec+LSTM model enhances the system's ability to meet students' needs and deliver tailored educational support.

3.4 Attention Mechanism Module

The attention mechanism is a powerful tool in neural networks, especially useful in processing sequential data in fields like natural language processing and computer vision [29]. Its principle involves dynamically adjusting attention weights for different positions based on current input and contextual information at each computational step. This allows the network to focus more closely on information relevant to the current task while disregarding irrelevant parts, which improves the model's ability to handle sequences of varying lengths and complexities, thus enhancing performance and generalization.

In our intelligent question-answering system, we utilize both multi-head attention and guided attention mechanisms to significantly improve the model's performance. Multi-head attention enables the model to focus on information from multiple subspaces simultaneously, helping capture complex relationships between features and improving feature representation. Additionally, we have introduced the guided attention mechanism for guiding interactions between different modalities, facilitating the integration and communication of information. The guided attention mechanism enables the system to better leverage multi-modal information, promoting interaction and fusion of information between different modalities, thus further enhancing the system's performance and effectiveness. The model can better explore the internal correlations between image and text features, enhancing sensitivity to detail and thereby strengthening the system's representational capacity. Meanwhile, the guided attention mechanism further enhances interaction between different modalities, enabling the model to comprehensively utilize multi-modal information, thus improving the system's overall performance. By using these attention mechanisms, the model can better understand the internal correlations between image and text features, increasing sensitivity to details and improving its representational capacity. In turn, this allows the system to generate more accurate answers in educational contexts, providing smarter and more efficient solutions for higher education management.

Attention mechanisms have become integral components in various deep learning architectures for tasks involving sequential or relational data. At

its core, attention mechanisms enable models to concentrate on particular sections of the input when making predictions or generating outputs.

Softmax attention mechanism: The softmax attention mechanism computes attention weights through a scaled dot-product between the current input and a context vector. The scaling factor, derived from the dimensionality of the key vectors, aids in stabilizing the gradient during the learning process. The mathematical formulation is expressed as follows:

$$\begin{aligned}&\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V})\\&\quad=\mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}+\mathrm{diag}(\mathrm{Q})\mathrm{K}-\mathrm{Kdiag}(\mathrm{Q})}{\sqrt{\mathrm{d_k}+\epsilon}}+\mathrm{M}\right)\mathrm{V}\end{aligned}\tag{17}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and $\mathrm{d_k}$ represents the dimensionality of the key vectors, used to scale the dot-product, and hence stabilize training dynamics.

Multi-head attention mechanism: This mechanism enhances the traditional softmax attention by parallel processing information across various representation suhspaces, allowing the model to capture multi-dimensional data features more comprehensively.

$$\mathrm{Multi\,head}(\mathrm{Q},\mathrm{K},\mathrm{V})=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head_h})\mathrm{W}^{\mathrm{O}}$$

where $\mathrm{head}_4=\mathrm{attention}(\mathrm{QW}_{\mathrm{i}}^{\mathrm{Q}},\mathrm{KW}_{\mathrm{i}}^{\mathrm{K}},\mathrm{VW}_{\mathrm{i}}^{\mathrm{V}}),\mathrm{W}_{\mathrm{j}}^{\mathrm{Q}},\mathrm{W}_{\mathrm{i}}^{\mathrm{K}},\mathrm{W}_{\mathrm{i}}^{\mathrm{V}}$ are linear transformation matrices for each head, and W^{O} is the final linear transformation matrix to combine heads.

Guided attention mechanism: Guided attention incorporates additional guidance signals to modulate the attention mechanisın's behavior, enhancing interpretability or direc:ting attention based on specific criteria.

$$\begin{aligned}&\mathrm{Guided\ attention}(\mathrm{Q},\mathrm{K},\mathrm{V},\mathrm{G})\\&\quad=\mathrm{softmax}\left(\frac{\mathrm{Q}((\mathrm{K}+\alpha\mathrm{G})\mathrm{W_k})^{\mathrm{T}}+\beta\mathrm{GW_g}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{VW_v}\end{aligned}\tag{18}$$

where G is the guidance signal, α is the weight. parameter for the guidance signal, β is another weight parameter for the guidance influence, $\mathrm{W_k}$ and $\mathrm{W_g}$ are learnable weight matrices for the keys and guidance signals respectively, and $\mathrm{W_v}$ is a learnable weight matrix for the values. Other variables remain the same as in the standard softmax attention mechanism.

4 Experiment

4.1 Dataset

Released jointly by Stanford University, the Indian Institute of Technology, and Microsoft Research, the VQA v2.0 dataset [30] aims to aid models in comprehending visual content and answering natural language questions. With over 200,000 image–question pairs, it covers diverse domains such as object recognition, scene understanding, reasoning, and common-sense inference. Questions vary in type, from factual to inferential, ensuring dataset complexity. Human-annotated answers accompany each question, providing diverse perspectives derived from image data and common-sense reasoning. Serving as a significant benchmark, VQA v2.0 fosters research in visual question answering, driving advancements in deep learning-based models and system performance evaluation.

4.2 Experimental Details

In our experiment, we initially selected the VQA v2.0 dataset as the benchmark for evaluating our model. This dataset comprises approximately 200,000 images and over 1.1 million questions. We utilized this dataset for both training and testing our model. Regarding model parameter settings, we employed a set of tuned parameters. For the image feature extraction model (Mask R-CNN), we configured the ROI Pooling size to be 7×7, learning rate as 0.001, batch size as 16, training epochs as 50, and utilized cross-entropy loss function with stochastic gradient descent (SGD) optimizer for parameter updates. For the text feature extraction model (Word2Vec+LSTM), we utilized 512 LSTM hidden units, 300-dimensional word vectors, learning rate of 0.01, batch size of 64, training epochs of 30, and initialized word embeddings using a pre-trained Word2Vec model. In the attention mechanism model, we set 8 attention heads and a hidden layer size of 512.

During the experimentation, we conducted data preprocessing initially, extracting features from images using the Mask R-CNN model and text features through the Word2Vec+LSTM model. Subsequently, we trained the ATT-MR-WL model on the training set of the VQA v2.0 dataset and evaluated the model on the validation set. Finally, we performed an in-depth analysis of the experimental results, including model performance across different question types, analysis of error cases, and potential improvement methods. Through these experiments, we comprehensively assessed our model's performance and generalization ability on the VQA task, providing valuable insights for further research and applications.

4.2.1 Model evaluation

We utilize various evaluation metrics to thoroughly evaluate the effectiveness of our model within the context of visual question answering (VQA). These metrics, namely Yes/No, Number, Other, and Overall, serve distinct purposes in evaluating the model's efficacy across various question types and offer valuable insights into its capabilities.

1. Yes/No: This metric evaluates the model's ability to correctly answer binary questions with a "yes" or "no" response. These questions typically require the model to understand basic concepts or properties depicted in the image.
2. Number: The "Number" metric assesses the model's performance in answering questions that require numerical responses. These questions often involve counting objects, estimating quantities, or providing numerical descriptions related to the image content.
3. Other: This category encompasses a variety of question types beyond binary or numerical answers. It evaluates the model's performance in responding to open-ended questions that may require descriptive or contextual answers beyond a simple "yes" or "no" or numerical response.
4. Overall: The "Overall" metric offers a thorough evaluation of the model's performance across all question types. It considers the accuracy of the model's responses to Yes/No, Number, and Other types of questions, providing a holistic view of the model's effectiveness in comprehending and answering questions based on visual content.

These evaluation metrics collectively provide insights into different aspects of the model's capabilities in visual question answering. By analyzing performance across these diverse question types, we gain a better understanding of the model's strengths and weaknesses in comprehending visual information and generating accurate responses.

4.2.2 Self-built dataset: EM dataset

To overcome the limitations of existing datasets in covering the diversity of the education domain and to provide comprehensive support for the application of intelligent question answering systems in education, we have decided to construct our own educational dataset. Existing datasets often focus on specific subjects or types of questions, whereas we aim to create a dataset that encompasses multiple disciplines, various question types, and rich educational scenarios to better meet the needs of intelligent question

answering systems in education. Moreover, creating our own dataset allows us to improve the quality and accuracy of the data, which in turn boosts the performance and effectiveness of the systems.

Based on the aforementioned considerations, we have constructed a multidimensional educational dataset covering various aspects of higher education. This dataset encompasses course content, teaching resources, academic achievements, teaching practices, student information, and educational policies. Specifically, we have gathered course outlines, teaching plans, textbook content, courseware, and teaching videos for various subjects as teaching resources. Additionally, academic papers, research project participation, and awards information for both teachers and students have been collected as part of the dataset. Furthermore, we have documented teachers' teaching activities and practical experiences, such as classroom teaching, laboratory sessions, and internship practices. Student personal information and academic performance, including names, ages, genders, student IDs, course selections, and learning progress, have also been included in the dataset. Finally, policy documents, management regulations, and reform plans from educational authorities have been compiled to comprehensively understand educational policies and management practices. Through these data collections, we ensure that the constructed dataset meets the requirements of intelligent question answering systems and provides sufficient support for model training and application.

Our dataset comprises a total of 5000 data samples. These data have undergone rigorous annotation and validation processes to ensure the quality and usability of the dataset. The image data underwent visual inspection to ensure clarity and quality. The textual questions and answer annotations were manually labeled and reviewed to ensure the accuracy of questions and consistency of answers. Furthermore, we conducted data analysis on the dataset, including statistical analysis of the distribution of different types of questions and answers, providing essential data support for subsequent model training and evaluation.

4.2.3 Data preprocessing

During the data preprocessing stage, we implemented a series of steps to prepare the image data and textual questions for effective processing by the ATT-MR-WL model. For the image data, we resized all images to a size of 256×256 pixels. Subsequently, we normalized the images, scaling the pixel values to the range [0,1], to accelerate model training and improve stability. To support these preprocessing steps, we utilized the Python Imaging

Table 1 Accuracy comparison of different methods on VQA v2.0 dataset

	Accuracy(%)			
Model	Yes/No	Number	Other	Overall
CNN-LSTM [31]	77.56	55.97	57.94	75
RNN-BiLSTM [32]	79.39	57.47	58.21	77.51
CAN [33]	80.23	58.71	58.94	78.34
BPI-MVQA [34]	79.79	58.52	58.72	78.02
K-PathVQA [35]	79.77	59.2	58.78	78.16
ATT-MR-WL (ours)	80.59	62.12	60.75	80.33

Library (PIL) for image processing and the NumPy library for image data manipulation and normalization. For the textual questions, we employed tokenization to split the questions into sequences of words or subwords, enabling the model to understand and process the textual data. Next, we converted the tokenized text into word vector representations, allowing the model to process the text data as input. To support these steps, we utilized natural language processing tools, including the Natural Language Toolkit (NLTK), and Gensim for tokenization and conversion of text into word vector representations. Following preprocessing, we conducted visualizations and statistical analyses of the preprocessed data to ensure data quality and consistency. These preprocessing steps and data analyses provided a reliable data foundation for model training and evaluation, ensuring the accuracy and performance of the model on our self-constructed dataset.

5 Results

5.1 Comparative Experiments and Result Analysis

This section verifies the effectiveness of the ATT-MR-WL model proposed in this article on the VQA v2.0 dataset by designing comparative experiments.

As shown in Table 1, there is a significant variation in the performance of different visual question answering (VQA) models on the VQA v2.0 dataset. In terms of accuracy for Yes/No questions, our method ATT-MR-WL leads with an accuracy of 80.59%, which is a 0.36 percentage point improvement over the closest competitor, CAN. For Number-type questions, our model demonstrates a clear advantage, achieving an accuracy of 62.12%, nearly 3 percentage points higher than the second-ranked K-PathVQA at 59.20%. In the category of Other questions, our method also surpasses other approaches with an accuracy rate of 60.75%, further proving the model's capability in comprehensively understanding nuanced queries. Looking at the overall

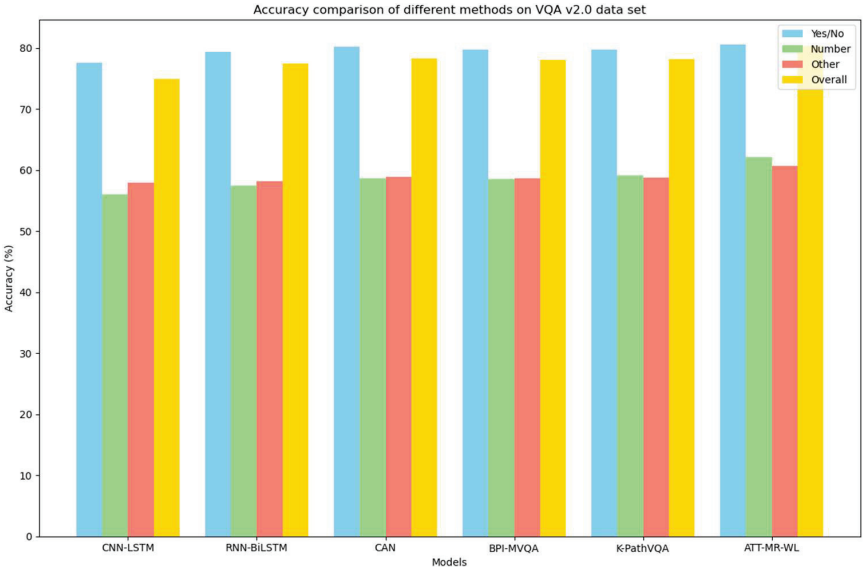

Figure 6 Accuracy comparison of different methods on the VQA v2.0 dataset.

Table 2 Performance of different iteration numbers in the modal information interaction module on the VQA v2.0 dataset

	Accuracy (%)			
Number of Iterations	Yes/No	Number	Other	Overall
1	79.85	59.8	69.27	78.02
2	80.3	61.59	70.49	79.02
3	80.37	61.9	70.53	80.11
4	80.36	60.65	69.66	78.51

accuracy, ATT-MR-WL reaches 80.33%, which is almost two percentage points ahead of the next best model, CAN, at 78.34%. In summary, the ATT-MR-WL model achieves the highest accuracy across all three categories – Yes/No, Number, and Other – and demonstrates a noticeable lead in overall performance. This underscores its potential and practical value in the realm of VQA tasks. Figure 6 visualizes the content of the table, intuitively highlighting the superiority of our method compared to other contenders, further emphasizing the significance of our approach in the field of visual question answering.

Based on the data in Table 2, we evaluated the performance of the multimodal information interaction module on the VQA v2.0 dataset with

different numbers of iterations. With one iteration, the model achieves an overall accuracy of 78.02%. Specifically, the accuracies for the "Number" and "Other" categories are 59.80% and 69.27%, respectively. This lower performance may be due to the model's insufficient use of multimodal information in a single iteration, leading to an incomplete understanding of both the question and the image. When the number of iterations is increased to two, the overall accuracy improves to 79.02%, with the "Number" and "Other" categories reaching 61.59% and 70.49% accuracy, respectively. This suggests that increasing the number of iterations helps the model better understand the relationship between the question and the image, thus improving performance. With three iterations, the model's overall accuracy further increases to 80.11%, reaching its peak. The accuracies for the "Number" and "Other" categories also increase to 61.90% and 70.53%, respectively. This implies that more iterations could further improve the model's performance, particularly with complex questions. However, the overall accuracy of the model slightly decreases to 78.51% when the number of iterations increases to four. This may be due to overfitting, leading to a decline in performance. In conclusion, based on the experimental results, the model performs best with three iterations, which is crucial for improving its performance. To provide a more visual representation of the data, we have plotted the data from the table into a line graph, as shown in Figure 7.

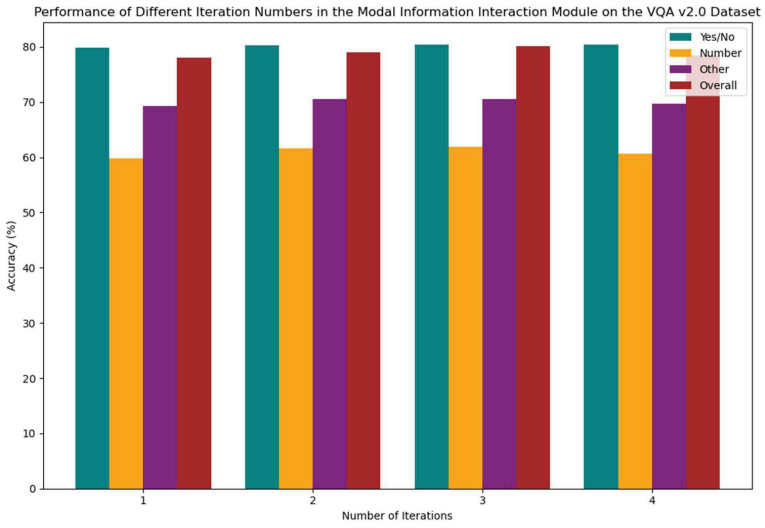

Figure 7 Performance of different iteration numbers in the modal information interaction module on the VQA v2.0 dataset.

Table 3 Accuracy comparison of different methods on the EM dataset

	Accuracy(%)			
Model	Yes/No	Number	Other	Overall
CNN-LSTM	77.82	56.23	68.2	75.26
RNN-BiLSTM	79.65	57.73	68.47	77.77
CAN	80.49	58.97	69.2	78.6
BPI-MVQA	80.05	58.78	69.08	78.28
K-PathVQA	79.91	59.46	69.04	78.42
ATT-MR-WL (ours)	80.85	62.38	71.01	79.59

5.2 Experimental Results on the EM Dataset

As depicted in Table 3, we present the accuracy comparison of different methods on our EM dataset. The table provides a comparison of model performance across various question types, including Yes/No, Number, and Other, as well as the overall accuracy. Our proposed ATT-MR-WL model stands out among the methods evaluated, achieving the highest overall accuracy of 79.59%. Notably, compared to the second-best performing model, our model exhibits a significant improvement of 1.25% in overall accuracy. Moreover, our model outperforms others in each specific question type: Yes/No (80.85%), Number (62.38%), and Other (71.01%). In summary, our ATT-MR-WL model demonstrates superior performance on our EM dataset, showcasing its effectiveness in addressing a variety of question types. Further details and visualizations of the results are provided in Figure 8.

In order to verify the Mask R-CNN, Word2Vec+LSTM and attention model designed in this article, this section conducts ablation experiments and analysis on this module. Table 4 shows the ablation experiment of ATT-MR-WL. Laboratory analysis will be conducted on the models with the multi-head attention module and the guided attention module removed respectively. Remove represents the removal of the multi-head attention module and the guided attention module from the ATT-MR-WL model.

Table 4 presents the results of the attention ablation experiment conducted on our EM dataset, where different components of the attention mechanism in our model were removed to evaluate their impact on performance. Our baseline model, ATT-MR-WL, achieves an overall accuracy of 80.05%. By removing the multi-head attention mechanism, the accuracy drops to 78.75%, indicating that multi-head attention contributes positively to the model's performance. Furthermore, when the guided attention mechanism is removed, the accuracy decreases even further to 75.08%, suggesting that guided attention also plays a crucial role in enhancing the model's accuracy.

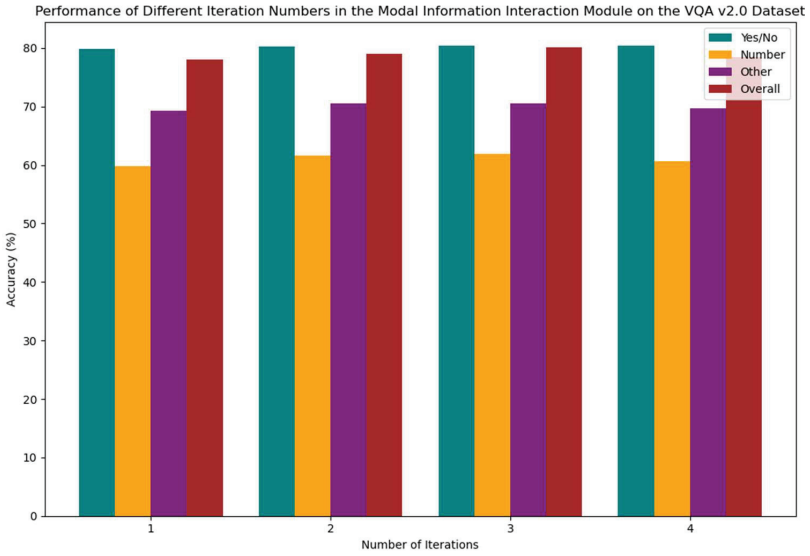

Figure 8 Accuracy comparison of different methods on EM dataset.

Table 4 Attention ablation experiment on EM dataset

	Accuracy(%)			
Model	Yes/No	Number	Other	Overall
ATT-MR-WL	80.54	65.58	77.69	80.05
Remove multi-head attention	76.94	63.44	73.59	78.75
Remove guided attention	75.29	62.94	71.48	75.08

Table 5 Ablation experiments of Mask R-CNN and Word2Vec+LSTM models

	Accuracy(%)			
Model	Yes/No	Number	Other	Overall
ATT-MR-WL	79.54	61.07	59.7	79.28
RetinaNet+WL	79.18	57.66	57.89	77.29
MR+Word2Vec+CNN	78.34	56.42	57.16	76.46

In conclusion, the attention ablation experiment demonstrates that both the multi-head attention and guided attention mechanisms significantly contribute to the overall performance of our model on the EM dataset. This highlights the importance of attention mechanisms in capturing relevant information and improving the model's ability to answer questions accurately.

As shown in Table 5, we conducted ablation experiments on the Mask R-CNN and Word2Vec+LSTM models to assess their importance for our

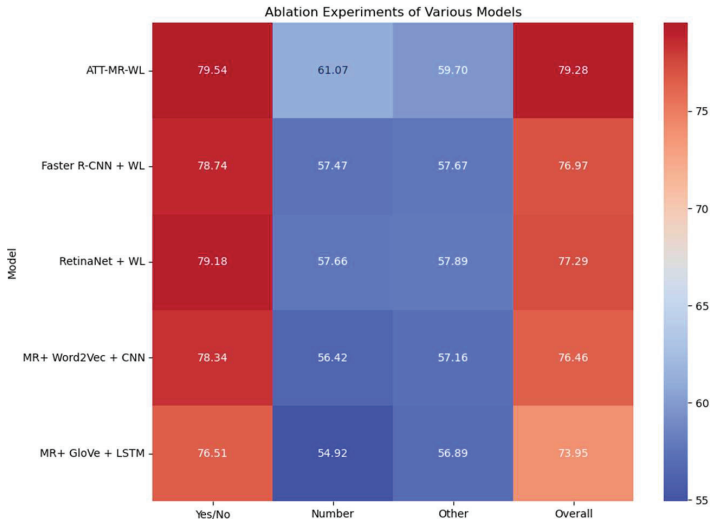

Figure 9 Ablation experiments of the Mask R-CNN and ord2Vec+LSTM models.

ATT-MR-WL model. We observe that in our ATT-MR-WL model, which includes all components, we achieved an accuracy of 79.54% for the "Yes/No" question type, significantly higher than the accuracies of other models. Particularly, compared to the Faster R-CNN + WL model with an accuracy of 78.74%, our model's accuracy increased by 3.0%. Our model also performs exceptionally well in other question types, such as achieving an accuracy of 71.07% for "Number" questions, while the highest accuracy achieved using other models was 57.66%. Additionally, we note that using the MR+Word2Vec+CNN and MR+GloVe+LSTM models resulted in accuracies of 78.34% and 76.51%, respectively, lower than our ATT-MR-WL model. This indicates the importance of the Word2Vec+LSTM models in our architecture, contributing to the improvement in accuracy. In summary, our ATT-MR-WL model demonstrates significant advantages in the "Yes/No" question type and overall accuracy, highlighting the importance of the Mask R-CNN and Word2Vec+LSTM models in our architecture. Figure 9 visualizes the tabulated data.

6 Visual Result Analysis

Upon evaluating the scenarios in Figure 10, the system exhibits a high degree of accuracy in identifying and interpreting structured educational

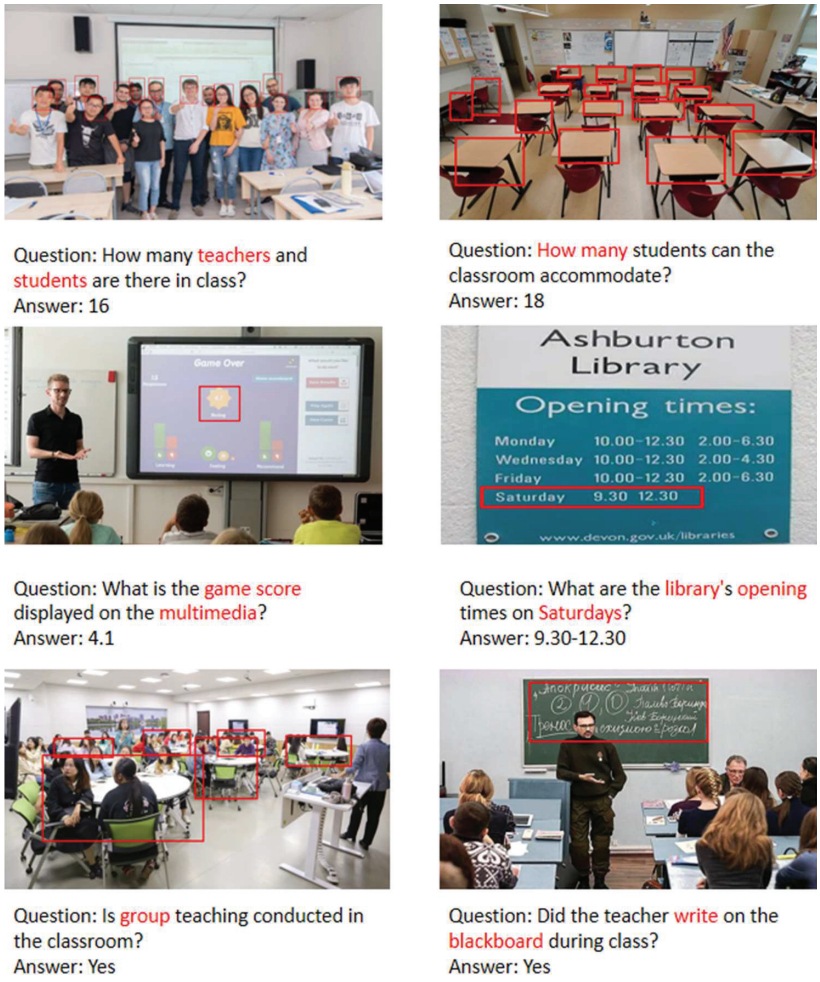

Figure 10 Model visualization results of model ATT-MR-WL on the EM dataset.

data. It not only quantifies classroom participants but also assesses room capacities and interprets written information such as library hours and game scores from digital displays. Additionally, it accurately detects group teaching dynamics and instances of writing on the blackboard. These competencies illustrate the model's versatility in processing varied visual inputs, signifying its comprehensive training and adeptness in visual data interpretation.

In the educational domain, the model enhances administrative efficiency by improving resource allocation and management. Its ability to analyze

classroom arrangements and interpret image-based information aids in space management and schedule optimization. In conclusion, the results validate the ATT-MR-WL model as an effective analytical tool for educational settings, aligning with the practical needs of academic administration and suggesting its capacity to streamline educational operations.

7 Conclusion

In this study, we proposed an intelligent question answering system model ATT-MR-WL based on multi-modal information fusion, and conducted experimental verification in the education field. By conducting experiments on the VQA v2.0 dataset and the self-built education management dataset (EM dataset), we verified the performance of the model under different question types. Experimental results demonstrate that the ATT-MR-WL model has strong accuracy and robustness in answering various types of questions in educational contexts, especially in handling both image and text-based queries effectively. This model achieves in-depth understanding and accurate answers to questions by effectively combining image and text information, providing intelligent management support for education. However, our model still has some shortcomings. First, for certain complex question types, the model may not fully comprehend the context, leading to incorrect answers. Although we considered multi-modal information fusion and attention mechanisms when designing the model, there are still some problems in specific scenarios that are difficult to handle. Second, when processing large-scale datasets, the model encounters efficiency and performance challenges. In real-time scenarios, the response speed may not meet user requirements, necessitating further optimization and enhancement.

Looking to the future, we will continue to refine and optimize our models to handle more complex educational challenges. First, we will further improve the deep learning capabilities of the model and strengthen the understanding and analysis of text and image information to improve the accuracy and robustness of the model. Additionally, we plan to explore more efficient model architectures and algorithms to improve processing speed and performance in large-scale data environments. We also aim to investigate the practical applications of the model in educational management, such as providing personalized learning support and assisting in decision-making. This will help us assess the model's real-world impact on education systems. Finally, we will continue to monitor developments in the education field

and upgrade our intelligent question-answering system to contribute to the ongoing informatization of educational management.

Funding

Hebei Provincial Department of Education, Research and Practice Project on Higher Education Teaching Reform in Hebei Province (2023GJJG225).

References

[1] Agbodike, O., C.-H. Huang, and J. Chen. *Cognitive attention network (can) for text and image multimodal visual dialog systems*. in *2020 6th International Conference on Applied System Innovation (ICASI)*. 2020. IEEE.

[2] Arif, A.M., N. Nurdin, and E. Elya, *Character Education Management at Islamic Grassroot Education: The Integration of Local Social and Wisdom Values.* Al-Tanzim: Jurnal Manajemen Pendidikan Islam, 2023. **7**(2): p. 435–450.

[3] Azmat, F., A. Jain, and B. Sridharan, *Responsible management education in business schools: Are we there yet?* Journal of Business Research, 2023. **157**: p. 113518.

[4] Zou, H.a.Z., Mafu and Farzamkia, Saleh and Huang, Alex Q, *Simplified Fixed Frequency Phase Shift Modulation for A Novel Single-Stage Single Phase Series-Resonant AC-DC Converter*. 2024 IEEE Applied Power Electronics Conference and Exposition (APEC). 2024: IEEE.

[5] Yadav, A. and A. Prakash, *Factors influencing sustainable development integration in management education: An Empirical Assessment of management education institutions in India.* The International Journal of Management Education, 2022. **20**(1): p. 100604.

[6] Gupta, A., et al., *Role of cloud computing in management and education.* Materials Today: Proceedings, 2023. **80**: p. 3726–3729.

[7] Zhou, Y.a.W., Zhaoqi and Zheng, Shirong and Zhou, Li and Dai, Lu and Luo, Hao and Zhang, Zecheng and Sui, Mingxiu, *Optimization of automated garbage recognition model based on ResNet-50 and weakly supervised CNN for sustainable urban development.* Alexandria Engineering Journal, 2024. **108**: p. 415–427.

[8] Nassiri, K. and M. Akhloufi, *Transformer models used for text-based question answering systems.* Applied Intelligence, 2023. **53**(9): p. 10602-10635.

[9] Do, P. and T.H. Phan, *Developing a BERT based triple classification model using knowledge graph embedding for question answering system.* Applied Intelligence, 2022. **52**(1): p. 636-651.

[10] Cho, J.W., et al. *Generative bias for robust visual question answering*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[11] Yusuf, A.A., F. Chong, and M. Xianling, *An analysis of graph convolutional networks and recent datasets for visual question answering.* Artificial Intelligence Review, 2022. **55**(8): p. 6277-6300.

[12] Nguyen, N.H., et al., *Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese.* Information Fusion, 2023. **100**: p. 101868.

[13] Pan, H., et al., *AMAM: an attention-based multimodal alignment model for medical visual question answering.* Knowledge-Based Systems, 2022. **255**: p. 109763.

[14] Wan, Q.a.Z., Zecheng and Jiang, Liheng and Wang, Zhaoqi and Zhou, Yan, *Image anomaly detection and prediction scheme based on SSA optimized ResNet50-BiGRU model.* arXiv preprint arXiv:2406.13987, 2024.

[15] Yang, T., et al., *Application of question answering systems for intelligent agriculture production and sustainable management: A review.* Resources, Conservation and Recycling, 2024. **204**: p. 107497.

[16] Zou, H.a.Y., Ruiyang and Anand, Rishab and Tong, Junhong and Huang, Alex Q, *A gan variable-frequency series resonant dual-active-bridge bidirectional ac-dc converter for battery energy storage system*. 2023 IEEE Applied Power Electronics Conference and Exposition (APEC). 2023: IEEE. 150–157.

[17] Ruiz, E., M.I. Torres, and A. del Pozo, *Question answering models for human–machine interaction in the manufacturing industry.* Computers in Industry, 2023. **151**: p. 103988.

[18] Peng, X.a.X., Qiming and Feng, Zheng and Zhao, Haopeng and Tan, Lianghao and Zhou, Yan and Zhang, Zecheng and Gong, Chenwei and Zheng, Yingqiao, *Automatic News Generation and Fact-Checking System Based on Language Processing.* arXiv preprint arXiv:2405.10492, 2024.

[19] Therasa, M. and G. Mathivanan, *ARNN-QA: Adaptive Recurrent Neural Network with feature optimization for incremental learning-based Question Answering system.* Applied Soft Computing, 2022. **124**: p. 109029.

[20] Echegoyen, G., Á. Rodrigo, and A. Peñas, *Study of a lifelong learning scenario for question answering.* Expert Systems with Applications, 2022. **209**: p. 118271.

[21] Tian, S., et al., *Continuous transfer of neural network representational similarity for incremental learning.* Neurocomputing, 2023. **545**: p. 126300.

[22] Wang, J., et al., *Towards robust lidar-camera fusion in bev space via mutual deformable attention and temporal aggregation.* IEEE Transactions on Circuits and Systems for Video Technology, 2024.

[23] Ran, H., et al., *Learning optimal inter-class margin adaptively for few-shot class-incremental learning via neural collapse-based meta-learning.* Information Processing & Management, 2024. **61**(3): p. 103664.

[24] He, H., et al., *Mask R-CNN based automated identification and extraction of oil well sites.* International Journal of Applied Earth Observation and Geoinformation, 2022. **112**: p. 102875.

[25] Bi, X., et al., *Iemask r-cnn: Information-enhanced mask r-cnn.* IEEE Transactions on Big Data, 2022. **9**(2): p. 688-700.

[26] Johnson, S.J., M.R. Murty, and I. Navakanth, *A detailed review on word embedding techniques with emphasis on word2vec.* Multimedia Tools and Applications, 2024. **83**(13): p. 37979-38007.

[27] Al Hamoud, A., A. Hoenig, and K. Roy, *Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(10): p. 7974-7987.

[28] Liu, Y., et al., *Analysis of the causes of inferiority feelings based on social media data with Word2Vec.* Scientific Reports, 2022. **12**(1): p. 5218.

[29] Ding, M., et al. *Davit: Dual attention vision transformers*. in *European conference on computer vision*. 2022. Springer.

[30] Patro, B.N. and V.P. Namboodiri, *Explanation vs. attention: A two-player game to obtain attention for VQA and visual dialog.* Pattern Recognition, 2022. **132**: p. 108898.

[31] Faseeh, M., et al., *Enhancing User Experience on Q&A Platforms: Measuring Text Similarity based on Hybrid CNN-LSTM Model for Efficient Duplicate Question Detection.* IEEE Access, 2024.

[32] Sangeetha, J. and U. Kumaran, *A hybrid optimization algorithm using BiLSTM structure for sentiment analysis.* Measurement: Sensors, 2023. **25**: p. 100619.
[33] Daniels, J. and C.P. Bailey. *Reconstruction and super-resolution of land surface temperature using an attention-enhanced cnn architecture*. in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023. IEEE.
[34] Liu, S., et al., *BPI-MVQA: a bi-branch model for medical visual question answering.* BMC Medical Imaging, 2022. **22**(1): p. 79.
[35] Naseem, U., et al., *K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering.* IEEE Journal of Biomedical and Health Informatics, 2023.

Biography

Ying Ba was born in Hebei, China, in 1983. From 2002 to 2006, she studied at Northeastern University and received her Bachelor's degree in 2006. From 2006 to 2013, she studied at Harbin Institute of Technology and received her Doctor's degree in 2013. She has published a total of 12 papers. Her research interests are included teaching reform and educational Management.