Mathematical Biosciences and Engineering

MBE, 20(8): 13581–13601.
DOI: 10.3934/mbe.2023606
Received: 05 April 2023
Revised: 12 May 2023
Accepted: 16 May 2023
Published:14 June 2023

http://www.aimspress.com/journal/MBE

Research article

A textual and visual features-jointly driven hybrid intelligent system for digital physical education teaching quality evaluation

Boyi Zeng[1], Jun Zhao[2,*] and Shantian Wen[2]

[1] Institute of Sport, Xihua University, Chengdu, Sichuan 610039, China
[2] School of Physical Education, Huzhou University, Huzhou, Zhejiang 313000, China

* **Correspondence:** Email: zhaojun@zjhu.edu.cn.

Abstract: The utilization of intelligent computing in digital teaching quality evaluation has been a practical demand in smart cities. Currently, related research works can be categorized into two types: textual data-based approaches and visual data-based approaches. Due to the gap between their different formats and modalities, it remains very challenging to integrate them together when conducting digital teaching quality evaluation. In fact, the two types of information can both reflect distinguished knowledge from their own perspectives. To bridge this gap, this paper proposes a textual and visual features-jointly driven hybrid intelligent system for digital teaching quality evaluation. Visual features are extracted with the use of a multiscale convolution neural network by introducing receptive fields with different sizes. Textual features serve as the auxiliary contents for major visual features, and are extracted using a recurrent neural network. At last, we implement the proposed method through some simulation experiments to evaluate its practical running performance, and a real-world dataset collected from teaching activities is employed for this purpose. We obtain some groups of experimental results, which reveal that the hybrid intelligent system developed by this paper can bring more than 10% improvement of efficiency towards digital teaching quality evaluation.

Keywords: hybrid intelligent system; textual features; visual features; smart cities

1. Introduction

With the development of the times and the progress of science and technology, along with the

rapid development and popularization of computer technology and network technology, modern educational technology has entered all aspects of human life, especially its application in education and teaching, which further reflects the modernization of teaching means. School physical education is an important part of higher education. The substantial expansion of school sports facilities, the continuous improvement and perfection of school physical education management, the continuous improvement of the quantity and quality of teaching staff, the increasing amount of scientific research, etc., all indicate that college physical education in China has achieved a leap forward in this process. On-line PET (physical education teaching) effect evaluation is an integral part of educational science. It systematically collected information using all feasible evaluation techniques. The value evaluation of the teaching effectiveness of physical education classroom teaching was conducted as a countermeasure, so as to provide a basis for making decisions to optimize physical education classroom teaching [1]. Especially with the application of computer-assisted instruction and the appearance of computer information networks, the teaching methods facing individual differences will develop more rapidly. Faced with lifelong sports and national fitness programs, a new sports information model is bound to permeate PET.

The core value of higher education is mainly judged by the quality of education and teaching, and the most common manifestation of the quality of teaching is the learning effect of students in the learning process and results. In this case, universities can no longer turn a deaf ear to the problems existing in sports, and we must take action to meet the challenges [2,3]. Guo et al. concluded that the development and practice of school-based curriculum of physical education can effectively improve students' learning effect in physical education class, positively change students' learning attitude towards physical education class, improve the spirit of cooperation and mutual assistance among students, and play a positive role in enhancing students' health awareness and behavior [4]. De-Kun et al. think that the factors that really play a role in improving the overall benefit of PET are actually the art between teaching content and organizing teaching methods and the art of teachers themselves [5]. Kidokoro et al. mentioned that evaluating students should not only evaluate the degree of knowledge acquisition and mastery, but also evaluate them from the intellectual and non-intellectual levels [6]. Vinnikov et al. evaluated the effect of teaching and examination according to the goals and requirements that the school can achieve in training students. Through evaluation, summary and improvement of teaching methods, college PET evaluation played a promoting role in PET [7]. Establishing a systematic and scientific evaluation system to determine reasonable evaluation indexes and giving corresponding normalized weights is the key to the success or failure of online PET effect evaluation and result value judgment. The original evaluation system cannot fully meet the requirements of the current PET development. It is necessary to establish a new online PET effect evaluation system, highlight the teaching concept of physical education and health courses, reflect the essence of PET as truly as possible, and reflect the evaluation, judgment and guidance function of evaluation on PET. Online PET effect evaluation plays a multi-faceted role in the process of college PET. It restricts the management and decision-making of college PET, regulates and controls the progress of PET activities as a whole, ensures the realization of PET activities, and plays an important guiding role in college PET. The integration and application of PET and modern educational technology can not only stimulate students' interest in learning, enhance students' intuitive understanding of specific movements, and form a complete concept of movements, but also effectively cultivate students' learning autonomy, fully develop students' abilities, and enable students to master the essentials of sports techniques in a short time and receive more relevant

knowledge.

This article develops a hybrid intelligent digital information quality evaluation system. Text data was used for visual teaching optimization. The innovative contributions of the research are as follows:

1) Digital teaching, driven by a combination of text and visual features, integrates different forms of teaching together.

2) Through the collection and simulation of actual datasets, it is shown that the hybrid intelligent system proposed in this article has brought significant efficiency improvements to digital teaching.

3) An online PET effect evaluation model was constructed by combining video description and NLP technology. A framework for weakly supervised video description and localization based on multi-level intra modal attention reconstruction is proposed. Research has shown that the accuracy of this method is higher than other methods. Under the accuracy index, this method greatly improves the retrieval results of a complete single topic.

2. Related work

The task of describing video with natural language is called video description. Its research combines NLP (natural language processing) with key technologies in the field of computer vision, and the research results promote the development of cross-modal analysis technology. In recent years, more and more researchers are engaged in the research of video description [8,9]. Video description makes it possible for people to understand video content more accurately from the semantic point of view. The research on video description has very important practical significance and many specific applications. Young et al. first used a visual recognition system to obtain the confidence of the targets, actions and scenes in a video, and then combined these detection confidence values with the probabilistic knowledge mined from the text corpus with a factor graph model to estimate the most likely subject, verb, object and place [10]. Taskin et al. proposed a video description model based on LSTM (long short-term memory) [11]. At first, they used a CNN (convolutional neural network) model pre-trained on large-scale object recognition image data sets to extract the frame-level features of videos, and then perform average pooling operations on the frame-level features of each video to obtain the video representation with fixed dimensions. However, this method only considers the characteristics of video frames, ignoring the dynamics and continuity of video. Li proposed a new weakly supervised intensive event description scheme, which input video-multiple description sentences for the training set, and established the relationship between video area and vocabulary tags through a vocabulary-full convolution network and weakly supervised multi-instance multi-tag learning [12]. Jararweh et al. studied various algorithms, and constantly tried to change the weights of feature terms. Based on the algorithm principle of a vector space model, they conducted closed training on data sets, and finally achieved a classification effect with a classification accuracy rate of 97% [13]. Kim et al. proposed a modulation classifier based on a neural network, which can distinguish band-limited modulation signals with noise from 13 types of digital modulation and analog modulation [14]. Clercq et al. used natural language processing to identify innovative technology trends related to food waste treatment, biogas and anaerobic digestion. The methods used include analyzing a large amount of text data mined from 3186 patents related to these three fields [15]. Marder supplemented clinical assessment with NLP and machine learning, providing a potential way

to address the limitations of personal language assessment [16]. Afzal et al. extended the previously validated NLP algorithm for portable Android device (PAD) recognition to develop and validate the sub phenotype NLP algorithm (CLI-NLP) for identifying CLI cases from clinical notes [17]. Krishnamurthy et al. proposed a constrained random context-free grammar (CSCFG) model. They proposed a particle filter algorithm that utilizes the CSCFG model structure to estimate the trajectory of the target [18]. Amaar et al. proposed a method that uses natural language processing and supervised machine learning techniques to detect fraudulent recruitment advertisements from online recruitment portals [19]. Hilal et al. proposed a robust English text watermarking and natural language processing method (RETWNLPA) based on a word mechanism and first-order Markov model to improve the accuracy of sensitive English text tamper detection [20]. Cheng et al. developed and tested a new integration solution for fire progression prediction data models. The scheme combines reduced order modeling, recurrent neural network (long-term and short-term memory), data assimilation and error covariance adjustment. Reduced order modeling and a machine learning surrogate model ensure the efficiency of the proposed method, while data assimilation enables the system to adjust the simulation through observation [21]. Zhang et al. proposed a digital dual fire model based on JULES-INFERNO, which uses ROM technology and deep learning prediction networks to improve the efficiency of global wildfire prediction. The iterative prediction implemented in the proposed model can use current year data to predict fires in the coming years [22]. The data used in the training process of any machine learning (ML) algorithm includes numerical error, approximation error and round-off error, which are trained into the prediction model. The integration of ML and DA improves the reliability of predictions by including physically meaningful information. Buizza et al. introduced data learning, a field that integrates data assimilation and machine learning to overcome the limitations of applying these fields to real-world data [23]. Soffer and Cohen's study examined the characteristics of students' participation in online courses and their impact on academic performance. Using learning analysis methods, four online courses were examined using a similar teaching model ($N_{students}$ = 646). The results showed a significant difference between students who completed the course and those who did not, among all 13 variables [24]. Oodaira et al. solved the challenge of estimating the importance of text in scene images. They constructed a scene image dataset including text and assigned importance to each text through subjective evaluation. Based on subjective evaluation, image features representing the importance of text content were determined, and an importance estimation model was proposed [25]. Previous methods mainly focused on the internal characteristics of users, while ignoring potential external features between users. Guo et al. proposed a deep PR method based on deep distributed learning for MENs. Through representative learning schemes, deep abstraction was performed on hidden feature components from local and global subspaces [26]. Guo et al. proposed a vehicle agent autonomous behavior decision-making framework using network physical social intelligence. First, they established a dynamic programming model with multiple objectives and constraints. This can be embedded into the control unit of the vehicle agent to give it appropriate social intelligence [27]. Hwang et al. adopted an online learning method based on social regulation to help students leverage the power of peers to achieve learning goals. The study conducted experiments on the experimental group and the control group. The experimental group used online learning methods based on social regulation, while the control group used traditional self-regulated learning methods [28]. Li et al. introduced an automatic detection method for fake news in the Chrome environment. This method can detect fake news on Facebook. Specifically, it used multiple features related to Facebook

accounts and some news content features to analyze account behavior through deep learning [29]. The text analysis module uses a mixture of ConvNet neural networks rich in SentiCircle contextual semantics to determine emotions. Zhang et al introduced an aggregation scheme to calculate mixed polarity. A support vector machine (SVM) classifier was trained using visual word bags (BoVW) for predicting visual content emotions [30].

3. Methodology

3.1. Establishment of effect evaluation index system

There must be a certain logical relationship among the indicators of online PET effect evaluation for college students, because the establishment of this indicator system is involved in the whole evaluation system. They not only reflect the main characteristics of college students' learning effect in physical education class from different aspects, but also reflect their internal relationship with their learning effect in physical education class. A comprehensive analysis of many existing factors makes each component of the evaluation system objective, thus truly embodying the characteristics that can effectively promote the teaching effect. Especially in the process of evaluation, it is required to be objective, fair and reasonable, so as to make a fair, just and realistic judgment on teachers' teaching and students' learning.

The information obtained from the evaluation results of online learning can enable the implementers and recipients of online teaching to find out the problems and make timely corrections and remedies, so that online teaching activities can be improved through constant circular corrections, which is more conducive to the smooth and effective conduct of online teaching activities. It is important to make teachers and students realize their own advantages and disadvantages, so they can constantly improve their own teaching process. This can stimulate their competitive desire, and stimulate their internal motivation and subjective initiative, so they can pursue better evaluation results, and thus play a role of teaching and learning. In addition, the evaluation results can encourage developers to design the teaching support platform more optimally.

As the subject of education, students have the most right to objectively evaluate the teaching quality of a teacher. Many universities combine the evaluation results of students with the evaluation results of teaching supervision teams as the basis for judging the teaching quality and level of teachers. Evaluation metrics include whether the teaching covers the content and expected purpose of the course, whether the teacher has comprehensive professional knowledge related to the course, whether the lesson preparation is sufficient and organized, and whether the teacher's teaching attracts students' interest. Through teaching evaluation, we can guide the improvement of teachers' quality, ensure education quality and teaching effect, and promote teachers' teaching ability and level to meet the requirements of university construction. It can be said that almost any comprehensive activity can be comprehensively evaluated. With the continuous expansion of people's fields of activities, the objects to be evaluated are becoming more and more complex. People cannot only consider one aspect of the objects to be evaluated, but must consider the problem from a holistic perspective.

The principle of operability requires that the established evaluation system, especially the selected evaluation indexes, can measure the PET phenomenon scientifically and reasonably, and at the same time, it is convenient to operate and implement. This requires careful analysis, research and precise and clear description of the evaluated object in the process of establishing the evaluation

system. The acceptability principle means that the constructed evaluation system can be accepted by the evaluated object and become the incentive factor of the evaluated object, which requires us to make the evaluation index practical and acceptable when constructing the evaluation system, especially the evaluation index system. According to the standard of evaluation content, it can be divided into primary, secondary and tertiary indicators. The importance of an indicator in the target is the corresponding index weight, the index scale reached by the evaluation object is the corresponding evaluation index, and the final evaluation of the evaluation object is the corresponding evaluation result.

The establishment of the index system should meet the following conditions: consistency with the target, direct measurability, mutual independence of the indexes in the system, overall completeness, comparability and acceptability of the index system. Online PET effect evaluation is a process of understanding and judging the online PET effect evaluation in colleges and universities, which needs comprehensive evaluation according to the objectives. The evaluation index system is simple and easy to operate. It helps the evaluation subject to make a concrete, quantitative and comprehensive evaluation of the evaluation object, and on this basis, obtains a comprehensive evaluation result, which is conducive to reducing the subjective randomness of evaluation and improving the objectivity of evaluation. The corresponding evaluation index and scientific calculation weight are the important support for seeking truth from facts evaluation, which is conducive to the good embodiment of PET values and the more scientific and specific requirements for PET quality and PET activities.

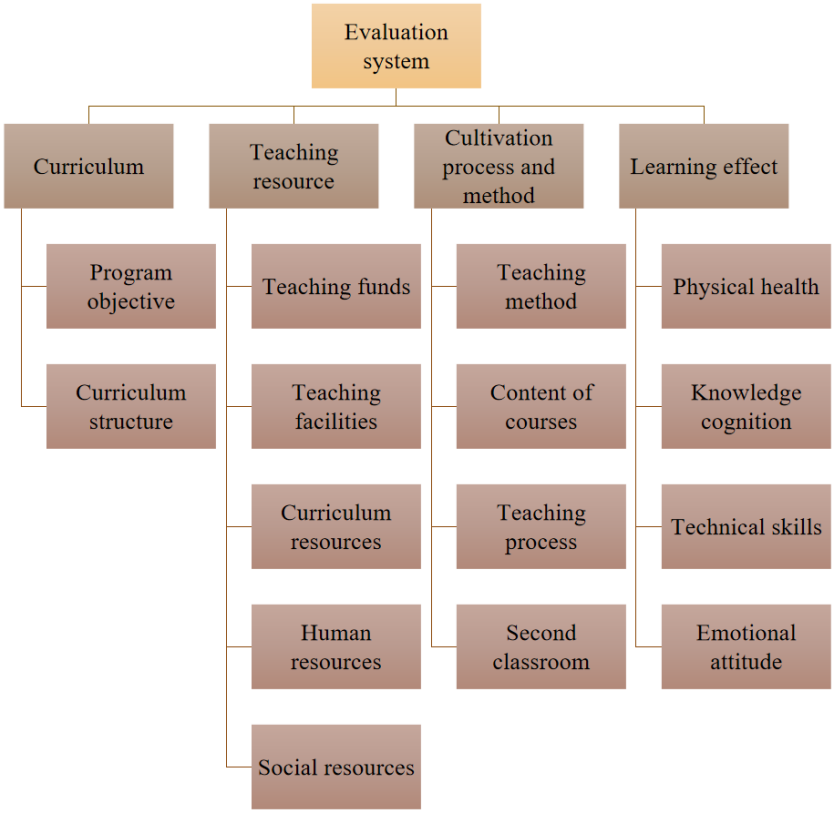

Figure 1. On-line PET effect evaluation system.

The teaching methods and means of modern educational technology are lively, novel and diverse, overcoming the traditional PET mode of teachers demonstrating and explaining at the same time. The use of modern educational technology not only accords with students' psychological characteristics of curiosity and innovation, but also creates better situations and emotional experiences. Combined with the new curriculum standards and the relevant theories and indicators of the new round of PET quality audit and evaluation in colleges and universities, after screening and sorting by experts, we set the first-level indicators into four items: the objectives and structure of curriculum setting, teaching resources, training process and methods, and students' learning effect. The second-level indicators included in each first-level indicator are shown in Figure 1.

According to the mathematical model of attribute evaluation, the weight of an online PET effect evaluation index should be calculated and optimized. Generally, weights can be divided into two categories: subjective weighting methods and objective weighting methods. The subjective weighting method is a method for decision-makers to weight according to subjective judgment of experience or subjective importance of each index. The obvious feature of the objective weighting method is that there are objective criteria for weighting, and the weight coefficient of an evaluation index is obtained by calculation, rather than being given artificially. However, the disadvantage of the objective weighting method is that sometimes the calculation results cannot be explained. They are complementary to each other [15,16]. In this paper, the optimal combination weighting method is used to deal with the weights in online PET effect evaluation.

There are m evaluation indexes and n evaluation objects. According to the principle of combining qualitative and quantitative analysis, a multi-object evaluation matrix about multiple indexes is obtained:

$$R^{'} = \begin{bmatrix} r_{11}^{'} & \cdots & r_{1n}^{'} \\ \vdots & \ddots & \vdots \\ r_{m1}^{'} & \cdots & r_{mn}^{'} \end{bmatrix} \tag{1}$$

where $R^{'}$ is standardized to get:

$$R^{'} = \left(r_{ij}\right)_{m\times n} \tag{2}$$

where r_{ij} is called the value of the j evaluation object above the index.

In the evaluation of (m,n), the entropy weight w_i of the i th index is defined as:

$$w_i = \frac{1-E_i}{m-\sum_{i=1}^{m} E_i} \tag{3}$$

From the information point of view, it represents the amount of useful information provided by this indicator in this problem. We then adjust, increase or decrease the evaluation index according to the entropy weight, so as to make a more accurate and reliable evaluation. At the same time, we can also use entropy weight to adjust the accuracy of some indexes' evaluation values, and, if necessary,

re-determine the evaluation values and accuracy [17].

3.2. Multi-modal information fusion-based smart evaluation method

3.2.1. Video description generation algorithm

Sports professionals make technical analysis through sports videos, study opponents' technical essentials and competition tactics, formulate corresponding competition strategies, and help coaches guide athletes to carry out targeted training, which plays an important role in rapidly improving technical level and competition results. The structure of sports video content is obvious, such as the first and second half of football, shooting, slow motion, etc., and often there are a lot of audience voices in the middle of the game. By analyzing the features of audio, video and subtitles in sports competitions, automatically identifying and extracting the wonderful scenes that the audience is interested in can save the time for the audience to enjoy sports competitions [18,19]. With the rapid increase of multimedia data, it is necessary to adopt corresponding methods and tools to add relevant semantics to multimedia data according to different forms and sources, so as to facilitate the management and use of multimedia content.

Machine learning deep learning is the mainstream way to realize artificial intelligence. Machine learning aims to model tasks in real scenes as mathematical problems, so as to realize further deduction and optimization. At the same time, the achievements in the field of artificial intelligence have inspired scholars to do in-depth research in the field of deep learning, resulting in many excellent research results. Different from the research of a single mode, a multi-mode task often requires the participation of multiple modes and the analysis of multiple modes at the same time to get the final result. Therefore, a multimodal model needs to understand the content of a single modality correctly at first. Video description and location is a multimodal task involving video modality and text modality, and it is an interdisciplinary research project between computer vision and NLP. At the same time, because the text description is more detailed and cannot be preset in advance like the action category, compared with the action detection, the text detection is often more uncertain, and it is necessary to distinguish the video segments corresponding to different time sequences in the video more finely according to the semantic information.

Inspired by adding semantic attributes to the image description task to improve the description results, many researchers try to add semantic attributes to the video description task, but in their work, they usually simply add the most common words in the sentences describing the video. At the same time, pronouns will also be removed for the resolved subject and object, because these words indicate a referential relationship, but they do not clearly indicate what it is. If these pronouns are used to construct a dictionary, it will not only improve the expressive ability of the dictionary, but also introduce noise, which will make the final sentence ambiguous.

Global semantic attributes represent the most common targets and actions in general videos. It involves not only predicting the target semantics in videos, but also the action semantics in videos. Therefore, this paper combines the visual target features and action features of videos to predict the global semantic attributes of videos [20]. Three types of semantic attribute prediction are expressed as follows:

$$g_a = MLP([f, m]) \tag{4}$$

$$o_a = MLP(f) \tag{5}$$

$$v_a = MLP(m) \tag{6}$$

where g_a represents global semantic attribute, o_a represents object semantic attribute, v_a represents action semantic attribute, f represents the feature after averaging video visual target features, m represents the feature after averaging video visual action features, and $[f, m]$ represents the splicing of two features.

The framework of weak supervised video description and location method based on multi-level intra-modal attention reconstruction is shown in Figure 2. The main parts of our model are: video segment sampling and representation module, intra-modal attention module and reconstruction module. These three modules are closely connected, so they can be trained end to end. The core of the video segment sampling and characterization module is to construct a learnable way to obtain the characterization of video segments for later processing. Compared with the object candidate frame in the image, the time series candidate frames in the video have larger variance. By adjusting the convolution kernel to reduce the dimension, the features of video segments in different positions and scales in the video can be obtained in a learning way.

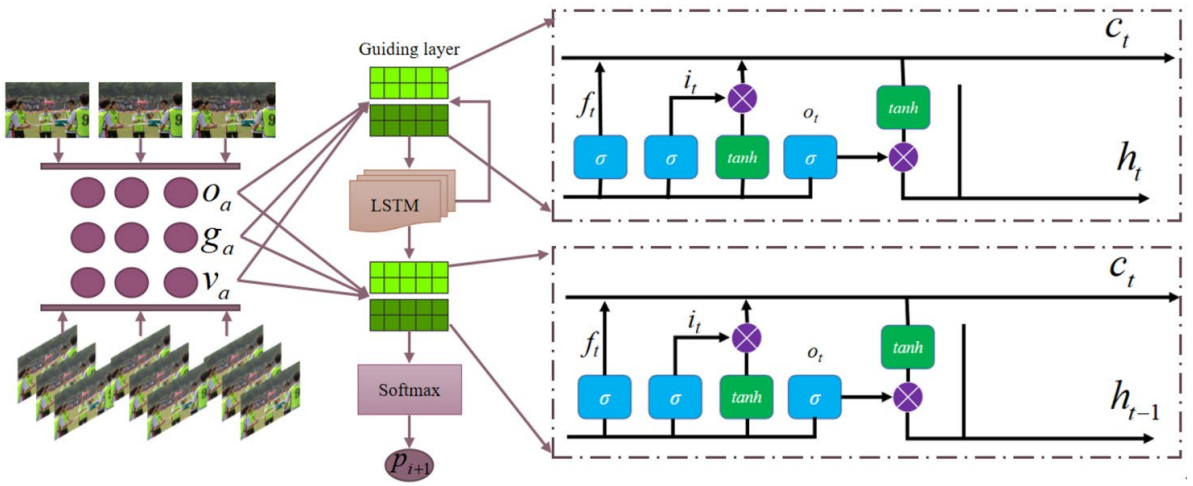

Figure 2. Framework of description and location method for weakly supervised video based on multi-level intra-modal attention reconstruction.

By adjusting the size of the convolution kernel and the number of convolution layers, the receptive field is directly or indirectly increased, corresponding to different forms of intra-modal correlation modeling, and the learned attention score is weighted with the video segment features to obtain the global video feature $F_p^{global} \in R^{l \times d_{vp}}$ associated with the given text:

$$F_p^{global} = F_{vp} \otimes Att_p \tag{7}$$

where $\otimes$ in the formula represents matrix multiplication, and F_{vp} is required when weighting with attention score.

Next, the model of language generation is introduced from the text-side feature construction and decoder part of video description generation. It can be obtained by pre-training the language model in a large number of unsupervised corpora, and to some extent, it can make up for the problems of rare words and out-of-set words caused by small data sets. Because the representation of the original word is only hot coding, the calculation process of the word vector embedding layer can be represented by a look-up table or the following linear transformation:

$$y_t^{emb} = W_{word} y_t + b_{word} \tag{8}$$

where $y_t \in R^{|V|}$ is the unique heat vector of the t word in the target text, $W_{word} \in R^{d_{model} \times |V|}$ is the parameter matrix composed of word vectors of all words, and $|V|$ represents the size of the vocabulary.

First, the language encodes them into fixed-length vector representations, and then the LSTM decoder decodes the vector representations into target sentence sequences. The goal of reinforcement is to minimize the expected sentence-level reward loss:

$$L_{SG} = -E_{S \sim p_\theta}[r(S)] \tag{9}$$

where θ is the LSTM network parameter of sentence generation, and the generation strategy of LSTM is p_θ. $r(S)$ represents the return of the generated sentence S and the standard sentence through the evaluation index.

We get the shot vector $S = (S_1, S_2, \cdots, S_n)$, which represents all the shot sequences corresponding to a player. The shot sequence of the i th athlete can be expressed as $S_i = (S_{i1}, S_{i2}, \cdots, S_{in}), \quad i = 1,2,\cdots,n$. In this way, we can get the expression of the video corresponding to a game as follows:

$$M = \sum_{i=1}^{n} P_i (S_{i1}, S_{i2}, \cdots, S_{in})^T \tag{10}$$

Based on this definition, we organize and classify video shots.

3.2.2. Multi-mode integration of sports video analysis

Although some preliminary information can be obtained by automatically extracting the exciting events in sports videos, such as dividing the video content into two parts, exciting and non-exciting, it is impossible to analyze the structure of the sports competition itself and related events, such as dividing the diving competition into several rounds and identifying the athletes' appearances and diving events. There are some semantic events in sports videos, such as diving in diving competitions and shooting in football matches. These semantic events are often the most valuable parts of the video, which need to be labeled for easy retrieval. Target events, such as diving

of athletes in diving competitions, shooting in football competitions, etc., refer to specific sports with ornamental features in sports competitions, which are usually represented by objects and their sports relations.

The channel modeling of the system mainly depends on the communication conditions of the actual channel. Although traditional channel modeling can capture some environmental characteristics, it is very difficult to reproduce the complex communication environment. The classical communication system is divided into several independent modules defined manually, and each module is optimized independently to solve the problem of signal transmission in the channel. Although optimizing each module can improve the overall communication performance, the optimal solution combination of each module does not necessarily generate the overall optimal solution. Language and symbols are closely related. A sentence is composed of multiple word symbols, and a paragraph of text and dialogue is composed of multiple sentence symbols. Word embedding is a distributed representation of words or phrases. A word embedding vector set can be learned from a large corpus. According to the probability of word-word co-occurrence, the objective function is trained, and the word vectors of words and contexts are obtained by making the dot product of two words vectors approach the logarithm of their common occurrences.

Traditional video content analysis methods extract objective perceptual features, but users often consume semantic content, which leads to the contradiction between computer automatic analysis and users' needs. With the help of domain knowledge, how to obtain users' high-level semantic information from low-level video features that can be automatically or semi-automatically obtained by computer is the focus of this paper. In this paper, hierarchical content association technology is specifically used to refer to such a frontier research field, which comprehensively utilizes the research results of digital signal processing, image understanding, pattern recognition, machine learning and other fields; and with the assistance of video production perception rules and domain knowledge, it completes the automatic recognition of computable semantics in video. When all kinds of single media are obtained, the synthesis relationship between the single media should be considered in the generation of video. In the final consumption process, the user perception model needs to be considered. Therefore, hierarchical content association technology needs the assistance of video production perception rules and domain knowledge.

The central task of NLP is to turn the inquiry and information of potential meaning and vague natural language into unambiguous internal expression, and to match and search based on this expression. In NLP-based retrieval systems, the most commonly used ones are analysis and processing at the lexical, syntactic and semantic levels, while other higher-level analysis is in the research stage. Traditional machine learning methods can be divided into supervised learning and unsupervised learning. In learning SVM, we should first learn what a linear classifier is. Let us record the data points as x and the categories as y. If we can get a hyperplane in the N -dimensional feature space, its equation is as follows:

$$w^T x + b = 0 \tag{11}$$

Because the high-frequency words in the training corpus contain more information, the word vectors of high-frequency words should be given higher weight. In this paper, let N be the number of words, k be the word frequency ranking position, and the assumed frequency of words with word frequency ranking of k is:

$$f(k,N)=\frac{\frac{1}{k}}{\sum_{n=1}^{N}\frac{1}{n}} \tag{12}$$

When combining word vectors, we use the assumed frequency of this word to weight the word vectors. Entries are made up of single words. Therefore, in a text document, the more times two adjacent words or strings appear, the greater the probability of combining them into one entry. We define the co-occurrence information of two terms as the probability of the co-occurrence of two terms, which is expressed by the formula as follows:

$$I(W_1,W_2)=\log\frac{P(W_1,W_2)}{P(W_1)P(W_2)} \tag{13}$$

where $P(W_1,W_2)$ is the probability that the words W_1,W_2 are adjacent to each other and appear together, and $P(W_1),P(W_2)$ respectively represents the probability that W_1,W_2 appears in the corpus. The mutual information shows how closely the entries are combined.

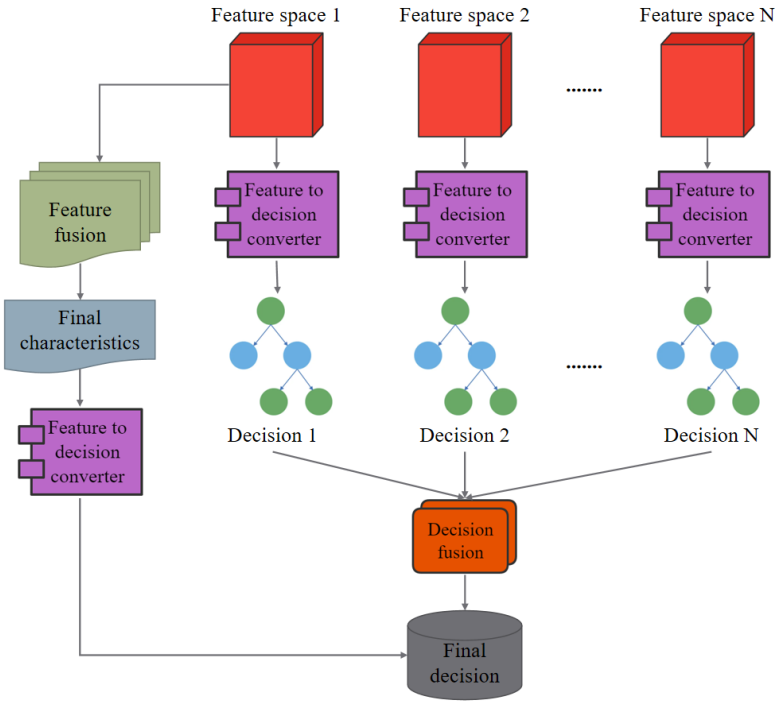

Figure 3. Feature fusion and decision fusion of multimodal information.

In the feature fusion method, feature vectors from different patterns are concatenated together to form a final feature, and then the final decision (classification result) is obtained by a decision maker (classifier), as shown in Figure 3. Because feature merging will lead to a higher-dimensional feature vector, in order to avoid the computational overhead caused by high-dimensional operation and the complexity of the classifier, feature dimensionality reduction transformation is usually needed after feature merging. That is, the semantic events at the bottom level can also be regarded as primitives

relative to the semantic events at the higher level, thus forming a multi-level relationship between events.

According to the principle of local feature invariance, the feature map obtained from the convolution result of the previous layer is sampled to further extract features, and at the same time, the size of the feature map is reduced, thus reducing the computational complexity. In this paper, the sampling method is used to extract the eigenvalues of the characteristic graph obtained by a specific convolution kernel:

$$\hat{s} = Max\{s\} = Max\{s_1, s_2, \cdots, s_{n-w+1}\} \tag{14}$$

At the same time, the two kinds of sampling can be adapted to different sizes of feature maps. Given two sequences X, Y, the similarity of TF-IDF (term frequency—inverse document frequency) indicates the sum of TF-IDF of common elements in two sequences in Y:

$$S(X,Y) = \sum_{t \in X \cap Y} c(t,Y) \log\left(\frac{N}{N_t + 1}\right) \tag{15}$$

where N represents the total number of sample sequences in the database, and N_t represents the number of sample sequences containing element t.

In traditional teaching methods, the mid-term and exams are often used as the standard to evaluate students' learning effect. Therefore, this paper takes the mid-term and final exams as a reference, and analyzes the correlation between other indicators and the results of the mid-term and final exams. During the collection process, 40 valid samples were collected by group evaluation. Using the Pearson correlation coefficient method, the following results were obtained (Table 1 and Figure 4). The results show that the indicators selected in this paper reflect the students' mid-term and final exam results to a certain extent, so teachers can track and supervise students according to their process indicators. For example, if a student's preview times before class are reduced, teachers can remind the student in time. Teachers can also encourage students to participate in class discussions, and encourage students to think more and ask questions to improve their learning effect.

Table 1. Correlation coefficients between process index and mid-term and final grades.

Index	Midterm examination	Final exam
Preview times	0.35625	0.72778
Preview video viewing time	0.45514	0.42727
Online answer situation	0.4782	0.54935
Discussion situation	0.201	0.70016
Classroom questioning times	0.49092	0.66212
Evaluation of learning effect	0.36486	0.80041
Satisfaction evaluation	0.36388	0.69739
Adaptability evaluation	0.46148	0.68014
Amount of speech data in discussion area	0.53811	0.4393
Number of questions in online discussion forum	0.3684	0.77186
Self-assessment	0.32922	0.54611
Group evaluation	0.4266	0.31601
Teacher evaluation	0.62539	0.49327

In order to extract the motion features, we calculate the motion vectors on each block in the image by block matching, and then calculate the motion intensity:

$$m = \frac{1}{K}\sum\sqrt{v_x^2 + v_y^2} \tag{16}$$

where K is the total number of motion vectors $v = \lfloor v_x, v_y \rfloor$.

4. Analysis and discussion of results

The traditional methods for feature extraction in computer vision mainly include the following:

1) Feature extraction methods based on manual design, including SFT, SURF, HOG, LBP, etc. These methods extract important features from the image by describing local features, textures, shapes, and other aspects of the image.

2) Feature extraction methods based on convolutional neural networks. CNN can automatically learn advanced features from original images.

3) Methods based on feature encoding, such as bag of words (BoW), Fisher vectors, and VI AD methods.

4) Deep learning based methods, including autoencoders, deep Boltzmann machines, convolutional autoencoders, variational autoencoders, etc. These methods learn the low dimensional representation of input images to obtain more robust and effective features.

Due to the differences between their different forms and approaches, integrating them together remains a real challenge when evaluating the quality of digital teaching. In fact, both types of information can reflect different knowledge from their respective perspectives. To fill this gap, this article proposes a hybrid intelligent system for digital teaching quality evaluation driven by text and visual features.

In this part, we will introduce the data set used: The Charades data set contains more than 10,000 videos with 157 types of actions. The Net-caption data set is the largest data set under the task of video description and positioning. There are more than 20,000 videos, including 100,000 video segment-sentence pairs. For the video segments with the same scale s, we set the sampling step size of $1/4s$, so that the overlap between video segments is reduced to 75%. Super-parameter $\alpha = 0.2$, and $\lambda = 1.5$ of the multilevel mechanism. The data visualization is shown in Figure 5. Table data visualization is shown in Figure 6. Firstly, we verify the effect of intra-modal attention module, and use the 2D convolution layer with convolution kernel of 3 × 3 to obtain the attention score of each video segment, which takes into account the video timing correlation.

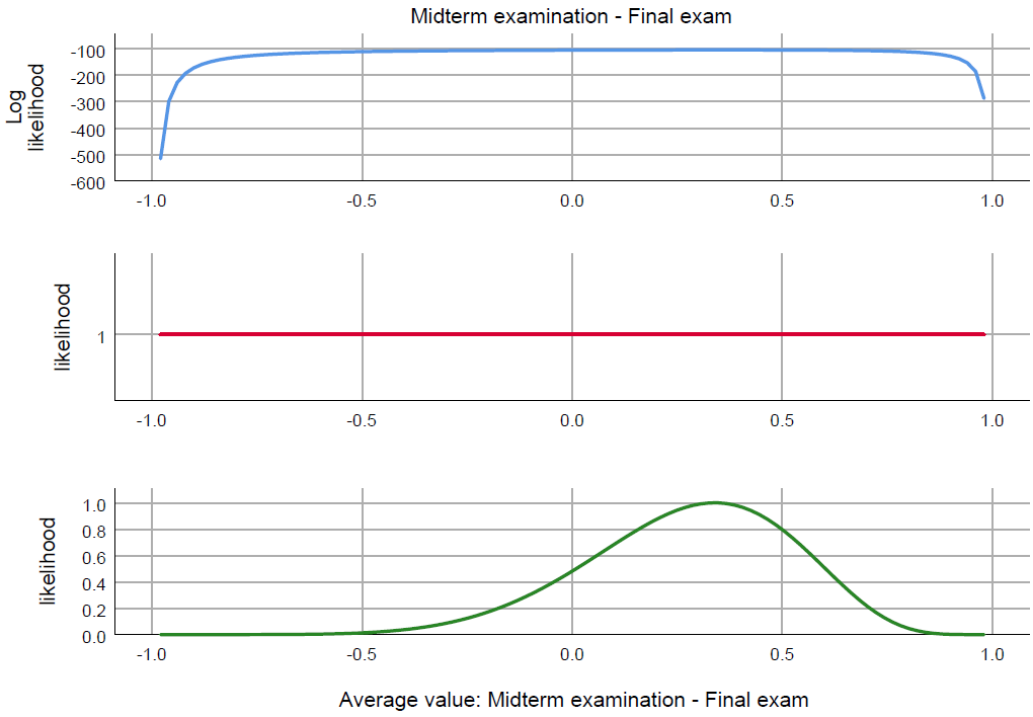

Figure 4. Posterior distribution characteristics of pairwise correlation.

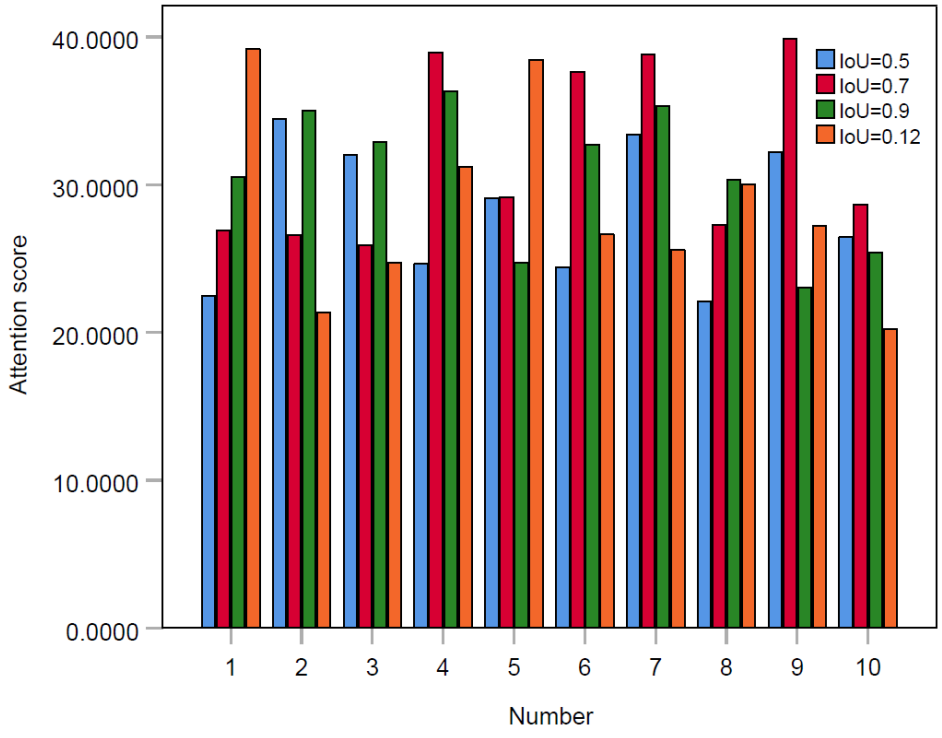

Figure 5. Charades-STA dataset data visualization diagram.

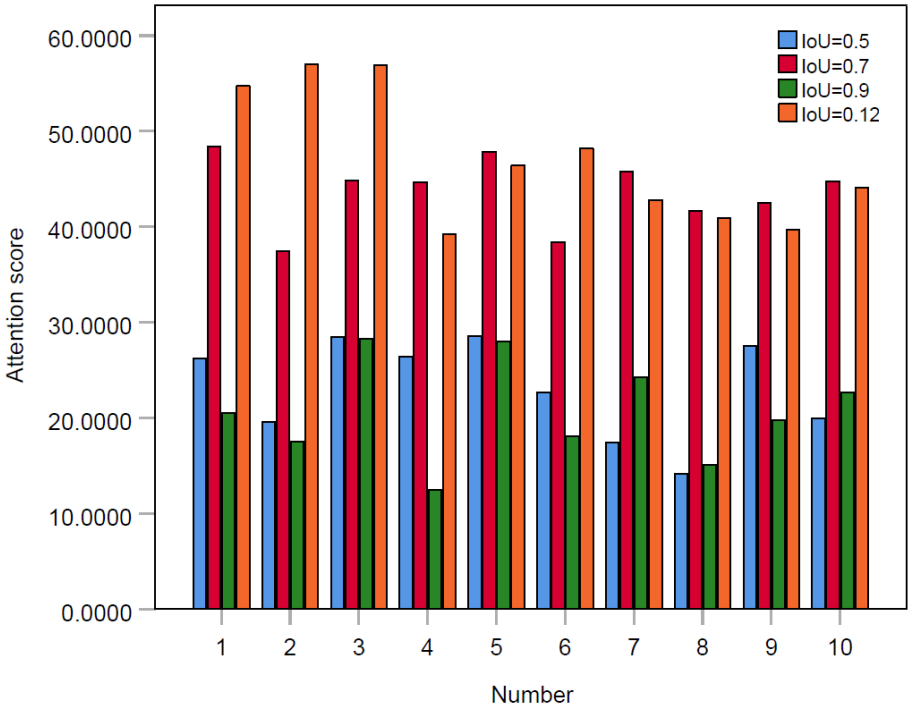

Figure 6. ActivityNet-Caption dataset data visualization diagram.

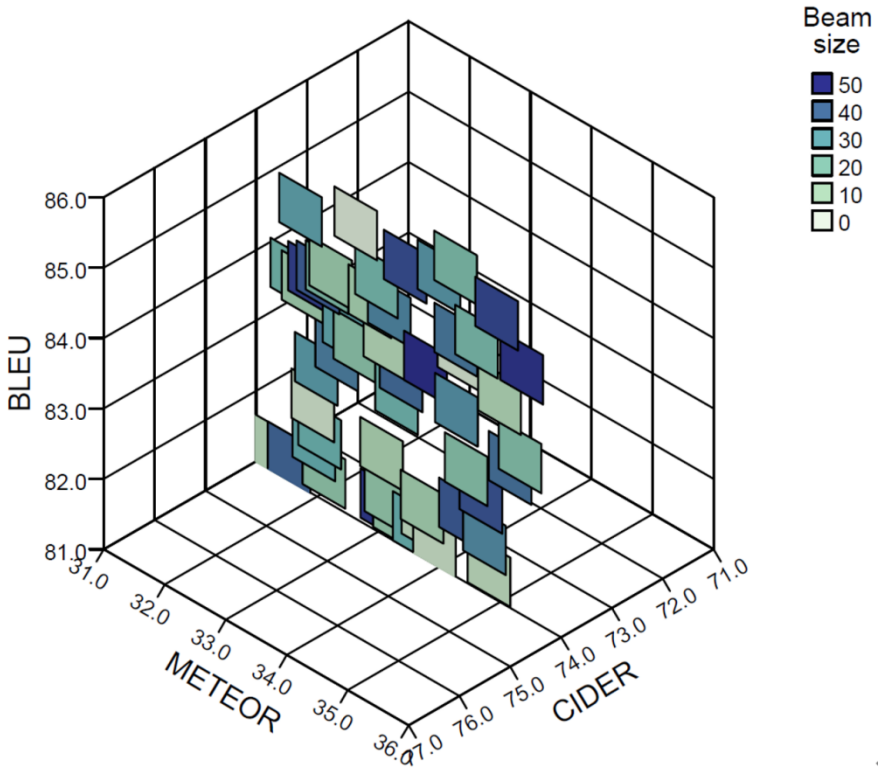

Figure 7. Comparison of results using different Beam size.

Comparing the experimental results of the first group and the second group shows that

multi-modal fine-grained association modeling is helpful to improve the effect of the weak supervised video description and location task. At the same time, by comparing the experimental results of the second group and the third group, it can be seen that when modeling the intra-modal correlation degree, selecting the appropriate length of correlation window has great influence on the final result.

This paper also studies the influence of different bundle sizes on the result of sentence generation in the test set. Figure 7 is obtained by changing the cluster size on the basis of the LSTM model. When the cluster size is 1, it is equivalent to the result obtained by a greedy algorithm. Every time, the word with the highest probability will be selected to generate a sentence, because it excludes all other possibilities.

Figures 8 and 9 show the comparison between the proposed method and other methods. It can be seen that the precision of the proposed method is higher than other methods. Under the precision index, the retrieval results of complete single-choice questions are greatly improved by the proposed method. Compared with CNN, the results of the proposed method are improved by 15.018%. This shows that the statistical query method is not applicable to either incomplete query or complete query.

The evaluation index system of college online PET effect is a multi-dimensional and dynamic model. In the process of popularization and use in various universities, it should be constantly adjusted, repeatedly used, revised and improved according to the actual situation. It is important to strengthen the supervision and management of the online PET effect evaluation system in universities, strengthen the organization and leadership of evaluation management, optimize and improve the evaluation procedures and methods, ensure the transparency of online PET effect evaluation, and strengthen the incentive assessment mechanism.

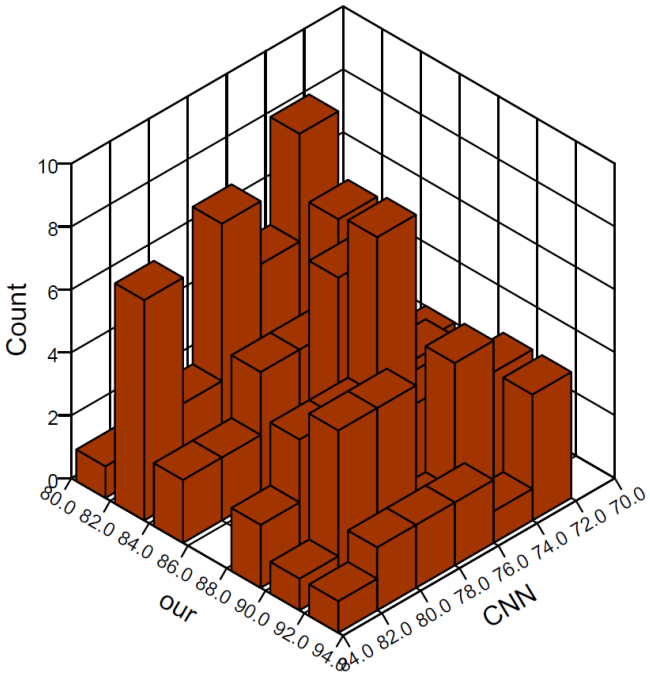

Figure 8. Precision comparison of incomplete query multiple choice questions.

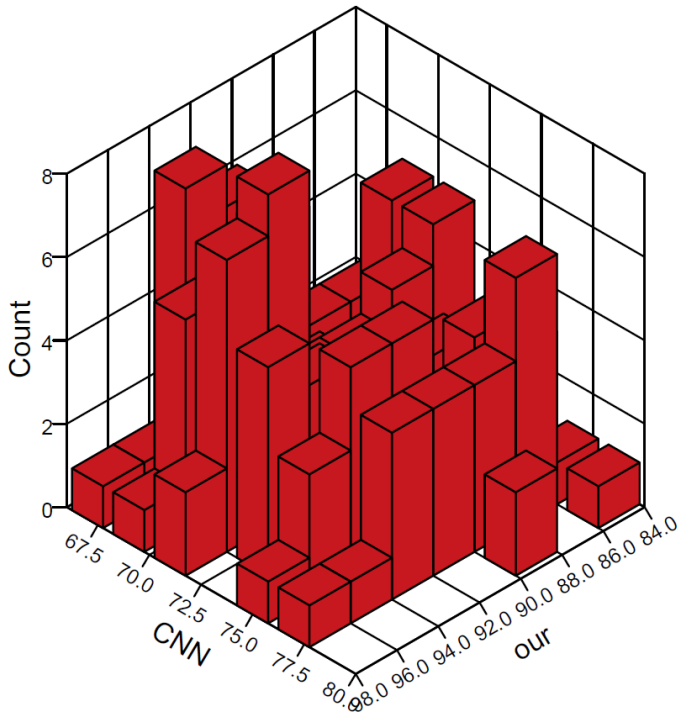

Figure 9. Precision comparison of complete query multiple-choice questions.

The content and proportion of evaluation indicators should be open and transparent. The collection process of evaluation information should also be supervised and controlled, the original information should be checked and disclosed, and supervision should be carried out through various channels and ways. Only in this way can it be open and fair, and can it get the support and cooperation of the evaluated unit.

5. Conclusions

Online PET effect evaluation is an integral part of educational science. It is a process that provides a basis for optimizing decision-making in physical education classroom teaching. This process takes physical education classroom teaching as a countermeasure, adopts all feasible evaluation techniques and means, systematically collects teaching information and makes Value judgment. Establishing a systematic and scientific evaluation system to determine reasonable evaluation indexes and giving corresponding normalized weights is the key to the success or failure of online PET effect evaluation and result value judgment. In this study, an online PET effect evaluation model combining video description and NLP technology was constructed. A description and location method framework of weakly supervised video based on multi-level intra-modal attention reconstruction was proposed. The research shows that the precision of the proposed method is higher than other methods. Under the precision index, the retrieval results of complete single-choice questions are greatly improved by the proposed method. Compared with CNN, the results of the proposed method are improved by 15.018%. This shows that the traditional statistical query method is applicable to neither incomplete query nor complete query.

However, the existing work mainly relies on the attention mechanism to predict the position of

the target corresponding to the noun while generating the image description through the regularization technology. There is a significant gap in the performance of these methods compared to fully supervised image description and localization. Most methods that rely on attention mechanisms often focus on the most discriminative local position of the target, which cannot fully predict the overall content of the target. Based on this, we will also conduct effective research and analysis on the mechanism of distributed attention in the future. After selecting multiple targets with the same semantics but not completely overlapping positions, one can aggregate them to obtain a more complete target position.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgment

This work was supported by The 2020 Huzhou University humanities and Social Sciences Research Project: From the birthplace of ancient silk to the birthplace of modern ethnic sports -- A feasibility study on the development of modern short-arm sports in Huzhou (No. 2020SKYY22).

Conflict of interest

The authors declared that there was no conflict of interests.

References

1. Z. Guo, K. Yu, A. K. Bashir, D. Zhang, Y. D. Al-Otaibi, M. Guizani, Deep information fusion-driven POI scheduling for mobile social networks, *IEEE Network*, **36** (2022), 210–216, https://doi.org/10.1109/MNET.102.2100394
2. Z. Shen, F. Ding, Y. Yao, A. Bhardwaj, Z. Guo, K. Yu, A privacy-preserving social computing framework for health management using federated learning, *IEEE Trans. Comput. Soc. Syst.*, (2022), 1–13, https://doi.org/10.1109/TCSS.2022.3222682
3. C. D. Wei, C. Liu, W. Shun, S. Wang, X. L. Wang, W. F. Wu, Research and application of multimedia digital platform in the teaching of college physical education course, *J. Intell. Fuzzy Syst.*, **34** (2018), 893–901. https://doi.org/10.3233/JIFS-169383
4. Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, A. Beheshti, Mixed graph neural network-based fake news detection for sustainable vehicular social networks, *IEEE Trans. Intell. Trans. Syst.*, (2022), 1–13. https://doi.org/10.1109/TITS.2022.3185013
5. J. De-Kun, F. H. Memon, Design of mobile intelligent evaluation algorithm in physical education teaching, *Mobile Networks Appl.*, **27** (2021), 527–534. https://doi.org/10.1007/s11036-021-01818-1
6. T. Kidokoro, Y. Kohmura, N. Fuku, Y. Someya, K. Suzuki, Secular trends in the grip strength and body mass index of sport university students between 1973 and 2016: J-fit$^+$ study, *J. Exercise Sci. Fitness*, **18** (2020), 21–30. https://doi.org/10.1016/j.jesf.2019.08.002

7. D. Vinnikov, Z. Romanova, A. Dushpanova, K. Absatarova, Z. Utepbergenova, Prevalence of supplement use in recreationally active kazakhstan university students, *J. Int. Soc. Sports Nutr.*, **15** (2018), 16. https://doi.org/10.1186/s12970-018-0220-4
8. D. Li, L. Deng, Z. Cai, K. Cai, Design of intelligent community security system based on visual tracking and large data natural language processing technology, *J. Intell. Fuzzy Syst.*, **38** (2020), 7107–7117. https://doi.org/10.3233/JIFS-179789
9. C. Themistocleous, K. Webster, A. Afthinos, K. Tsapkini, Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing, *Am. J. Speech-Language Pathol.*, **30** (2021), 466–480. https://doi.org/10.1044/2020_AJSLP-19-00114
10. T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, *IEEE Comput. Intell. Mag.*, **13** (2018), 55–75. https://doi.org/10.1109/MCI.2018.2840738
11. Z. Taskin, U. Al, Natural language processing applications in library and information science, *Online Inf. Rev.*, **43** (2019), 676–690. https://doi.org/10.1108/OIR-07-2018-0217
12. H. Li, Deep learning for natural language processing: Advantages and challenges, *Natl. Sci. Rev.*, **5** (2018), 24–26. https://doi.org/10.1093/nsr/nwx110
13. Y. Jararweh, M. Al-Ayyou, E. Benkhelifa, Advanced arabic natural language processing (ANLP) and its applications: Introduction to the special issue, *Inf. Process. Manage.*, **55** (2019), 259–261. https://doi.org/10.1016/j.ipm.2018.09.003
14. H. Kim, S. Lee, A video captioning method based on multi-representation switching for sustainable computing, *Sustainability*, **13** (2021), 2250. https://doi.org/10.3390/su13042250
15. D. D. Clercq, Z. Wen, Q. Song, Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach, *Sci. Total Enviro.*, **673** (2019), 402–413. https://doi.org/10.1016/j.scitotenv.2019.04.051
16. S. R. Marder, Natural language processing: Its potential role in clinical care and clinical research, *Schizophr. Bull.*, **48** (2022), 958–959. https://doi.org/10.1093/schbul/sbac092
17. N. Afzal, V. P. Mallipeddi, S. Sohn, H. Liu, R. Chaudhry, C. G. Scott, et al., Natural language processing of clinical notes for identification of critical limb ischemia, *Int. J. Med. Inf.*, **111** (2018), 83–89. https://doi.org/10.1016/j.ijmedinf.2017.12.024
18. V. Krishnamurthy, S. Gao, Syntactic enhancement to vsimm for roadmap based anomalous trajectory detection: A natural language processing approach, *IEEE Trans. Signal Process.*, **66** (2018), 5212–5227. https://doi.org/10.1109/TSP.2018.2866386
19. A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, S. Ludi, Detection of fake job postings by utilizing machine learning and natural language processing approaches, *Neural Process. Lett.*, **54** (2022), 2219–2247. https://doi.org/10.1007/s11063-021-10727-z
20. A. M. Hilal, F. N. Al-Wesabi, A. Abdelmaboud, M. A. Hamza, M. Mahzari, A. Q. A. Hassan, A hybrid intelligent text watermarking and natural language processing approach for transferring and receiving an authentic english text via internet, *Comput. J.*, **65** (2021), 423–425. https://doi.org/10.1093/comjnl/bxab087
21. S. Cheng, I. C. Prentice, Y. Huang, Y. Jin, Y. K. Guo, R. Arcucci, Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting, *J. Comput. Phys.*, **464** (2022), 111302. https://doi.org/10.1016/j.jcp.2022.111302

22. C. Zhang, S. Cheng, M. Kasoar, R. Arcucci, Reduced order digital twin and latent data assimilation for global wildfire prediction, *EGUsphere*, (2022), 1–24. https://doi.org/10.5194/egusphere-2022-1167
23. C. Buizza, C. Q. Casas, P. Nadler, J. Mack, S. Marrone, Z. Titus, et al., Data learning: integrating data assimilation and machine learning, *J. Comput. Sci.*, **58** (2022), 101525. https://doi.org/10.1016/j.jocs.2021.101525
24. T. Soffer, A. Cohen, Students' engagement characteristics predict success and completion of online courses, *J. Comput. Assist. Lear.*, **35** (2019), 378–389. https://doi.org/10.1111/jcal.12340
25. K. Oodaira, T. Miyazaki, Y. Sugaya, S. Omachi, Importance estimation for scene texts using visual features, *Int. Inf. Sci.*, **28** (2022), 15–23. https://doi.org/10.4036/iis.2022.A.06
26. Z. Guo, K. Yu, N. Kumar, W. Wei, S. Mumtaz, M. Guizani, Deep-distributed-learning-based poi recommendation under mobile-edge networks, *IEEE Int. Things J.*, **10** (2022), 303–317. https://doi.org/10.1109/JIOT.2022.3202628
27. Z. Guo, D. Meng, C. Chakraborty, X. Fan, A. Bhardwaj, K. Yu, Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence, *IEEE Trans. Comput. Soc. Syst.*, (2022). https://doi.org/10.1109/TCSS.2022.3212864
28. G. Hwang, S. Wang, Chiu-Lin Lai, Effects of a social regulation-based online learning framework on students' learning achievements and behaviors in mathematics, *Comput. Educ.*, **160** (2021), 104031. https://doi.org/10.1016/j.compedu.2020.104031
29. Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, K. Yu, Smart assessment and forecasting framework for healthy development index in urban cities, *Cities*, **131** (2022), 103971. https://doi.org/10.1016/j.cities.2022.103971
30. Q. Zhang, Z. Guo, Y. Zhu, P. Vijayakumar, A. Castiglione, B. B. Gupta, A deep learning-based fast fake news detection model for cyber-physical social services, Pattern Recognition Letters, **168** (2023), 31–38. https://doi.org/10.1016/j.patrec.2023.02.026

AIMS Press