A Novel Online Teaching Effect Evaluation Model Based on Visual Question Answering

Yanqing Cui[1], Guangjie Han[2*], Hongbo Zhu[3]
[1]School of Information and Business Management, Dalian Neusoft University of Information, China
[2]College of Internet of Things Engineering, Hohai University, China
[3]School of Information Science and Engineering, Shenyang Ligong University, China
cuiyanqing@neusoft.edu.cn, hanguangjie@gmail.com, hombochu@sina.com

Abstract

The paper proposes a novel visual question answering (VQA)-based online teaching effect evaluation model. Based on the text interaction between teacher and students, we give a guide-attention (GA) model to discover the directive clues. Combining the self-attention (SA) models, we reweight the vital feature to locate the critical information on the whiteboard and students' faces and further recognize their content and facial expressions. Three branches of information are encoded into the feature vectors to be fed into a bidirectional GRU network. With the real labels of the students' answers annotated by two teachers and the predicted labels from the text and facial expression feedback, we train the chained network. Experiment reports a couple of competitive performance in the 2-class and 5-class tasks on the self-collected dataset, respectively.

Keywords: VQA, Facial expression recognition, Online teaching effect evaluation, GRU, Guide-attention

1 Introduction

With the rapid development of networking and the popularization of online education, many businesses of the traditional education have been informationized, such as course choosing, performance evaluation, etc [1]. From teacher's perspective, the AI-empowered technologies are even charging the traditional teaching mode. Especially for one-to-many teaching, the progress cannot be consistent and the effectiveness are always difficult to be investigated timely in the past online education. The performance evaluation are relied on questionnaires. The above problems are the key barrier to hinder the further development. Thanks to deep learning (DL) [2], its outstanding performance in the fields of text analysis and image understanding offers a series of strides for the basic technologies for online education. Based on DL-based schemes, many effective applications are designed and implemented successively, e.g., the computers can review examination papers automatically via few manual intervention [3], and some novel online interactive model take the place of face-to-face brainstorming [4-5]. But unfortunately, the existing modes are almost based on the analysis of the single information source, which are difficult to obtain a comprehensive teaching effect feedback. Thus, how to fuse multi-source information is crucial for the next stage of the AI-empowered educational technology.

Visual question answering (VQA) is a solution on solving an image understanding problem with the text questions. Since its emergence in 2015, a large number of models are based on the joint embedded (JE) model, which dominates the development trend of the VQA tasks. The model achieves a three-stage preprocessing for raw inputs, specifically denoting a multi-source information fusion from the images and texts to a learnable feature vector. The vectors are gathered and fed into at least one classifier. Due to the convenience of the model training, more researchers have improved more feature extraction methods to diversify image and text characterizations, or even tried some novel fusion of the feature modes [6-11]. It is a more perfect way for online education. As a feedback of the teacher's questions, the facial expressions of the students can be recognized with a VQA-based classification framework. More abundant and subtle expression changes are captured and classified into the labels of comprehension. With more detailed interpretation of teachers, the series of changes can offer the traceability for thinking logic. The feedback determine whether the current knowledge point need be rehashed.

Intuitively, the above task seems to be decomposed into three sub-tasks, sequentially including image captioning (IC) [12] machine reading comprehension (MRC) [13] and facial expression recognition (FER) [14]. They are used to translate images into the readable questions, find the best answers, and monitor the student feedback, respectively. However, different students have the different durations to understand the presentations in a common teaching procedure. Thus, the frameworks have to own a better comprehensive ability for the content of the images and give the further inferring. Besides, the outputs must be short and explicit, which are easier to be compared with the ground truths.

Refer to a classic survey of VQA [15], we can divide the existing methods into four groups. Except the JE-based schemes, the other mainstream schemes contains attentive model, compositional model and the models with some external knowledge databases. The former two model are used to simple the image and text tasks based on attentive module and grammar parser, which are working as a guide to point the remarkable information for the machines, respectively. Moreover, external databases give the other knowledge for the detected targets, and take the machines becoming smarter.

In this paper, we design a special VQA scheme, and embed it into an online teaching effect evaluation model. The contributions of our works contains two aspects:

1. We design a novel attentive model based on a guide by the textual QA interaction among the system users.

*Corresponding Author: Guangjie Han; E-mail: hanguangjie@gmail.com
DOI: 10.53106/160792642022012301009

Combining with the self-attention models in the sub-tasks of facial expression recognition and teaching content analysis, the model can improve the performance of the teaching effect estimation.

2. We build a mapping from the facial expressions of the students and the teaching content to the comprehensive level for the question interpretation by a VQA-based framework.

The rest of the paper can be organized as follows: Sec. 2 formulates the main problem of our proposal. Sec. 3 introduces the modules of the solution from top to bottom. Sec. 4 reports the experiment data and results and the performance comparison with the schemes based on the state-of-the-art methods. We conclude our work and share some future ideas in Sec. 5.

2 Problem Description

In our proposal, the problem covers four participants: real-time videos from white boards/monitors of teachers, audios of teachers, monitors of students and texts of chat boxes. The participants can be arranged to describe the whole procedure into a timing diagram (shown in Figure 1). Note that there three interactions in Figure 1, containing white board (WB)-to-student, teacher-to-student and chat box (CB)-to-student, which can affect the changes of students' facial expressions. During the steps 6 and 13, we can integrate text, audio and video information into a learnable vector in this paper. For a better computational efficiency, a column of parallel encoders are used to extract three types of feature representations. The entire system have only one input for the students to discuss the questions with the teacher and the students. All the questions and answers are shared on the chat box. Thus, three are also three QA (Questioning-Answering) modes, which are triggered with video/audio explication, chat conversation and audit, respectively. The aims of the proposed system can be summarized as follows:

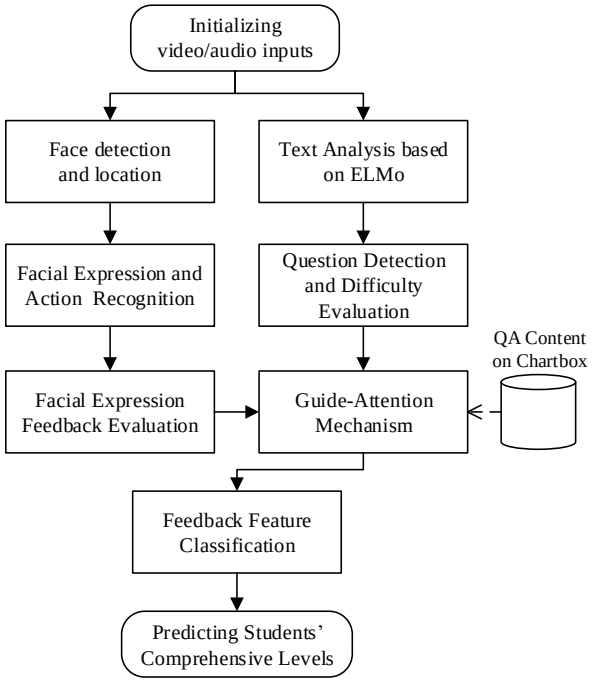

Figure 1. The flowchart of the proposed system for teaching effect evaluation

1. The interpretation process of the teachers in the online video is the classic self-QA processes, which can be used to extract the questions and the solving processes.

2. The chat conversation between the teacher and the students on the chat box is the reviews for the pervious interpretation, which can bring a focus on the intricate questions.

3. For the other students looking on the chat box, the QA interaction can also drive their self-examination.

The relations between the inputs and outputs of the above three encoding modes can be illustrated in Fig. 2. From the left to the right feature groups, three branches filter the extracted features via the attentive models. Self-attention (SA) [16] model is used to reweight the original features to highlight their important part. For the indigestible content on the white board, the feedback of chat box give a guideline for another attention mechanism. Guide- attention (GA) models point out the questions which needs to be retold and detailed. Meanwhile, with detailing the knowledge points, they can locate the questions more precisely.

3 Framework

In Figure 2, the branches consist of two video encoders and one text encoder. For the video of the white board, the main tasks is finding the being explained region, which can be formulated as a target detection task. For the videos of cameras, SA models are used to discover the subtle features to recognize the facial expressions. With the text questions on the chat box and external knowledge databases, the related texts and images of the questions are refined better.

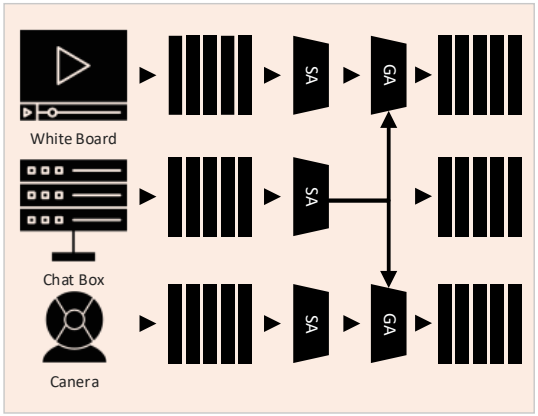

Figure 2. Three encoding branches with SA&GA

3.1 Encoding the content on the white board

Target detection is one of the important applications in the field of machine vision. Its key tasks is to locate the targets from the images, and identify the categories and attributes. Traditional detection algorithms, such as deformable component model method, select regions, extract features and classify them by sliding a sampling window. Most of them have poor region selection effect and time-costing, due to the nature of their hand-crafted features. Besides, the extracted low-level features leads poor generalization of the model. Benefited from the better performance of the DL-based schemes in the visual tasks, the schemes have become the dominant methods for target detection. The existing models can be roughly divided into two main categories. The first

category is two-stage detection model, which consists of candidate region proposal network (RPN) [17] and region feature extraction, such as R-CNN series models [17-19]. The second type is single-stage detection model, which uses end-to-end training without RPN, such as YOLO [20-21] series models. Considering the demand of our proposal, we use Faster R-CNN for a target detector for the content on white board. We retrain a ResNet-101 on a self-collected dataset of primary mathematics. For a given frame I of the video, a sampling window of RPN is sliding on it for screening the M candidate targets $X_I = \{x_1, x_2, ..., x_m\}, x_i \in \mathbb{R}^D$ with a non-maximum suppression and IoU (Intersection over Union) measurement. The screened feature are jointed into a set of final feature for the frame I.

3.2 Recognizing the text on the chat box

The existing VQA-based systems analyze texts by a trained network to translate the words into the vectors sequentially, i.e. each word is assigned by an element of a vector. These systems perform worse on the dataset of the polysemy texts. As solutions, the bidirectional recurrent networks are proposed recently, in which Bidirectional Encoder Representation from Transformers (BERT) [22] and Embedding from Language Models (ELMo) [23] are more effective and remarkable. Due to the higher computation complexity of BERT, we select ELMo as the inferring framework for our proposal. It is an encoder-decoder network. The encoder is a bidirectional deep model to obtain the contextualized representations. The decoder can be running at each recurrent unit to translate its hidden state into the semantic information. The objective function is jointly maximizing the log likelihood of the forward and backward branches. The network architecture is shown in Figure 3.

For the forward branch, it is a chain of RNN units. To further reduce the computational cost, we replace the LSTM (Long Short Term Memory) units [24] of the original ELMo to GRUs [25], which has two gates (update gate z_t and reset gate r_t). Their update can be computed as the following equations:

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \quad \textbf{(1)}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \quad \textbf{(2)}$$

$$h_t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t]) \quad \textbf{(3)}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \quad \textbf{(4)}$$

$$y_t = \sigma(W_o \cdot h_t) \quad \textbf{(5)}$$

For a N length of sentence $\{t_1, t_2, .., t_N\}$, the models of the forward and backward beranches can be formulated as Eq.6 and Eq. 7, respectively.

$$p(t_1, t_2, .., t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, .., t_{k-1}) \quad \textbf{(6)}$$

$$p(t_1, t_2, .., t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, .., t_N) \quad \textbf{(7)}$$

A couple of their logarithmic computation results can be summed to optimize the parameters of the GRU units by maximizing

$$\sum_{k=1}^{N} (\log p(t_k \mid t_1, t_2, .., t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{GRU,}, \Theta_s) + \log p(t_k \mid t_{k+1}, t_{k+2}, .., t_N; \Theta_x, \overleftarrow{\Theta}_{GRU,}, \Theta_s)) \quad \textbf{(8)}$$

where $\overrightarrow{\Theta}_{GRU,}$ and $\overleftarrow{\Theta}_{GRU,}$ are the parameters of the sequential GRUs in the forward and backward branches, respectively. Θ_x is the shared parameters of two branches, and Θ_s is the weights of the units.

3.3 Facial expression recognition

In the section, we also give a STGCN-based inferring network for facial expression recognition (FER) [26]. Strictly speaking, we build an extent network to monitor the facial expression with the emotions and the actions of upper body for the students. Here, we first extract the skeleton of the upper body which include face and upper torso with the Openpose toolbox [27]. To boost the performance of FER, we improve the 2D template of Openpose, and the updated template contains 15 joints, in which the additional joints are located at canthus, nose, and corner of her mouth.

We used a GCN-based encoder to generate multi-scale graph feature from a spatio-temporal view. A classifier are working to point out the current learning states of the students. Considering their different reaction times, the durations of the time sampling windows are also set as 3, 5, 7, 9, and 11, respectively.

3.4 Attention mechanisms

For the above multi-source features, we follow the combination of attention mechanisms in MCAN [28] to produce higher-level feature representations. The feature of the last layer of self-attention model are used as the query matrix of the other layer. Concretely, for a M length of the reweighted image features $X^L = \{x_1^L, x_2^L, ..., x_m^L\}$ and a N length of text characterizations $Y^L = \{y_1^L, y_2^L, ..., y_m^L\}$, the model are normalizing two weight sets, which can be computed by

$$\alpha = \text{soft}\max(\text{MLP}(X^L)) \quad \textbf{(9)}$$

$$\beta = \text{soft}\max(\text{MLP}(Y^L)) \quad \textbf{(10)}$$

where α and β are the weight vectors of the image features and text characterizations. Their weighted fusion computation can be denoted as

$$\chi = \text{Norm}(W_x^T x + W_y^T y) \quad \textbf{(11)}$$

where W_x^T and W_y^T are the linear mapping matrixs. After a softmax activation, the final fusion are assign the scores for the answer set. Compared with the refered answers, a cross entropy loss function is used to update network parameters.

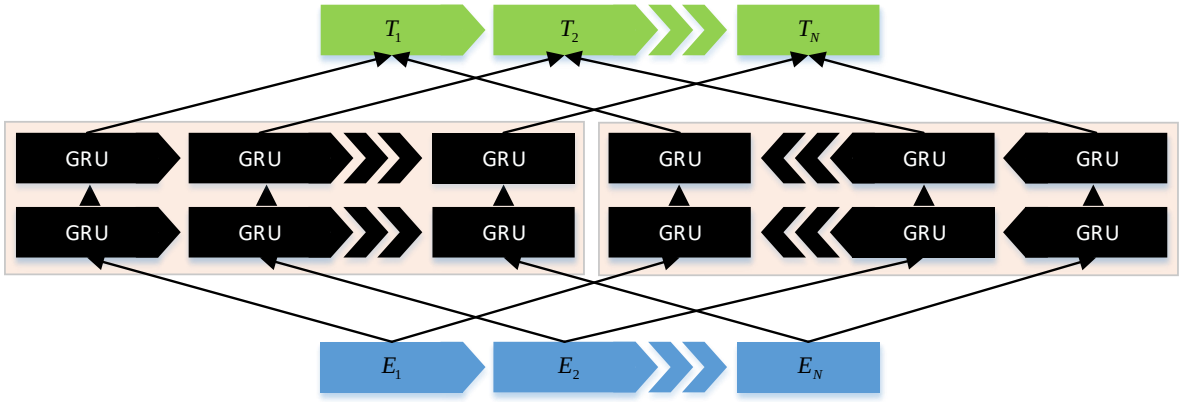

Figure 3. The network structure of ELMo

4 Experiments

In this section, we first introduce the self-collected dataset, following by experimental setup. The parameter and implementation details are listed. Evaluation criteria and results are reported in the last two subsections.

4.1 Dataset

We selected the 41 teaching videos on the Bilibili website, and the ages of audiences are limited from 7-8 years old. Meanwhile, we called 19 age-appropriate students to watch the videos after informing consent form from the students and their guardians. 13 students finished watching all the videos. The cameras recorded the actions and facial expressions. The total length is 39 hours. To simulate an interaction on chat box, we employ two primary school teacher to answer the questions from students via WeChat.

4.2 Experiment Setup

We set the number of the candidate regions on white board, and each region has the 1024 dimension of features. The atentive modules with the hidden layers of 256 neurons are divided into 8 heads. We set 5-layer MCAN to optimize the output answers. The answers are packaged into a dictionary after screening out the answers whose occurrence amounts are more than 8. The activation function is ReLU (Rectified Linear Unit) [29], and the parameters of the Adam optimizer are set as 0.9 and 0.95.

The aims of the experiment contain two aspects: estimating the understanding level of the students for the teaching knowledge points, and estimating the teaching effect. For the first task, we need classify the diffculties of the questions. To obtain the refere labels of the diffculties, two teacher give a 5-degree annotation. According to the change of the subtle facial expressions and body language from the cameras of the students, we need to infer their understanding levels with a guide of the real QA procedures on the chat box, and the refered labels of the understanding levels are surveyed by a exam. Meanwhile, the average of exam results are used to estimate the teaching effect as the refered labels to compare with the outputs of the VQA-based system.

4.3 Evaluation Criteria

For the above two tasks of our experiment, we compute the correct rates, respectively. For understanding level estimation, the predicted difficulty of the question are compared with the score (5-degree) annotated by two teachers. If the difference value of the score pair is less than 1, we consider the infered result is correct, otherwise is incorrect. Also, the average of the inferred results are compared with those of the exam.

4.4 Results

We report an overall teaching estimation result in Table 1. We use glove and word2vec as a baseline for text characteristic extractor, combining with a template parser for facial expression recognition. In the baseline schemes, the content on the white board are detected and tracking by a backbone of Faster R-CNN [18]. Also, a scheme of the winner in the 2019 VQA challenge is listed in the third row of the table. The last two rows show the results of the two counterparts of the proposed framework.

Table 1. Quantitative comparison between the global accuracy rates (GARs) of the state-of-the-art schemes and the proposed schemes in the 5-class (Comprehensive Levels) and 2-class (Comprehensible/Incomprehensible) tasks.

Schemes	GAR-5 (%)	GAR-2 (%)
glove+Faster R-CNN	69.34	79.25
word2vec+Faster R-CNN	69.87	80.31
2019 winner (glove)	72.41	88.27
Proposed (Small-scale)	74.11	90.23
Proposed (Large-scale)	72.82	89.15

We note that the top two rows give a couple of worse results in the comparison, about 3-5% and 9-10% of falling behind the other rows in the columns of GAR-5 and GAR-2, respectively. The winner of the VQA challenge 2019 is the watershed in the results of the comparison. However, the scheme also used a static word vector generator, which causes a gap with the proposed schemes. We note that the comparison contains two counterparts of the proposed scheme. The main differences between the two counterparts is

sizes of the GRUs and ELMo, and the larger sizes are about 4 times of the small ones.

Locally, for a rigorous analysis for all the modules of the proposed framework, we give three groups of performance comparisons.

4.1.1 Effectiveness of word embedding

First, we focus on the effectiveness of the dynamic feature vectors. The experiment will build models using different text characterization methods to evaluate the impact of dynamic and static word vectors on the accuracy of the results. The experiment used pre-trained word2vec and glove word vectors in static word vectors were selected, and a single layer GRU network was casculated to transform its feature dimension into 256 dimensions, which was convenient for subsequent fusion with image features. word2vec word vector is the word vector obtained by word2vec model on Sougou lab. The word2vec word vector is composed of 300 dimensional vectors representing 2.4 million words and phrases. Glove word vectors are trained from the corpus, and contain 2.1 million 300 dimension vectors. To use static word vectors, the input sample text is first cropped to a sequence of words of length 10, and then a query table is used to get a static vector for each word. For the words without pre-trained word vectors in the dataset, their word vectors are initialized as zero vectors. Table. 2 shows the amounts and dimensions of the word vectors with the different schemes of Table 1.

The ELMo model uses character-level encoding, so even for words that do not exist in the corpus, the initial word vector can still be obtained, and then the ELMo-based word vector can be obtained. Therefore, the number of ELMo-based word vectors is infinite theoretically.

Table 2. Amounts and dimensions of three models for generating word vector

Schemes	Amounts	Dimensions
Glove	2.1 M	300
Word2vec	2.4 M	300
ELMo (Small-scale)	/	256
ELMo (Large-scale)	/	1024

4.1.2 Effectiveness of facial expression recognition

For facial expression and action recognition, we simple the recognition problems as two binary classification tasks to determine whether the students are listening to the class and whether the students are understanding to the current interpretation, respectively. Thus, we give two comparisons of the template-based recognition. The first comparison is based on the 8-joint and 15-joint templates to recognize the actions with the students' facial expressions, which is shown in Table 3. Also, the second group is used to compare the FER results between the 7-joint and the 15-joint schemes.

Table 3. Peformance comparison of binary classification for facial expression (FER) and action recognition (AR) with the different template-based schemes

Schemes	mAPs
7-joint template for FER	0.792
8-joint for template AR	0.854
15-joint template for FER	0.863
15-joint template for AR	0.856

In Table 3, due to simplifying the classification tasks, all the results of the template-based schemes perform the higher mean Average Precisions (mAPs) about 79.2-85.6% than those of performing on the other FER datasets. Besides, we note the results of the 7-joint and the 15-joint schemes have a gap of 7.1% in the FER task, as well as a gap of 1.2% between the 8-joint and the 15-joint schemes in the AR task. The difference between two gaps reveals that the template with more joints give the better performance than those of less joints. However, there is only a small improvement in the AR task, because the actions are weakly relied on the changes of the facial expression. But in another task, the facial expressions of the students are changing, when the students feel the difficulty of comprehending the teacher's interpretation. Some body pose adjustment are always occurred with the time-varying facial expressions. Naturally, the boosting mAP demonstrates the 15-joint scheme can further help classifying the facial expression in the binary classification task. To protect the privacy of the students, we follow the parental expectations, and cannot show the detailed cases of the facial expression and action recognition.

To capture the facial expressions and actions, each student is asked to use a web camera. The sampling frequency of the cameras is between 15Hz and 24Hz. Thus, temporal sampling window sizes have to be considered. Thus, we also give a changing of mAP with the growing window sizes in Figure 4.

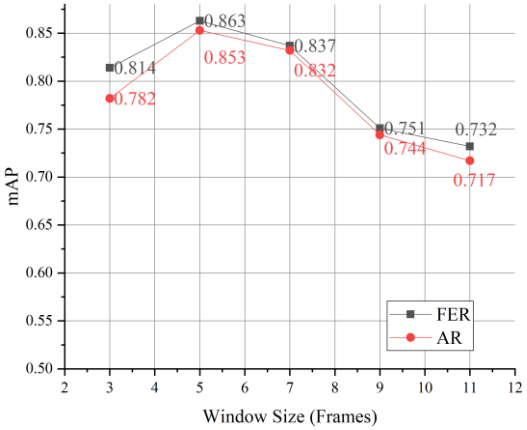

Figure 4. Changing curves in the FER and AR traces with the growing sizes of the temporal sampling windows

In Figure 4, we can see that the best mAP is shown at the window size of 5 frames both in the FER and AR traces. The main reason is the reasonable size of context encoding. The other sizes of windows may cover some confusable facial expressions and upper body actions, which can cause the failure of FER and AR. So we define the size of the sampling window as 5. To demonstrate the enough competitiveness of the proposed FER scheme, we also give a comparison among the state-of-the-art (SOTA) methods and ours in Table 4.

In Table 4, we list four groups of mAPs for the compared methods on the self-collected dataset. The third scheme performs best in this comparison, which benefits from their consideration on the geometrical and textual contents of the joints on the face templates [31]. However, more sub-tasks can bring more computational complexity and less applicability for some clients with lower computing power.

The second scheme is based on the original pose template. Compared with STSGN [31], it has a simple facial graph as well as the lower mAP but faster running speed. The proposed scheme give an appropriate joint amount to balance speed and precision. We note that the additional joints can effectively boost the recognition performance than the original STGAN [26]. Also, the guide-attention (GA) module can give a help for performance impartment.

Table 4. Performance comparison on the self-connected dataset by using the SOTA methods and the proposed schemes

Schemes	mAPs
STM-ExpLet [30]	0.715
STGCN [26]	0.811
STSGN [31]	0.882
STGCN-GA (Ours)	0.863

4.1.3 Effectiveness of attention mechanism

We give two types of attention mechanism in the proposed framework. Self-attention and Guide-attention models are fused to reweight the important information. The former is running on the feature encoders to point out the key content on the white board and the chat box which cause a series of subtle expressions. Meanwhile, part of the subtle expressions are screened to encode into a feature set. Integrating a temporal factor, the facial expressions are decoded into the labels of comprehensive levels. We list the recognition results with/without self-attention model in Table 5.

Table 5. Performance comparison of four modules with (w)/without (o) self-attention (SA) models

Schemes	mAP o/SA	mAP w/SA
ELMo for WB	0.691	0.812
ELMo for CB	0.646	0.785
FER	0.693	0.801
AR	0.637	0.814

In Table 5, all the modules perform better than their originals after appending self-attention models. It is worthy to note that the greatest heighten is in the AR module. Because that the self-attention model can find the remarkable local actions and avoid the interference of unconscious movement. In contrast, for the template with few joints, the performance improvement is minor. Also, the recognition results for the content of white board (WB) and chat box (CB) are performing the similar state, due to the difference between the semantic integrity of prepare lesson and instant reaction.

Guide-attention model is based on the clue from the text information on the chat box to boost the attentions for the interactive part on the white board and the change of the subtle expression. We give a performance comparison between the schemes with/without guide-attention models in Table 6.

Table 6. Performance comparison of three modules with (w)/without (o) guide-attention (GA) models

Schemes	mAP o/GA	mAP w/GA
ELMo for WB	0.812	0.859
FER	0.801	0.863
AR	0.814	0.856

In Table 6, the three modules benefit from the guide-attention models, and perform better than the schemes with self-attention model only. As we expected, there is only a minor increasing performance of the AR trace compared to the other two traces because the text of the chatbox can offer the clues, which lead more attention to local expressions, emotion, and location of the whiteboard but actions.

5 Limitation

There are some limitations for the proposed framework in its further applications, because it is a VQA-based exploratory work. The detailed limitations contain the following aspects:

(1) Currently, the learning-based system can only analyze the elementary course with mass visual interpretation. Because the VQA model offer the answers, including the classes and numbers of the objects.

(2) For the different students, their personalities and habits cause the diversiform feedbacks when they try to understand the logic operations. The mapping between the emotions and facial expressions is not stable. With increasing age, the more pronounced the phenomenon.

6 Conclusion

With the growing trend of online education, the teaching effect is vital to identifying qualified teachers. A practical model for online teaching effect evaluation is indispensable. In the paper, we proposes a novel visual question answering (VQA)-based online teaching effect evaluation model. Based on the text interaction between teacher and students, we give a guide-attention (GA) model to discover the directive clues. Combining the self-attention (SA) models, we reweight the vital feature to locate the critical information on the whiteboard and students' faces and further recognize their content and facial expressions. Three branches of information are encoded into the feature vectors to be fed into a bidirectional GRU-based inferring chain. We train the chained network with the real labels of the online videos annotated by two teachers and the predicted labels from the text and facial expression feedback. It can build a mapping between the facial expression and the VQA-based answering to form the online teaching effect evaluation model. Experiment reports the predicted results of 74.11% and 90.23% in the 5-class and binary classification tasks for the students' comprehensive levels based on their text interaction on the chat box and the facial expressions. The results demonstrate that the proposed VQA-based model can preliminarily satisfy the demand of the online teaching effect evaluation system.

References

[1] T. S. Chen, P. S. Chiu, Y. M. Huang, W. J. Hsieh, The Effect of Context-Aware Mobile Learning on Chinese Rhetoric Ability of Elementary School Students, *Journal of Internet Technology*, Vol. 17, No. 7, pp. 1309-1316. December, 2016

[2] C. C. Wu, C. C. Li, C. F. Tsai. Factors Determining of Effects of Teachers' Web-Based Teaching Platform Usage-Using UTAUT to Explore, *Journal of Internet*

Technology, Vol. 14, No. 6, pp. 919-928, November, 2013.

[3] L. Logeshvar, A. B. Premnath, R. Geethan and R. Suganya, AI based Examination Assessment Mark Management System, *International Conference on Secure Cyber Computing and Communications (ICSCCC)*, Jalandhar, India, 2021, pp. 144-149.

[4] C. Ni, The Human-Computer Interaction Online Oral English Teaching Mode Based on Moodle Platform, *IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, 2021, pp. 633-635.

[5] G. Li, T. Lu, X. Ding, N. Gu, Predicting Collaborative Edits of Questions and Answers in Online Q&A Sites, *Journal of Internet Technology*, Vol. 17, No. 6, pp. 1187-1194, November, 2016.

[6] M. Malinowski, M. Rohrbach, M. Fritz, Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images, *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1-9.

[7] Y. Zhu, J. J. Lim, L. Fei-Fei, Knowledge Acquisition for Visual Question Answering via Iterative Querying, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6146-6155.

[8] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu, Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering, *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 2, Cambridge, MA, USA, 2015, pp. 2296-2304.

[9] H. Noh, P. H. Seo, B. Han, Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 30-38.

[10] S. Takada, R. Togo, T. Ogawa, M. Haseyama, Estimation Of Visual Contents Based On Question Answering From Human Brain Activity, *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020, pp. 61-65.

[11] M. Gu, Z. Zhao, W. Jin, D. Cai, F. Wu, Video Dialog via Multi-Grained Convolutional Self-Attention Context Multi-Modal Networks, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 12, pp. 4453-4466, December, 2020.

[12] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, T. Chua, More is Better: Precise and Detailed Image Captioning Using Online Positive Recall and Missing Concepts Mining, *IEEE Transactions on Image Processing*, Vol. 28, No. 1, pp. 32-44, January, 2019.

[13] F. Xu, W. Zhang, H. Du, S. Zhou, Enhancing Machine Comprehension Using Multi-Knowledge Bases and Offline Answer Span Improving System, *Journal of Internet Technology*, Vol. 22, No. 5, pp. 1095-1107, September, 2021.

[14] F. Setiawan, A. G. Prabono, S. A. Khowaja, W. Kim, K. Park, B. N. Yahya, S.-L. Lee, J. P. Hong, Fine-grained emotion recognition: fusion of physiological signals and facial expressions on spontaneous emotion corpus, *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 35, No. 3, pp. 162-178, October, 2020.

[15] K. Khurana, U. Deshpande, Video Question-Answering Techniques, Benchmark Datasets and Evaluation Metrics Leveraging Video Captioning: A Comprehensive Survey, *IEEE Access*, Vol. 9, pp. 43799-43823, February, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, Attention is all you need, *Proceedings of the 31th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 2, Long Beach, CA, USA, 2017, pp. 6000-6010.

[17] R. Girshick, Fast R-CNN, *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448.

[18] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June, 2017.

[19] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 2, pp. 386-397, February, 2020.

[20] J. Redmon, A. Farhadi, YOLO9000: Better, Faster, Stronger, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517-6525.

[21] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox and A. Farhadi, IQA: Visual Question Answering in Interactive Environments, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 4089-4098.

[22] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint*, arXiv:1810.04805v2, May, 2019.

[23] M. E. Peters, M. Neumann, M. Lyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, *arXiv preprint*, arXiv:1802.05365, March, 2018.

[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, December, 1997.

[25] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *arXiv preprint*, arXiv 1412.3555, December, 2014.

[26] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-Based Action Recognition With Shift Graph Convolutional Network, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 180-189.

[27] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172-186, January, 2021.

[28] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077-6086.

[29] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, *Proceedings of the 14th International*

Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Vol. 15, pp. 315-323, January, 2011.

[30] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 10, pp. 1683-1699, October. 2007.

[31] J. Zhou, X. Zhang, Y. Liu, X. Lan, Facial Expression Recognition Using Spatial-Temporal Semantic Graph Network, *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020, pp. 1961-1965.

Biographies

Yanqing Cui is currently an Associate Professor of Dalian Neusoft University of Information, Dalian, China. She received the B.E. and M.S. degrees from Harbin Institute of Technology, Harbin, China in 2003 and 2005, respectively. Her research interests include internet of things, machine learning and cloud computing.

Guangjie Han is currently a Professor with the Department of Internet of Things Engineering, Hohai University, Changzhou, China. He received his Ph.D. degree from Northeastern University, Shenyang, China, in 2004. His current research interests include Internet of Things, Industrial Internet, Machine Learning and Artificial Intelligence, Mobile Computing, Security and Privacy.

Hongbo Zhu is currently an Associate professor with School of Information Science and Engineering, Shenyang Ligong University. He received the B.Sc., M.E., Ph.D. degree from Northeastern University, Shenyang, China, in 2009, 2012, and 2020, respectively. His research interests include medical image computing and deep learning.