Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 22 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i22.44103

PAPER

Research on the Teaching Method of College Students' Education Based on Visual Question Answering Technology

Fang Lin(✉)

Training Center, Zhengzhou Shengda University, Zhengzhou, China

Fang_Lin2023@outlook.com

ABSTRACT

With the changes of the times, visual question answering (VQA) technology has gradually been widely used in many fields, such as intelligent bionic robots, student learning and education, and visual assistance for the disabled. In view of this, the study proposes a method of college student education and teaching system based on VQA technology. First, a multi-scale fusion feature method is proposed for the representation of image features, and then an improved mixed attention mechanism is proposed based on this. Solve the network noise problem; finally, introduce the VQA system. The results show that the research method tends to a stable state when the iteration reaches the 28th time, and the corresponding loss function converges to 2.38, which is 0.5% lower than the traditional model; in the application effect analysis, on the data set test-dev, the research model The accuracies of the four types of questions about "total, whether, count, and others" are 69.94%, 86.21%, 50.16%, and 59.61%, respectively; among the example outputs of the VQA dataset, the research model can accurately infer the categories of each target object. Analyzing the change of students' comprehensive literacy, when the teaching times of the constructed model reach six times, taking students' enthusiasm as an example, the score ratios of innovation literacy in subjects such as English, mathematics, and Chinese are 96.32, 93.14, and 88.56. The above results all show that the accuracy of the research method is better, and it has better feasibility and effectiveness in the field of education and the teaching of students.

KEYWORDS

visual question answering (VQA), education and teaching, college students

1 INTRODUCTION

With the rapid development of artificial intelligence technology and the further popularization of smart devices and multimedia applications, people's production and lives have undergone earth-shaking changes. Entering the new century,

Lin, F. (2023). Research on the Teaching Method of College Students' Education Based on Visual Question Answering Technology. *International Journal of Emerging Technologies in Learning (iJET)*, 18(22), pp. 167–182. https://doi.org/10.3991/ijet.v18i22.44103

Article submitted 2023-08-16. Revision uploaded 2023-09-29. Final acceptance 2023-09-29.

more and more intelligent technologies have become research hotspots in the field of education and teaching and play an important role in students' daily learning. Scholars have begun to study the expression and interaction of cross-media data, and visual question answering (VQA) technology is one of the hot issues [1–2]. VQA belongs to the intersection of computer vision and natural language processing. It combines the characteristics of computer vision and natural language and needs to process the given image text and questions to finally generate the correct natural language answer [3–4]. However, many studies have shown that most of the existing VQA technology models are top-down visual answers, which seriously ignore the content expression of the image text itself, which seriously leads to the loss of image features. For example, the traditional VQA technology model ignores the dynamic relationship between semantic information and the rich spatial structure between different visual areas under dual modalities. Moreover, the calculation of the results by the traditional method is relatively simple, ignoring the factors that affect the specific teaching effect, which leads to the decline of the quality of education and teaching [5–6]. Therefore, in the problem of education and teaching of college students, to improve the effective recognition ability of text images and improve students' learning enthusiasm, the research proposes a teaching method for college students based on VQA technology. For the problem of image feature representation, the research uses multi-scale features to improve the information and considering that the complete image features will have information redundancy such as noise introduced by the network, the study improves the attention mechanism by strengthening the text sentence. The structure achieves the purpose of improving the overall performance of the model. Finally, the traditional VQA technology model is improved by integrating multi-scale and improved mixed attention mechanisms. The obtained method can avoid problems such as insufficient detail information and finally obtain a positive application effect.

2 RELATED WORKS

The continuous advancement of science and technology has made artificial intelligence and multimedia widely used in the education and teaching of college students. In recent years, a variety of methods and VQA technology have crossed paths, and many valuable results have been obtained in education. Scholars such as Liu C proposed a method based on BP neural networks and stress testing for the quality evaluation of postgraduate education. In this method, the four dimensions of teaching attitude, content, method, and the teacher's basic characteristics are used as variables to construct an index library. Using scenario analysis to explore the background of presupposed educational quality, it was found that the model constructed can provide certain theoretical and empirical support for future research [7]. To improve the teaching quality of physical education, Han proposes to use an artificial neural network to build a quality evaluation model of physical education in colleges and universities. The model quantifies the evaluation indicators during the experiment and inputs them into the model as actual data to assist in data analysis with relevant software. The data show that the model can effectively improve the evaluation quality, avoid the influence of subjective factors in the evaluation process, and obtain a relatively satisfactory evaluation result [8]. To overcome the shortcomings of traditional teaching methods, Naim proposed using convolutional neural network (CNN) to build an educational learning model. The study uses an applied lightweight model to identify student engagement in distance

education learning and improves the utility of the method through multiple trials. The results show that, compared with the traditional model, the research model can effectively improve students' classroom participation, and it is also the most suitable application model for the real teaching environment [9]. For the quality evaluation efficiency of scientific research personnel in universities, researchers such as H combined BP neural networks and fuzzy theory to build a talent evaluation model. The model uses the two-level index evaluation system to determine the weight of the index, then normalizes all samples and uses fuzzy theory to fuzzy-determine the relationship between each judgment domain. The results show that the model has high feasibility and can provide a reference for subsequent related experiments [10]. Li proposed to build an English learning system model based on the improved fuzzy-layered neural network to improve the ability to learn English independently. This model is actually a computer-aided learning platform, which can provide more teaching content and promote real-time monitoring of education. The results show that this model is an interactive education system model, which can significantly improve the efficiency of teachers' classroom teaching and the ability of students to learn independently [11].

Visual question answering technology has been widely concerned and applied in various fields. Zheng S. et al. proposed an end-to-end deep learning framework model to identify novel drug-protein interactions. In the experiment, the CNN training system was introduced, and the self-attention mechanism was added to automatically extract language features; the process followed the visual interrogation mode. The results show that the model has excellent performance [12]. Li and other scholars proposed a new VQA system technology to make better use of the information in the visual language task. The model is able to highlight conceptually related regions, reducing the gap between vision and language. The application performance shows that this method has significantly excellent performance and can effectively improve the efficiency of image recognition [13]. Zhang's team proposes an alternating attention network to deal with visual language problems in multimedia intelligence, which can fully pay attention to frame regions, words, and video frames of images in multiple rounds. Experimental results show that this method produces better video frame parameters and outperforms traditional models [14]. To solve the situation where the image and language domains are limited by factors, Cao Q proposed a new neural network model. This model also known as the Parsing tree guided reasoning network (PTGRN), can conduct a more comprehensive analysis of the problem. PTGRN interprets language domain questions by constructing a VQA system. The results show that this method is superior to VQA and has outstanding language interpretation performance [15]. Gao and other scholars proposed an end-to-end structured multi-modal attention neural network system to improve the system's ability to read text and visual reasoning. The method uses structural graphs to describe image content, and designs a multimodal graph attention network to infer text language. The results show that the research model has excellent performance in text visual reasoning, and has a strong task processing ability for datasets [16].

To sum up, VQA technology has produced relatively sufficient basic theories and references in many fields. Most of these studies focus on the convenience of neural network systems for text information and visual reasoning, and few studies use basic methods such as attention mechanisms and multi-scale to improve VQA system technology. In view of this, the study proposes a teaching method for college students based on VQA technology. During the experiment, the multi-scale fusion and attention mechanism were used to improve the VQA technology. It is urgent to build

a model that can further promote the quality of education and teaching of college students in my country and the enthusiasm of students for independent learning.

3 RESEARCH ON VQA SYSTEM DESIGN FOR COLLEGE STUDENTS' EDUCATION AND TEACHING

3.1 Image feature extraction of VQA system based on multi-scale fusion deep learning

Different features in the neural network have different depths and corresponding meanings, which have strong generalizations. To solve the problem of incomplete representation of the output feature information in the process of task classification and student target detection, the multi-scale feature fusion method was introduced in the image feature extraction of the VQA system to promote the education and teaching of college students, and the analysis was carried out based on existing research [17–18]. First, image features are extracted using deep learning, as shown in Figure 1.

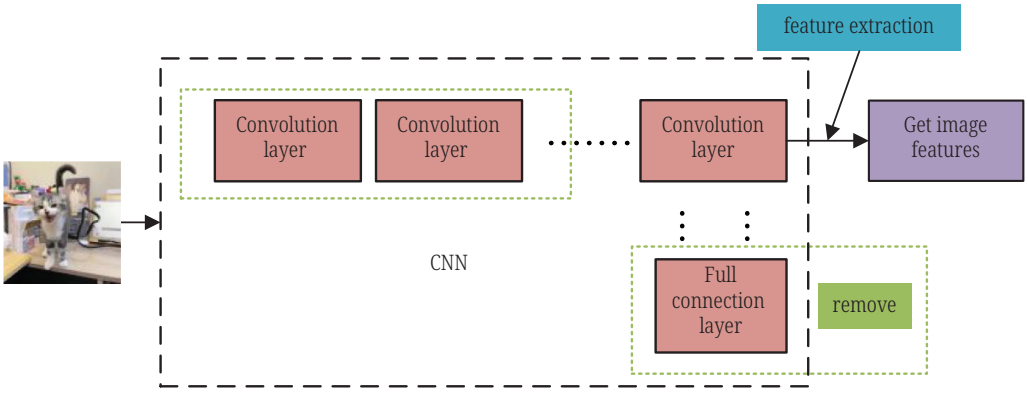


Fig. 1. DL-image feature extraction

In Figure 1, the fully connected layer in the deep neural network is removed, and the output of the CNN layer is used as the required image features. However, when there is a large difference between the original data set and the training data set, transfer learning is usually used to fix the fixed part of the neural network parameters and fine-tune the remaining neural network to obtain the network parameters required for the experiment. Finally, the process method of removing the fully connected layer and directly extracting the output of the convolutional layer can obtain the image features required for the final experiment. The traditional method uses CNN to process image data in a grid, and research uses a recurrent neural network (RNN) to extract text information from image features. The system-state transition expression is shown in equation (1).

$$\begin{cases} h_i^{(t)} = \tanh(s_i^{(t)}) * q_i^{(t)} \\ s^{(t)} = f(s^{(t-1)}, x^{(t)}; \theta) \end{cases} \tag{1}$$

Equation (1), $s^{(t)}$ represents the stated performance of the $x^{(t)}$ system at a time and t represents the driving signal of the external environment. $f(\cdot)$ Indicates that the current state depends on the past state and the current signal. However, after

many iterations of RNN, the gradient tends to disappear or explode. The study then proposes a long-short-term memory network (LSTM) and a gated recurrent unit network (GRU) to solve this problem. See equation (2) for the specific operation.

$$\begin{cases} f_i^{(t)} = \sigma\left(b_i^f + \sum_j A_i^f, {}_j x_j^{(t)} + \sum_j B_i^f, {}_j h_j^{(t-1)} \right) \\ g_i^{(t)} = \sigma\left(b_i^g + \sum_j A_i^g, {}_j x_j^{(t)} + \sum_j B_i^g, {}_j h_j^{(t-1)} \right) \\ s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j A_i, {}_j x_j^{(t)} + \sum_j B_i, {}_j h_j^{(t-1)} \right) \\ h^{(t)} = \tanh(s_i^{(t)}) \end{cases} \tag{2}$$

In equation (2), $x^{(t)}$ represents t the input at the current moment; $h^{(t)}$ represents the current hidden layer; $f_i^{(t)}$ represents the forget gate; A represents the input weight; and B represents the forget gate weight. Due to the complex structure of LSTM, the research uses GRU to simplify the overall feature extraction process and assist in the combination of question text features and image spatial information. In the process of extracting text feature information, to obtain convolution kernels and word vectors under the three scale windows, see equation (3).

$$\bar{v}_{a,t}^p = conv\left(W_c^a v_{t,t+a-1}^w \right), a \in \{1,2,3\} \tag{3}$$

In equation (3), it a represents the size of the convolution kernel, corresponding to three different scales, which satisfies the feature extraction of "phrases" of different lengths; $v^w = \left\{ v_1^w, v_2^w, \cdots, v_N^w \right\}$ represents the sentence contains N a word; N represents the sentence scale; W represents the weight parameter. After the convolution operation of equation (3), each word can be represented by features at three scales. Then combine the outputs under all windows and use the maximum pooling process to obtain the "phrase" feature, see equation (4).

$$v_t^p = \max\left(\bar{v}_{1,t}^p, \bar{v}_{2,t}^p, \bar{v}_{3,t}^p \right), t \in \{1,2,\cdots,N\} \tag{4}$$

In equation (4), t represents the phrase-sentence scale. For text-scale features, the research uses RNN to extract the features after maximum pooling, and N iterates to obtain sentence features. The skip – thought model is inserted in the process of feature acquisition, and the probability calculation method of similarity between words is extended to sentences. The original sentence phrase is given in the entire model (s_{i-1}, s_i, s_{i+1}), which s represents i the moment sentence and the adjacent context sentence; represents the x_i^t word embedding of N the first word in the entire text sentence; the sentence length is set to. The model experiment process includes encoding and decoding operations. Apply the model to GRU-RNN, see equation (5).

$$\begin{cases} r^{(t)} = \sigma(A_r x^{(t)} + B_r h^{(t-1)}) \\ z^{(t)} = \sigma(A_z x^{(t)} + B_z h^{(t-1)}) \\ \bar{h}^{(t)} = \tanh\left[Ax^{(t)} + B(r^{(t)} \otimes h^{(t-1)}) \right] \\ h^{(t)} = (1 - z^{(t)}) \otimes h^{(t-1)} + z^{(t)} \otimes \bar{h}^{(t)} \end{cases} \tag{5}$$

In equation (5), it $h_i^{(t)}$ represents the hidden feature i of the first word of the first sentence t; correspondingly h_i^N represents the hidden feature of the entire sentence; ⊗ represents bitwise multiplication. The decoding part is also expanded by GRU-RNN. The difference is that the result of the encoder is inserted into the input and controlled together. The expression is shown in equation (6).

$$\begin{cases} r^{(t)} = \sigma(A_r x^{(t)} + B_r h^{(t-1)} + C_r H_i) \\ z^{(t)} = \sigma(A_z x^{(t)} + B_z h^{(t-1)} + C_z H_i) \\ \bar{h}^{(t)} = \tanh\left[Ax^{(t)} + B(r^{(t)} \otimes h^{(t-1)}) + Ch_i \right] \\ h_{i+1}^{(t)} = (1 - z^{(t)}) \otimes h^{(t-1)} + z^{(t)} \otimes \bar{h}^{(t)} \end{cases} \tag{6}$$

In equation (6), h_i represents the output of the encoder; h_{i+1} represents t the hidden layer at the moment. The multi-scale feature extraction and fusion methods are introduced into the VQA system. First, word embedding is performed on the text to map it to a continuous numerical feature space, and then a continuous word vector is obtained. Use the skip-thought model to extract sentence features and phrase feature sentences; use different scales for convolution operations; and use the maximum pooling output as phrase features [19–20]. The feature extraction operation is shown in Figure 2.

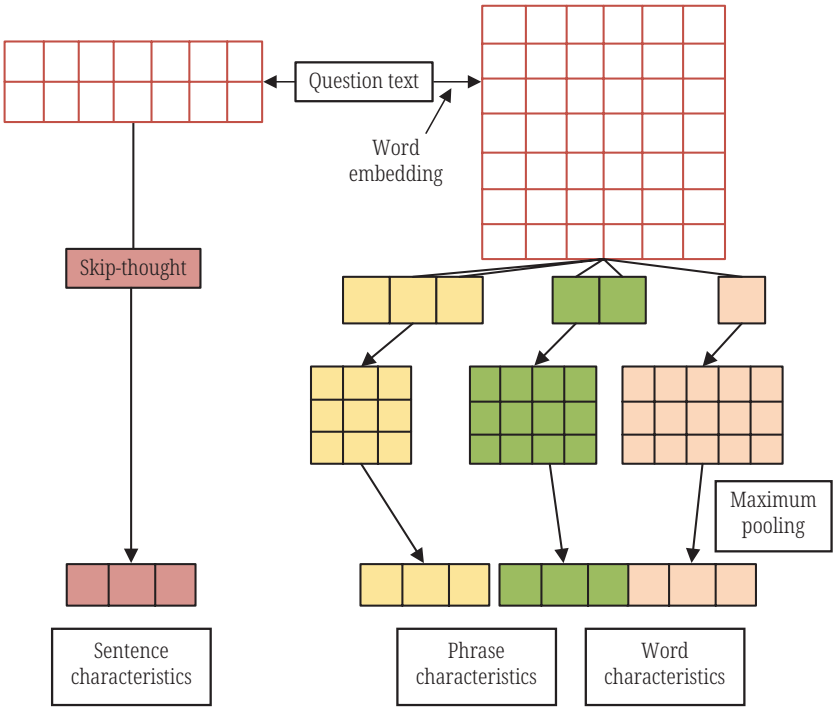
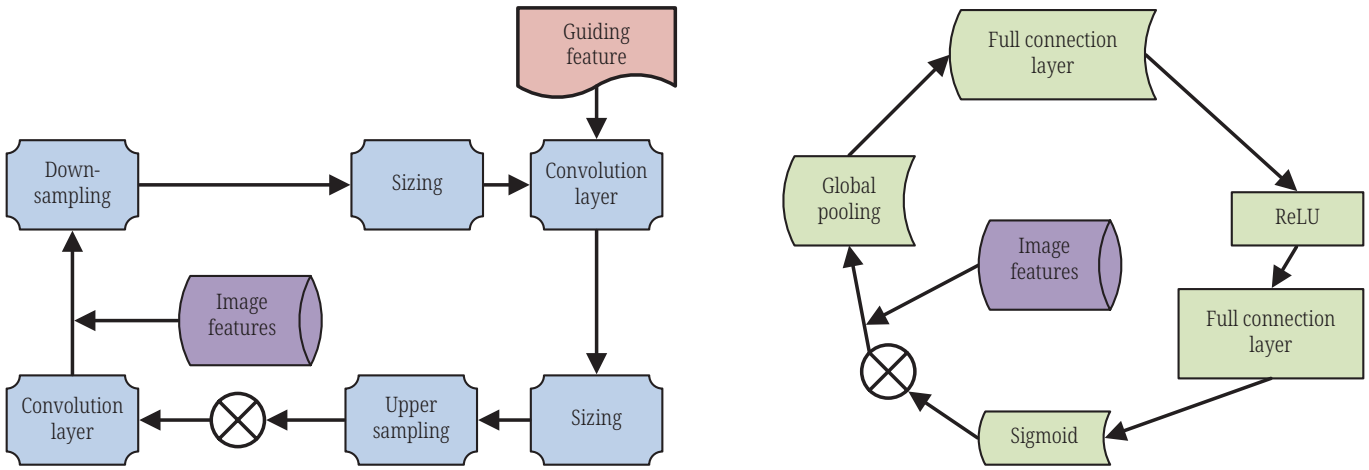


Fig. 2. Operation diagram of multi-scale image feature extraction

3.2 Improvement of VQA system based on attention mechanism

In the computer vision language processing system, the addition of the attention mechanism is beneficial to improving the accuracy of the results. When human beings observe an image, the distribution of attention to all the

information on the image is not uniform, and most of the attention is focused on the part that is most interesting to humans, and there is a significant shift in the state relationship between different contents. Some scholars have proposed an attention mechanism research model based on RNN. This model can make the research focus on the area of human interest through the attention mechanism, and the current learning state will be affected by the state of the previous step, which is beneficial to reduce the calculation process. The complexity and storage parameters step in. The visual attention mechanism is widely used in VQA system tasks, which can use the extracted semantic features as guiding features for the attention weight calculation of image features. The attention mechanism can be divided into spatial domain attention mechanisms and channel attention mechanisms [21–22]. The basic operation flow of the two attention mechanisms is shown in Figure 3.



a) Spatial attention mechanism **(b) Channel domain attention mechanism**

Fig. 3. Basic operation flow of the attention mechanism

In the spatial domain attention mechanism, the image features are first down-sampled, and the global image semantic information obtained is shown in equation (7).

$$v_I = \max pooling\ (v_i) \tag{7}$$

In equation (7), v_i represents the image feature and v_I represents the global image semantic information. Then, the maximum pooling feature is obtained through LSTM, and the attention weight is calculated, and the image text feature is processed and input to the softmax layer, as shown in equation (8).

$$\begin{cases} h_z = \tanh(W_{I,Z} v_I \oplus (W_{Q,A} v_Q + b_A)) \\ p_I = soft\max(W_p h_z + b_p) \end{cases} \tag{8}$$

In equation (8), v_Q represents the feature code and P_I represents the attention vector, which includes the attention weights of all image regions corresponding to the feature code. Adjust the attention weight layer so that the resulting weights are restored to the image feature size. In the operation of the channel attention mechanism, it is assumed that I it is an image feature (among them $I \in R^{H' \times W' \times C'}, V \in R^{H \times W \times C}$),

and multiple channel features are obtained through multiple convolution kernel filtering operations V as represented in equation (9)

$$v_k = f_k * I = \sum_{s=1}^{C'} f_k^s * i^s \tag{9}$$

In equation (9), the obtained feature $V = \{v_1, v_2, ..., v_C\}$; the image I has C' a channel; $*$ represents the convolution operation; f_k^s represents the corresponding first s channel. The obtained features are independent of each other. To explore the feature relationships of different channels, the study uses global pooling to compress features and performs averaging operations from different dimensions of feature length and width to generate new features. And adding an activation function to generalize and reduce the image features. See equation (10) for details.

$$\begin{cases} z_k = (1/HW)\sum_{i=1}^{H}\sum_{j=1}^{W} v_k(i,j) \\ P_c = \sigma(W_2\delta(W_1 z)) \end{cases} \tag{10}$$

In equation (10), RELU is selected as the activation function. σ represents an activation function; W_1 represents a fully-connected layer and contains an attenuation index r, the purpose of which is to reduce the dimensionality and generalize the original features. W_2 represents the final output layer, the purpose of which is to rearrange the dimensions introduced into the required dimensions when combining features. After getting all the attention weights, blend the image features and refine the weighted blend attention. To avoid the information loss after the attention mechanism and iterative transformation of features, the research adds the original features of the image according to a certain proportion, and the obtained image features are H_{Att}. Refer equation (11).

$$H_{Att} = A(P_I + P_c + b) * H_I \tag{11}$$

In equation (11), P_I represents the spatial domain attention mechanism; P_c represents the channel attention mechanism; and b represents the small weight. The resulting attention mechanism module is shown in Figure 4.

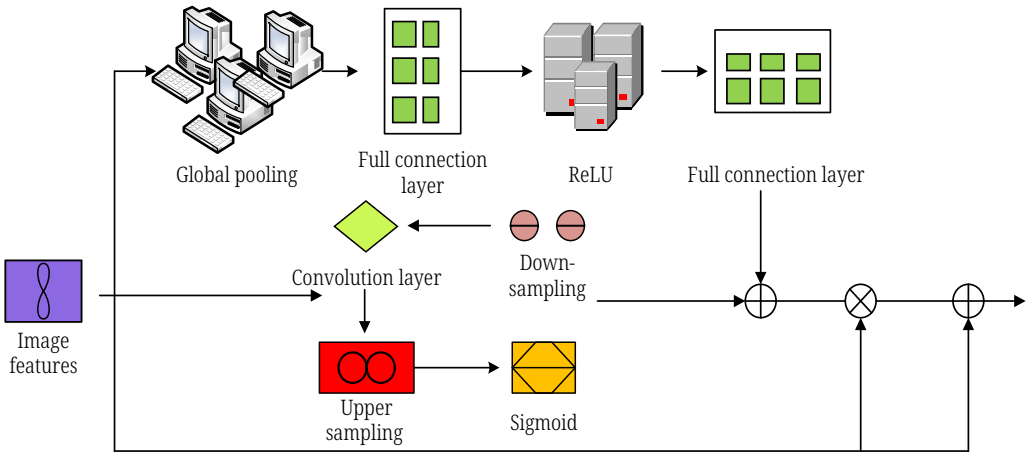


Fig. 4. Mixed attention mechanism module

In the process of improving the overall education and teaching method, because the traditional method does not consider the problem of language and text, the research proposes to add the text attention mechanism to the VQA technology system to solve this problem. After the features are extracted, the attention mechanism is inserted for weighting. Different image features use different attention mechanisms. The improved VQA technology system proposed in the study mainly includes feature extraction of image objects and feature extraction of natural language. The expression of multi-scale feature extraction for images and texts is shown in equation (12).

$$\begin{cases} V_I = \mathrm{Re}\,sNet_{conv3,conv5}(x) \\ Q = Conv_{N-d}(q) \oplus skip - thougut(q), N = 1,2,3 \end{cases} \tag{12}$$

In equation (12), V_I represents multi-scale image features; Q represents multi-scale text features; and $\oplus$ represents the splicing method between dimensions. By using the mixed attention mechanism to integrate the VQA technology system, the image feature extraction process is performed on the spatial and the channel domains. Softmax is used to normalize the data to obtain attention weights. At the same time, the channel attention mechanism is used to average all the hierarchical features of the image, and the extracted student image features are weighted on the channel [23]. The processed text image features are used as the guiding features of the channel attention mechanism, and the attention weights and VQA image features are calculated as shown in equation (13).

$$\begin{cases} z_t = \left(1/H \cdot W\right) \sum_{i=1}^{H} \sum_{j=1}^{W} V_I(i,j) \\ h_c = \tanh(W_{I,C} z_I \otimes (W_{I,Q} v_Q + b_q)) \\ P_c = soft\max(W_c h_c + b_c) \\ H_{Att} = A(P_I + P_c + b) * H_I \end{cases} \tag{13}$$

In equation (13), z_t, h_c, P_c represents the attention weight and H_t represents the image feature. In the end-to-end VQA system, the multimodal fusion method is used to fuse the image and text features, and the fusion features are obtained as shown in equation (14).

$$H = MLB\,(A_{tt}, Q) \tag{14}$$

In equation (14), H represents the fusion feature and MLB represents the multimodal feature fusion method. The ultimate goal of carrying out education and teaching for college students is to use the network to monitor students in real time. The process of VQA technology monitoring students' learning can be regarded as a classification task. The task employs a shallow classification network to generate answers. The fused features are classified and trained, and finally, the softmax layer is used to obtain the answer probability as seen in equation (15).

$$P = soft\ \max\,(CNN(H)) \tag{15}$$

In equation (15), P represents the generated answer probability. The final improved VQA system model is illustrated in Figure 5.

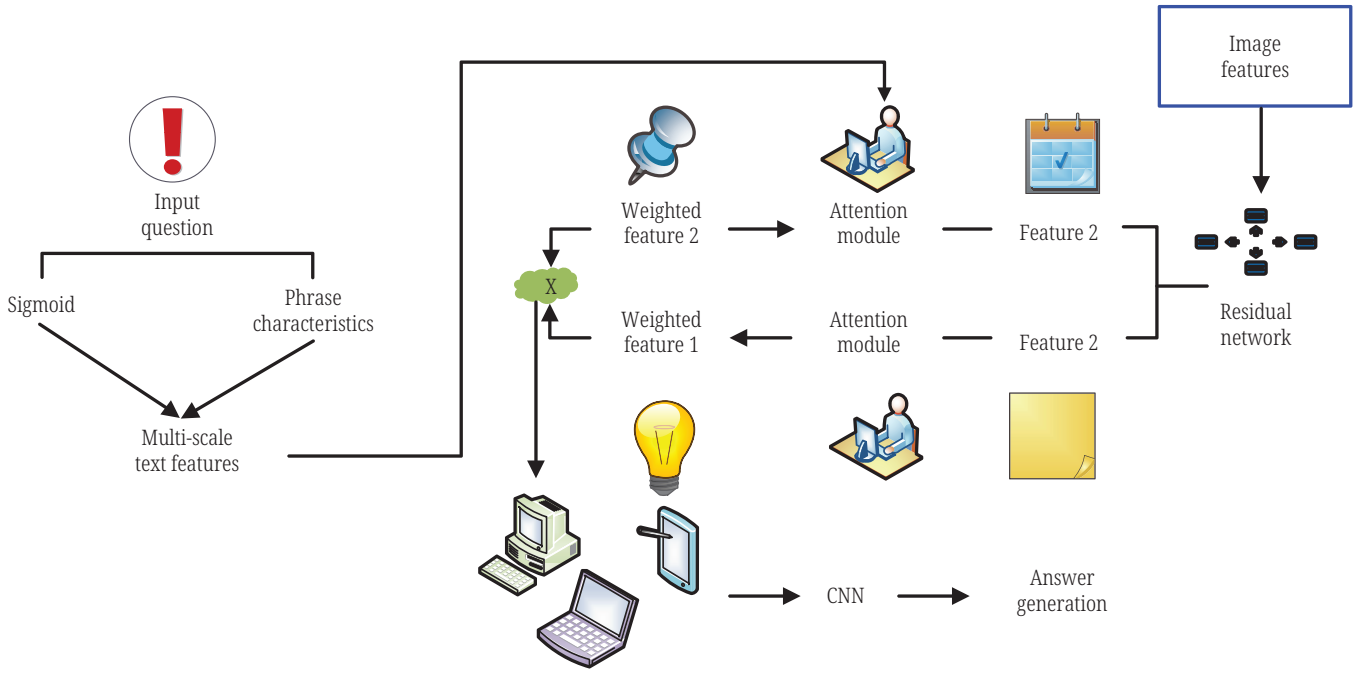


Fig. 5. Improved VQA system model

4 PERFORMANCE AND APPLICATION EFFECT OF EDUCATION TEACHING MODEL UNDER VQA TECHNOLOGY

4.1 Perform performance simulation test on the built model

The research selects top-down attention question answering model (BUTD) and bilinear attention question answering model (BAN) as baseline models. The loss functions of the research method and other VQA models are compared using the VQA dataset. See Figure 6.

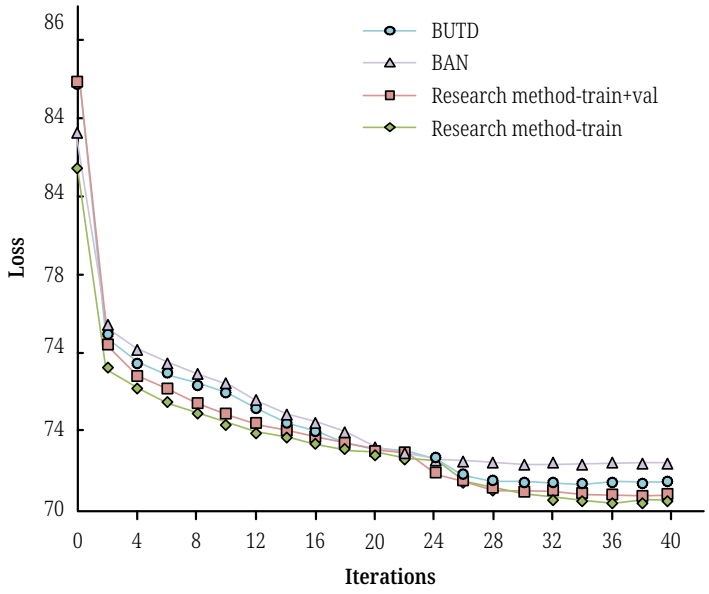


Fig. 6. Loss function

As can be seen from Figure 6, all methods go through 40 iterations. At the end of the model, train means that the model is only trained on the training set; train+val means that the model is trained on both the test set and the training set; val means the accuracy of the verification set. The loss values of all methods show a clear downward trend as the iterations progress. When the iteration reaches the 2nd and 24th times, the loss curve fluctuates due to the change in the learning rate of the research object; when the iteration reaches about the 28th time, the loss values of all methods begin to stabilize. The convergence speed of the loss function of the research method during the training process is significantly better than other functions. When the iteration reaches about the 28th time, it starts to converge. The corresponding loss function converges to 2.38, which is nearly 0.5% lower than the loss value of the baseline model. The accuracy rate of the data set test was compared between the research method and the classic Mutan model, as shown in Figure 7.

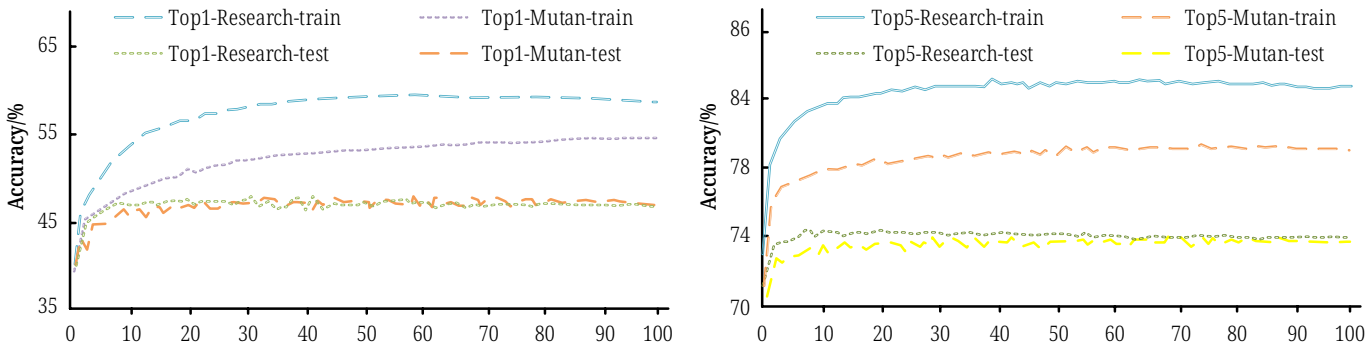


Fig. 7. Top1 and Top5 data set accuracy

It can be seen from Figure 7 that the research method and the Mutan model method are applied to the evaluation of the accuracy of the Top1 and Top5 test sets, respectively. In Figure 7a, as the number of iterations increases, the accuracy of the datasets under both methods begins to increase. Among them, the accuracy rate of the research method model on the training set is about 30 iterations, and the highest accuracy rate is 60%. The highest accuracy rate of the training set under the Mutan model is 55%, and the corresponding iteration number is about 95 times. And under the generalization processing of the test set, there is not much difference in the accuracy results of the two methods, both close to 48%. Figure 7b shows that under the Top5 evaluation method, the accuracy rate and the number of iterations under the two methods show a positive correlation. The highest accuracy rate of the training set results of the research method is 83%, and the accuracy rate of the Mutan model is 79%, and the gap between the two is close to 4%. And the test result accuracy rate of the research method on the test set is 74%, which is nearly 0.5% higher than the result of the Mutan model. During the whole process of the Top5 data set test, the research method starts to converge at the 10th iteration, while the Mutan model starts to converge at the 60th iteration. The convergence speed of the research method is significantly better than that of the classic Mutan model. This shows that the research method has high convergence speed and high feasibility.

4.2 Analysis of the application effect of the research method

To verify the effect of the built model on student education and teaching, the data set parameters were input into the online server of the VQA challenge for evaluation, as shown in Figure 8.

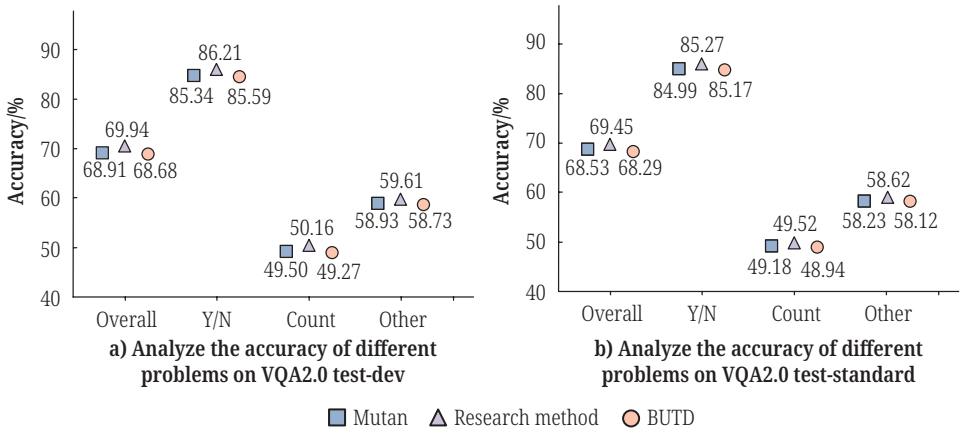


Fig. 8. Accuracy of the method on different questions on different VQA datasets

In Figure 8, on the data set test-dev, the accuracy of the research model on the four types of questions of "overall, whether, count, and others" is 69.94%, 86.21%, 50.16%, and 59.61%, respectively. On the data set test standard, the accuracy of the research method model on the four types of questions of "overall, whether, count, and others" is 69.45%, 85.27%, 49.52%, and 58.62%, respectively. The accuracy of the Mutan and BUTD methods is significantly lower than the research method. The above results show that the mixed attention mechanism and multi-scale VQA research method can significantly improve the accuracy of answering questions and are highly compatible with educational and teaching problems. The research has widely applied the model to question-answering tasks in various fields to verify its applicability, as shown in Figure 9.

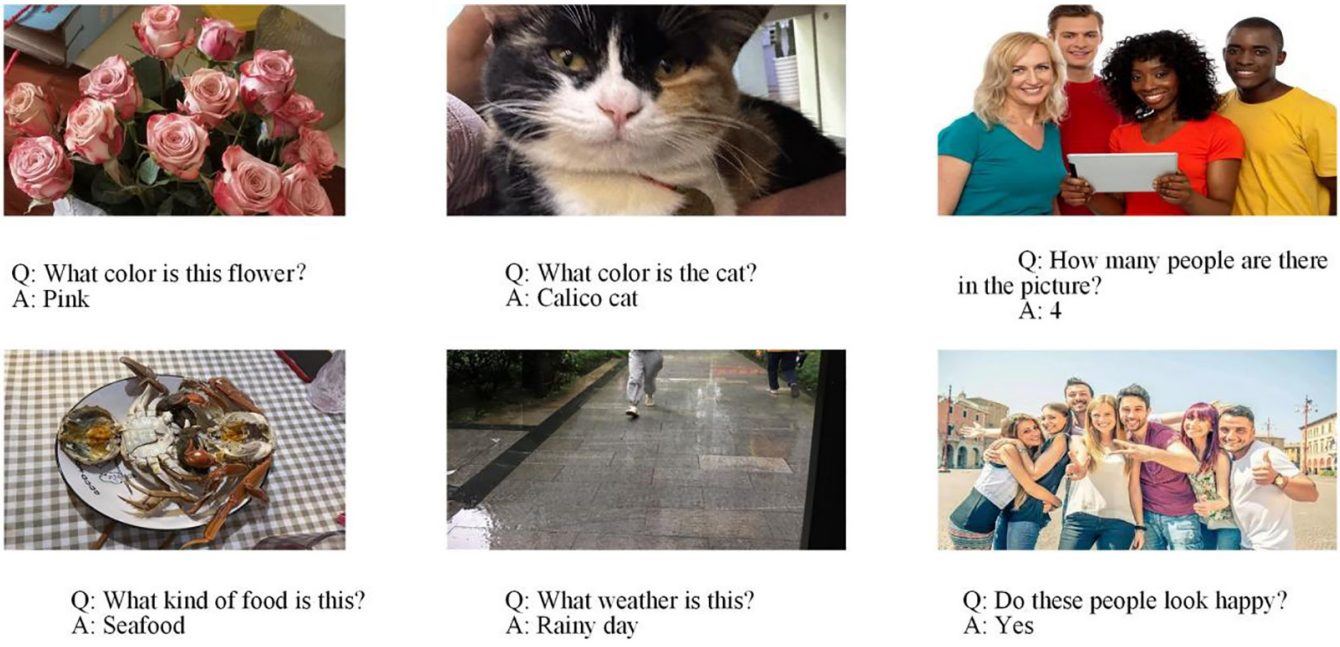


Fig. 9. *(Continued)*

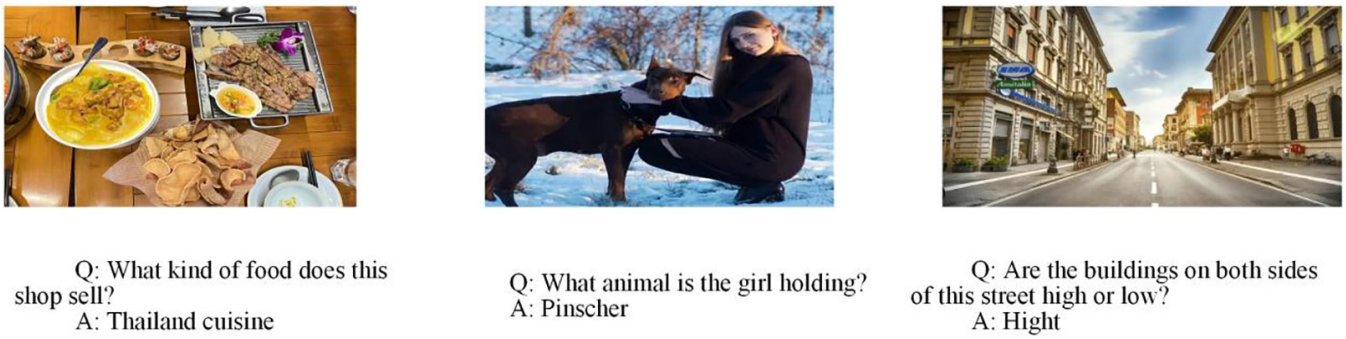


Fig. 9. Example output of the research model on the VQA dataset

In Figure 9, nine pictures represent the output cases of the research model in the VQA task. For the pictures in the first row, the questions mainly include color, animal species, and object counting. The research model can accurately detect the category of each target object and discriminate the number of objects. In the second line, in addition to distinguishing simple objects, the model also needs to infer that the weather is rainy and people are happy based on ground conditions and changes in human expressions. In the third row, the model mainly answers questions about the type of food, the breed of dog, and the height of street buildings. It can be seen that the model accurately identifies that the food is Thai food, the dog is a Doberman, and the street buildings are tall. The above results show that the research model has high applicability and can be applied to the education and teaching of college students. Then use the research method to carry out multiple teachings, and use the comprehensive literacy of the students as the change index to observe the application effect, as shown in Figure 10.

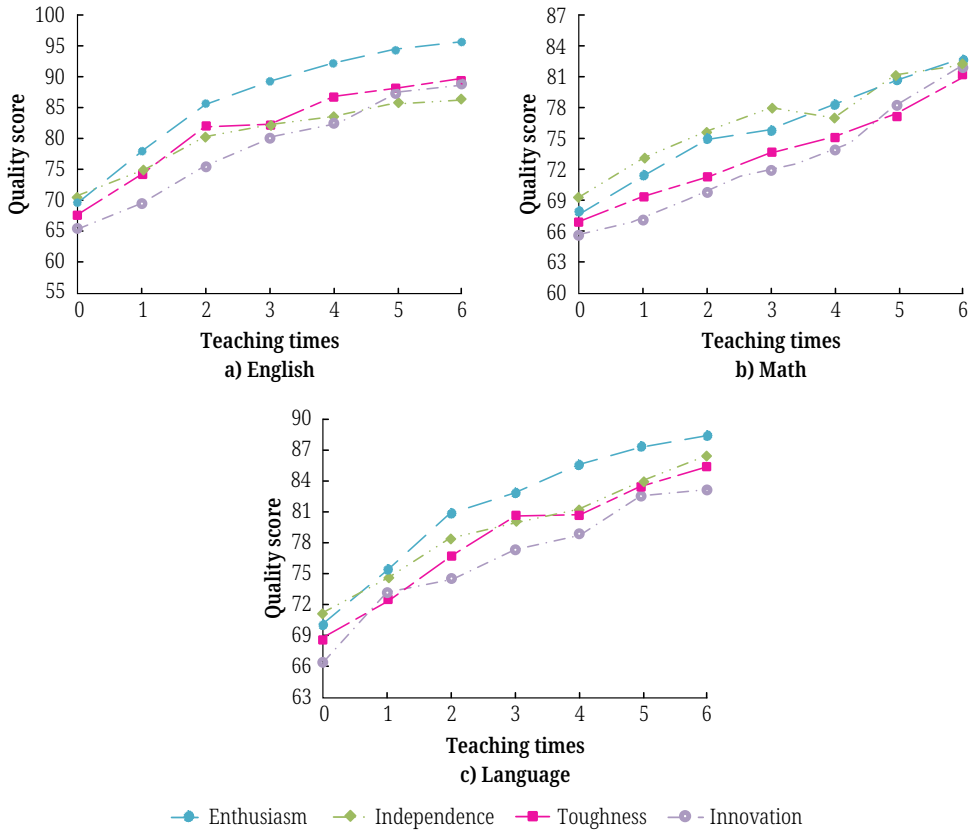


Fig. 10. Changes in students' comprehensive quality of subjects

In Figure 10, the change in students' comprehensive literacy increases with the number of models taught. When the number of models taught reaches six times, each subject has the maximum enthusiasm score, and the score ratios of innovation literacy in subjects such as English, mathematics, and Chinese are 96.32, 93.14, and 88.56; the persistence scores for learning were 88.28, 83.31, and 85.69; and the scores for students' innovative literacy for each subject were 89.91, 82.11, and 83.27. The literacy scores of the above subjects are all higher than 80 points, which shows that the research model can effectively improve students' independent learning ability, improve students' learning methods, and promote the healthy development of students' physical and mental health.

5 CONCLUSION

To better improve the teaching quality of college teachers and use teaching to enhance students' positive emotions, such as learning enthusiasm, this study proposes a teaching quality improvement method based on VQA technology. In the process, the method of multi-scale fusion features is used to improve the large-scale data information features; the mixed attention mechanism method is improved to solve the problem of information redundancy caused by the noise brought by the network; and finally, the VQA technology is introduced to build an educational model. The research results show that when the iteration reaches the 28th or so, the research method begins to converge, and the loss value is 2.38, which is nearly 0.5% lower than the traditional baseline model. On the TOP1 training set, the research method has the highest accuracy rate of 60% when the model is iterated about 30 times, while Mutan has the highest accuracy rate of 55% when the iteration is about 95 times. Under the TOP5 evaluation method, the research method starts to converge at the 10th iteration, and the result accuracy rate is 83%; while the Mutan model starts to converge at the 60th iteration, and the result accuracy rate is 79%. In the application effect analysis on the data sets test-dev and test-standard, the accuracy of the research method on the four types of questions of "overall, whether, counting and others" is 69.94%, 86.21%, 50.16%, 59.61%, and 69.45%, 85.27%, 49.52%, and 58.62%. The research model can accurately discriminate between the information in the VQA dataset and the category of the target object. And the experiment using the model to teach shows that the number of teaching times is positively correlated with the comprehensive quality of students, and the quality scores of the main subjects are all higher than 80 points. The above results show that the research method can positively promote the effect of students' education and teaching and, at the same time, improve the comprehensive quality of students. However, the current research involves relatively single textual data and the content of applied disciplines. In the future, the disciplines can be appropriately expanded to make the research results more scientific and reasonable.

6 REFERENCES

[1] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, and M. Zheng, "Transformer CPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, 2020. https://doi.org/10.1093/bioinformatics/btaa524

[2] T. Saito and Y. Watanobe, "Learning path recommendation system for programming education based on neural networks," *International Journal of Distance Education Technologies (IJDET)*, vol. 18, no. 1, pp. 36–64, 2020. https://doi.org/10.4018/IJDET.2020010103

[3] S. Bagila, A. Kok, A. Zhumabaeva, Z. Suleimenova, A. Riskulbekova, and E. Uaidullakyzy, "Teaching primary school pupils through audio-visual means," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 22, pp. 122–140, 2019. https://doi.org/10.3991/ijet.v14i22.11760

[4] Q. Qiu, "The application of neural network algorithm and embedded system in computer distance teach system," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 148–158, 2022. https://doi.org/10.1515/jisys-2022-0004

[5] E. Bahadir, "Prediction of prospective mathematics teachers' academic success in entering graduate education by using back-propagation neural network," *Journal of Education and Training Studies*, vol. 4, no. 5, pp. 113–122, 2016. https://doi.org/10.11114/jets.v4i5.1321

[6] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1971–1981, 2019. https://doi.org/10.1109/TMM.2019.2894964

[7] C. Liu, Y. Feng, and Y. Wang, "An innovative evaluation method for undergraduate education: An approach based on BP neural network and stress testing," *Studies in Higher Education*, vol. 47, no. 1, pp. 212–228, 2022. https://doi.org/10.1080/03075079.2020.1739013

[8] Q. Han, "Using neural network for the evaluation of physical education teaching in colleges and universities," *Soft Computing*, vol. 26, no. 20, pp. 10699–10705, 2022. https://doi.org/10.1007/s00500-022-06848-9

[9] A. Naim, "E-learning engagement through convolution neural networks in business education," *European Journal of Innovation in Nonformal Education*, vol. 2, no. 2, pp. 497–501, 2022.

[10] H. He, H. Yan, and W. Liu, "Intelligent teaching ability of contemporary college talents based on BP neural network and fuzzy mathematical model," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 4, pp. 4913–4923, 2020. https://doi.org/10.3233/JIFS-179977

[11] H. Li, "Improved fuzzy-assisted hierarchical neural network system for design of computer-aided English teaching system," *Computational Intelligence*, vol. 37, no. 3, pp. 1199–1216, 2021. https://doi.org/10.1111/coin.12362

[12] S. Zheng, Y. Li, S. Chen, J. Xu, and Y. Yang, "Predicting drug–protein interaction using quasi-visual question answering system," *Nature Machine Intelligence*, vol. 2, no. 2, pp. 134–140, 2020. https://doi.org/10.1038/s42256-020-0152-y

[13] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Transactions on Cybernetics*, vol. 51, no. 2, pp. 913–926, 2019. https://doi.org/10.1109/TCYB.2019.2914351

[14] W. Zhang, S. Tang, Y. Cao, S. Pu, F. Wu, and Y. Zhuang, "Frame augmented alternating attention network for video question answering," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1032–1041, 2019. https://doi.org/10.1109/TMM.2019.2935678

[15] Q. Cao, X. Liang, B. Li, *et al.*, "Interpretable visual question answering by reasoning on dependency trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 887–901, 2019. https://doi.org/10.1109/TPAMI.2019.2943456

[16] C. Gao, Q. Zhu, P. Wang, H. Li, Y. Liu, A. Van den Hengel, and Q. Wu, "Structured multimodal attentions for TextVQA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, p. 9603–9614, 2021. https://doi.org/10.1109/TPAMI.2021.3132034

[17] X. S. Yang, J. J. Zhou, and D. Q. Wen, "An optimized BP neural network model for teaching management evaluation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3215–3221, 2021. https://doi.org/10.3233/JIFS-189361

[18] E. Okewu, P. Adewole, S. Misra, *et al.*, "Artificial neural networks for educational data mining in higher education: A systematic literature review," *Applied Artificial Intelligence*, vol. 35, no. 13, pp. 983–1021, 2021. https://doi.org/10.1080/08839514.2021.1922847

[19] X. S. Yang, J. J. Zhou, and D. Q. Wen, "An optimized BP neural network model for teaching management evaluation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3215–3221, 2021. https://doi.org/10.3233/JIFS-189361

[20] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention siamese network," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3068–3080, 2019. https://doi.org/10.1109/TCYB.2019.2936503

[21] X. Xu, T. Wang, Y. Yang, A. Hanjalic, and H. T. Shen, "Radial graph convolutional network for visual question generation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1654–1667, 2020. https://doi.org/10.1109/TNNLS.2020.2986029

[22] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L. J. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1893–1908, 2019. https://doi.org/10.1109/TPAMI.2018.2890628

[23] X. S. Yang, J. J. Zhou, and D. Q. Wen, "An optimized BP neural network model for teaching management evaluation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3215–3221, 2021. https://doi.org/10.3233/JIFS-189361

7 AUTHOR

Fang Lin, Training Center, Zhengzhou Shengda University, Zhengzhou, 450000, China (E-mail: Fang_Lin2023@outlook.com).