Application of Short Video Description Technology in College English Teaching

Xiaoyan Shi, Jilin Province Economic Management Cadre College, China*

ABSTRACT

In order to avoid students' negative learning mood, contemporary teachers are required to abandon the application of spoon-feeding teaching method in English classroom teaching, adopt micro-class teaching method, highlight the teaching characteristics of being close to the people, and create an efficient, short, and special teaching space to meet students' learning needs. In this study, short video description technology is applied to college English teaching, and a generation model of short video natural language description based on Attention mechanism is established. The video feature sequence may be out of sync with the generated word sequence, that is to say, the order of objects and behaviors appearing in the video may be different from their positions before and after the description sentence. In this article, a new generation model of short video natural language description based on attention mechanism is designed.

KEYWORDS

English Education, Neural Network, Short Video, Video Description

INTRODUCTION

For any college English teacher, it is of great significance to deeply explore the application of new teaching modes in college English teaching. Different from traditional teaching forms, micro-courses mainly refer to courses in which the teachers use a video recording to highlight key knowledge or difficult knowledge according to the specific requirements of practical guidance or teaching standards (Li, 2021). However, in the current context of junior high school English teaching, an important teaching error lies in the fact that some teachers focus too heavily on improving student achievement and ignore the important communicative function of English as a language subject, which greatly suppresses students' interest in learning and seriously affects the English teaching effect (Li, 2019). In order to avoid students' negative attitude toward learning, English language teachers must abandon traditional teaching methods, adopt the use of micro-courses, focus on connecting with students, and create an efficient and unique teaching space to meet students' learning needs. Therefore, finding a new teaching mode and structure through exploration and practice to expand students' exposure to English, create the best English learning situation, and stimulate students' interest in learning has become one of the key factors for students to achieve good teaching results.

DOI: 10.4018/IJWLTT.338717

*Corresponding Author

Short video technology organically integrates language learning and cultural background knowledge learning. It organically combines phonetic learning with appreciating plots and pictures, which can effectively improve students' interest in learning English. As a language discipline, the learning goal of English should be for students to use it as a communication tool and to achieve the comprehensive development of listening, speaking, reading, and writing. In order to improve teaching efficiency and obtain higher test results, teachers in traditional classrooms usually transfer knowledge to students in one way. Students rarely have the opportunity to interact with teachers in language, and it is difficult to create an effective language learning environment. Applying short videos to English teaching can create an active and interesting classroom atmosphere. The pictures, sounds, scenes, and other aspects in the short video will have a sensory impact on students, attract students' attention, and enhance students' enthusiasm for learning English, achieving the effect of teaching and entertainment. Students can participate in the process of making short videos. The process of making short videos can improve pronunciation, improve their own accent standards, and enhance their confidence in speaking English through imitation of standard English accents.

A video consists of a group of shots, and the reprocessing of this video clip usually cuts these shots, so it is important to explore the shot structure of videos through video content analysis. As a language subject, English cannot be interpreted perfectly if it is taught only through traditional classroom methods (Li et al., 2018; Bera et al., 2021). The short video teaching can build a good language environment for students, encourage them to feel the connotation of English in real scenes, and strengthen their own learning by imitating English accents, which will help students strengthen their oral expression abilities. Moreover, even if there is oral English teaching, only some simple communicative situations are set up, and the teaching content is limited to exercises in textbooks, which limits the space for students' oral English development, hinders students' development, and further affects students' interest in English learning. This paper begins with an overview of the main characteristics of short video technology, then analyzes the value of short video technology in college English teaching, and designs a new generation of short video natural language model based on attention mechanism. Various modal features are input into an independent encoder for calculation, which can better learn the underlying semantic information contained in various features.

RELATED WORK

Short video technology is a novel and promising research field that aims to automatically generate natural language descriptions for short videos. It has various potential applications in education, entertainment, social media, and accessibility. In this paper, we review the recent literature on short video description technology and its application in college English teaching. We focus on the following aspects: (a) the challenges and opportunities of short video description technology; (b) the main methods and models for short video description generation; (c) the evaluation metrics and datasets for short video description; and (d) the pedagogical implications and benefits of using short video description technology in college English teaching. Short video description technology faces several challenges, such as dealing with complex visual scenes, diverse linguistic expressions, cross-modal alignment, temporal coherence, and domain adaptation (Chen, 2020). However, it also offers many opportunities for advancing natural language processing, computer vision, multimodal learning, and human-computer interaction. Recent methods for short video description generation can be broadly categorized into two types: encoder-decoder models and transformer-based models (Gan, 2017). Encoder-decoder models use separate encoders to extract features from different modalities (e.g., visual, audio, textual) and a decoder to generate descriptions based on these features. Transformer-based models use a single transformer network to jointly encode and decode multimodal inputs and outputs. Both types of models often employ attention mechanisms to learn the relevance between different modalities or sequences. The evaluation of short video description generation is usually conducted using automatic metrics (e.g., bleu, rouge, meteor) or human judgments (e.g., fluency, relevance, informativeness). However, these metrics have limitations in capturing the quality and diversity of generated descriptions (Li, 2018). Therefore, some researchers have proposed new metrics or frameworks that consider multiple aspects of evaluation (e.g., accuracy, diversity, novelty). The datasets for short video description are mainly collected from online platforms or movies. They vary in size, domain, language, and annotation style. Some datasets also provide additional information such as audio tracks, subtitles, or scene boundaries (Liu, 2019).

The application of short video description technology in college English teaching can enhance students' interest, motivation, and engagement in learning English. It can also improve students' listening, speaking, reading, and writing skills by exposing them to authentic and diverse English materials and providing them with feedback and guidance. Moreover, it can foster students' critical thinking, creativity, and collaboration skills by involving them in activities such as creating or editing short videos or descriptions.

Some examples of using short video description technology in college English teaching are: (a) using generated descriptions as captions or summaries for short videos; (b) using generated descriptions as prompts or scaffolds for writing exercises; (c) using generated descriptions as stimuli or references for oral presentations; and (d) using generated descriptions as sources or targets for translation tasks.

Application of short video to English teaching is an innovative development. The new micro-course teaching mode based on short video teaching has gradually attracted national attention and led to curriculum design and research in model schools in various provinces and cities (Hari et al., 2024). Taskin et al. (2019) claimed that the essence of micro-curriculum lies in practicing teaching objectives and learning methods in a shorter time and forming a more focused learning experience. Hang (2018) asserted that there are some limitations in micro-courses, stating that it is difficult for micro-courses to meet the needs of teaching for many complicated courses that require detailed explanation or excessive content. He argued that the purpose of micro-courses is not to teach students everything, but to let students use micro-courses to build their own knowledge system. Jararweh et al. (2019) explored the importance of using short videos in English teaching; they concluded that using short videos can enrich classroom teaching content on weekdays and that using rich teaching resources can stimulate students' thirst for knowledge in learning English, comprehensively improve students' abilities in listening, speaking, reading and writing, and promote the revision of English teaching methods and optimize teaching ideas. Huang et al. (2020) created an English learning video library, which aims to guide English listening, speaking, pronunciation, grammar, and other learning, and divided it into different topics. Zeng et al. (2022) supplemented or expanded Richards' model by combining the research results of pragmatics, psycholinguistics, and linguistics. This model takes "words" as the center to describe related knowledge elements. Luo and Chong (2020) claimed that if subjects are not told prior to performing a task that they will take a vocabulary test afterwad, then their vocabulary learning is incidental. On the contrary, if subjects are informed there will be a test, they engage in intentional vocabulary learning. Zhang et al. (2020) proposed a short video natural language description generation model based on conditional random field (CRF). Firstly, semantic packets containing various objects and behavior tags are generated, then semantic representations are obtained by CRF, and finally semantic representations are connected into sentences by machine translation. However, this model is only suitable for several known behaviors in a specific scene and cannot adapt to the task of video description in open domain. Aguirre et al. (2021) studied a drop-out classifier with better generalization ability, which can predict many short video courses, innovatively used this classifier to implement intervention based on questionnaire survey and reported the results. Chen et al. (2021) extracted the spatial information and time information of the video based on the dual-stream neural network model, and the accuracy of this method on the data set reached 90.6%. Because the video features are encoded, this kind of algorithm requires fewer computing resources and can improve the training speed.

MAIN THEORETICAL VIEWPOINTS

The theoretical background of this study is rooted in pedagogical theories, applied linguistics, and technology in education. In the field of pedagogy, constructivism is a key theoretical idea that supports the use of student-centered and interactive teaching methods in order to foster students' active learning and cultivate their critical thinking skills. In the field of applied linguistics, communicative language teaching is a widely used approach that emphasizes the importance of authentic communication in language learning and teaching.

Moreover, with the development of technology in education, the use of multimedia and digital tools has become increasingly prevalent in language teaching. Short video description technology, as a new learning tool, has the potential to promote learners' engagement and motivation in the learning process. In addition, the attention mechanism, a key technique in natural language processing, can effectively capture the underlying semantic correlations between different features, making it a suitable choice for modeling short video description generation.

Overall, the theoretical background of this study is based on the principles of constructivism, communicative language teaching, and the application of technology in education, which guide the design and implementation of short video description technology in college English teaching.

Currently, there is still a lack of research that explores the application of short video description technology in college English teaching. Therefore, this article examines how to effectively integrate short video description technology into college English teaching, create an efficient and special teaching space to meet students' learning needs, and establish a natural language description generation model for short videos based on attention mechanisms, which can accurately describe video content.

RESEARCH METHOD

Design Ideas of Short Video Course Teaching Mode

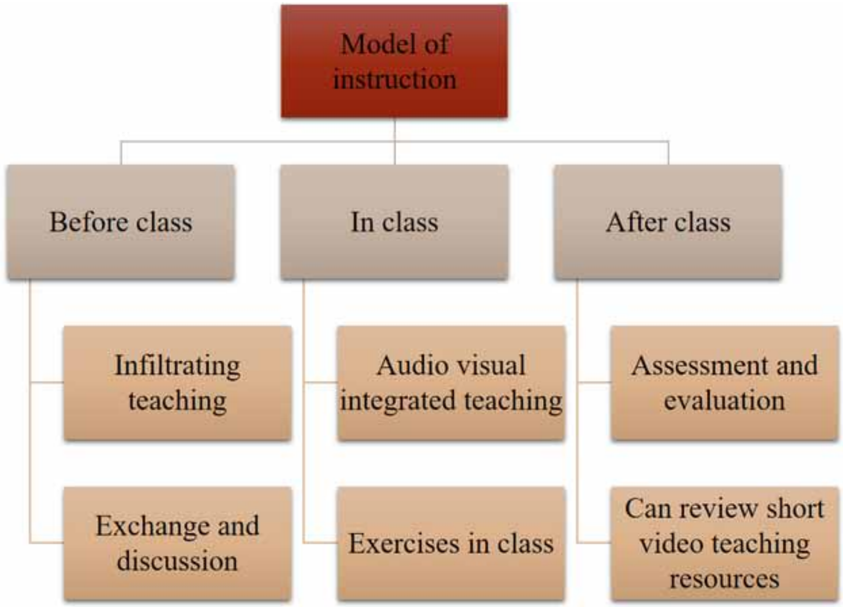
The application of short video description technology to students' English learning is to organically combine short video description technology with English teaching through students' English courses, thus changing the traditional teaching concepts, methods, and evaluation. To achieve this effective combination, we must make overall and systematic planning and arrangement—that is, we must carry out instructional design. Instructional design integrates basic elements in the teaching process, such as teaching objectives, content, objects, media, and evaluation, systematically analyzes and studies teaching needs, and designs a solution (Kabooha & Elyas, 2018). The general steps of the design process are, under the guidance of the viewpoint and method of system science, analyzing the characteristics of teaching content and teaching objects to determine the reasonable choice of teaching objectives and designing teaching media, giving full play to teachers' leading role and students' main role, forming an optimized teaching structure design and teaching evaluation, modifying the steps of the teaching process, and realizing the optimal classroom teaching plan (Hyunho et al., 2022).

A speech system mainly processes information about speech, while an image system mainly deals with nonverbal objects or pictures. According to dual coding theory's general model, the memory of graphics is better than that of words (graphic superiority effect), and the semantic generation of images in graphics is faster than that from the pronunciation of words. The new information that learners notice must go through three stages—sensory memory, short-term memory, and long-term memory—before it can be stored. Comparatively speaking, phonetic processing and spelling processing of new words are relatively shallow forms of processing, which will not lead to the retention of memory traces for a long time; however, the semantic processing of new words is a relatively deep-seated processing, which will lead to the retention of memory traces for a longer time.

Short videos can help students make effective use of fragmented time, learn through short videos anytime or anywhere, and give play to their learning initiative. They can also continue to learn in the fast-paced and demanding context of college life. Students can watch some short English learning videos by taking advantage of such trivial times as waiting in line in the canteen, taking a break between classes, and waiting for the bus. This not only allows them to kill time and relax, but also to effectively use time to absorb easily digestible knowledge points or consolidate learned knowledge. In addition, English teacher teams can register accounts on the short video platform, and students can follow the accounts of teaching teams. Even after class, students can still use the gap between waiting for the bus and lunch to learn and consolidate knowledge points. The combination of information technology and education makes the presentation methods of short videos flexible and diverse in teaching. Research shows that different presentation styles of short videos can achieve different learning effects. From the perspective of cognitive theory, micro-learning refers to learning activities with short time and fine content, emphasizing the fragmentation and loose connection of knowledge. Therefore, the characteristics of micro-learning, such as less content, short time, high efficiency, and more media, have attracted many learning enthusiasts. Students are the main focus of learning, and the guiding role of teachers should not be ignored. Teachers create a good learning environment for students, guide students to become knowledge builders through independent inquiry and cooperative learning, and gradually improve their learning ability (Wu et al., 2019). By designing colorful activities in the classroom, we can help students complete their inquiry tasks. Students can also change from passive acceptance of knowledge to active construction of knowledge and acquire new knowledge through autonomous learning and cooperative inquiry.

The short teaching video is generally 1-5 minutes in length, explaining a certain knowledge point or skill point. The length of the micro-class is generally 5-30 minutes. It gives a comprehensive and in-depth explanation of a topic or course and pays more attention to the interactive teaching method. The production of short videos is simpler than that of a micro-class. The initial starting point and ultimate goal of micro-curriculum design are to effectively promote students' English learning ability and help them learn efficiently. At the same time, teachers themselves should have the ability to find and solve problems and have their own research and ideas on curriculum development. Once the micro-curriculum experiment is carried out, it will definitely be a new opportunity for teachers' growth and development, and it will definitely bring great improvement to participating teachers. We designed the teaching model of a college English short video course, as shown in Figure 1.

Figure 1. Design of teaching model of short video course in college English



The core of the short video course teaching model lies in the cooperation and echo between a short video before class and classroom teaching, so the content design of the short video is the key to the success of micro-course experiment. To apply this model, a teacher would first record the explanation of, for example, superlative adjectives and adverbs into a short video, introduce the video effectively, arrange the short video for students to watch one day in advance. Secondly, the topic is introduced, and real-life scenes are created through short videos. Students can improve their oral expression ability through role-playing short videos. Then, the teacher would play a short video to emphasize the correct pronunciation and intonation to improve students' English pronunciation and check it at any time after class for consolidation.

Compared with the traditional oral English teaching, an important feature of short video oral English teaching is that it can create a real language environment for students' oral English learning. Meanwhile, college English teachers can share other teaching content, such as college English teaching ideas, teaching expressions, teaching achievements, teaching skills, and so on. Micro-courses are good at making use of modern network technology and short video description technology, so that college students can use computers or mobile phones anytime, anywhere and log on to relevant electronic devices to play the learning videos repeatedly. This not only helps college students avoid the restrictions of fixed classrooms and fixed class hours, but also helps college students to use spare time to learn English independently.

In the whole teaching process, teachers, as promoters, first analyze the learning objectives and learning content, select and make multimedia courseware in the media design stage; play the role of leading, nudging, guiding, and promoting in the teaching activity stage; and finally evaluate the teaching effect. Multimedia courseware plays such an important role in teaching, so whether the courseware is properly selected, whether it can arouse students' interest in learning, and whether it meets the teaching objectives directly affects the success of this course. The English teaching curriculum design mainly includes the cleaning and editing of the camera screen before playing. Students watch the homework design and key points in the process, as well as the selection of key points for explanation and discussion topic design after the broadcast. In consideration of students' knowledge limitations, sometimes we need to make some necessary introductions to the historical and cultural background of the video in advance to improve students' ability to interpret the works in depth.

Automatic Short Video Classification

With the extension of Internet use trends, short video English teaching will continue to expand in the future. In this context, how to quickly process short videos with different sources and categories and label and classify them for effective management is the key point in short video business. Compared with the single image understanding method, the video content understanding method has just started, and the related technologies are not mature enough. Because video contains a large amount of information and different themes in different time periods, the dynamic feature of video data sets video apart from static images (Waluyo & Apridayani, 2021). According to the shot organization and feature index, the relationship between shots is studied by video clustering and the shots with similar content are combined to gradually narrow the search range until the required video data is found. After the video is abstracted as key frames, video retrieval is converted into key frame retrieval according to some similarity. On this basis, the support vector machine (SVM) classifier is used to analyze the internal relations among features and the discrimination power of feature combinations, so as to select a variety of visual features. Under the condition of ensuring the classification results, the feature dimension is reduced, the computational complexity is reduced, and the classification performance of the classifier is improved.

Automatic video classification is the core requirement of this classification system, and it is necessary to automatically add classification labels to videos (Baltova, 1999). After the video is uploaded to the file storage system, some preprocessing work must be done before the algorithm is used for classification, including extracting the basic information of the video and generating features that can be directly used by the classification algorithm. After automatic classification of short videos, effective management should be carried out. The main users of video management function are annotators (Xu & Tan, 2021). According to the video content, it can be preliminarily divided into concrete video or abstract video for analysis. Concrete video refers to the video whose main content is concrete and non-abstract entities, such as animals, food, and so on. On the contrary, it is difficult to represent the entities in its content, so it can be called abstract video (Yan et al., 2020).

If the color distribution in a picture conforms to a certain probability distribution, the moment of color can be used as a feature to distinguish different color distributions. Compared with color histogram, another advantage of this method is that there is no need to vectorize features, thus speeding up the processing speed. The two lower moments of color are mathematically expressed as:

$$\sigma = \left(\frac{1}{M \times N} \sum_{i=1,j=1}^{i=M,j=N} p_{i,j} \left(p_{i,j} - \mu \right)^2 \right)^{1/2} \tag{1}$$

$$S = \left(\frac{1}{M \times N} \sum_{i=1,j=1}^{i=M,j=N} p_{i,j} \left(p_{i,j} - \mu \right)^3 \right)^{1/3} \tag{2}$$

where $p_{i,j}$ is the color component value of the pixel located at the coordinate $\left(i, j\right)$ in the image, and M, N is the length, width, and pixel number of the image, respectively.

The problem to be solved by classification is to classify an event or object. In use, the model can be used to analyze the existing data, and it can also be used to predict the future data. To construct a classifier, a training sample data set is needed as input. There are three kinds of classifiers: decision tree (DT) classifier, selection tree classifier, and evidence classifier.

Using DT to generate classifier has the following advantages: understandable rules can be generated, the amount of calculation is relatively small, it can handle continuous and discrete fields, and it can clearly show which fields are important. The information entropy after classification based on attribute A is:

$$I\left(A\right) = \sum_{i}^{v} \frac{p_i + n_i}{p + n} I\left(p_i, n_i\right) \tag{3}$$

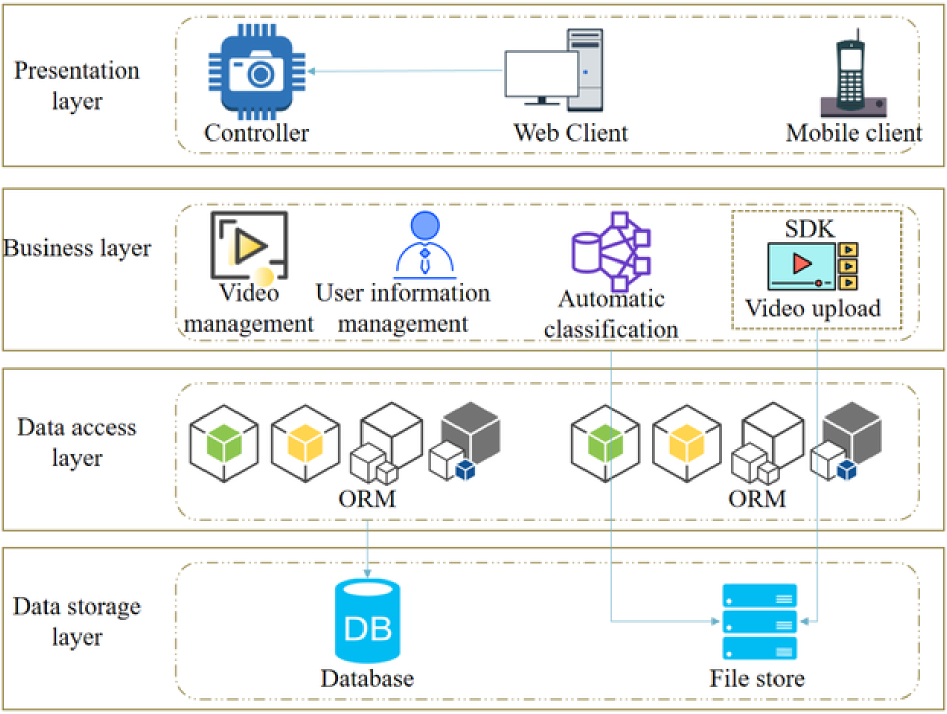
Therefore, the information gain based on A is:

$$Gain\left(A\right) = I\left(p, n\right) - E\left(A\right) \tag{4}$$

ID3 selects the attribute $A^{'}$ that makes $Gain\left(A\right)$ the largest (that is, $E\left(A\right)$ the smallest) as the root node. The above procedure is recursively called for the V subset E_i of E corresponding to different values of $A^{'}$ to generate the sub node $B_1, \cdots, B_v$ of $A^{'}$.

This topic designs the short video classification system according to the hierarchical architecture shown in Figure 2, which consists of data storage layer, data access layer, business layer, and presentation layer from bottom to top. The data layer includes database and file storage; the database is used to persist the business data of the classification system; and the file storage is used to store the video files uploaded by users, the picture files generated in the video preprocessing process, and

Figure 2. Short video classification system



the feature files. The presentation layer is the user interaction layer, including the controller of the server application, the web client, and the mobile client.

This module is to calculate the feature vectors of video files in the local system. There is no feature extraction on the distributed platform, and the optical flow and trajectory on the background are eliminated by estimating the camera motion estimation. Because the camera is moving, there are many trajectories on the background, and the trajectory of people is also greatly affected by the camera motion. In this paper, two eigenvectors are used to accomplish this function.

The gray images at time t and $t+1$ are I_t, I_{t+1}, respectively, and the projection transformation matrix H is obtained by calculating the two images:

$$I_{t+1} = H \times I_t \tag{5}$$

Then use the inverse of H to transform I_{t+1}, that is:

$$I_{t+1}^{warp} = H^{-1} \times I_{t+1} \tag{6}$$

where I_{t+1}^{warp} represents the image at the moment $t+1$ when there is no camera motion. I_{t+1}^{warp}, I_t can be used to calculate the optimal optical flow.

Distance measurement is to calculate the intra-class and inter-class distances of samples to analyze and study the strength of feature discrimination. For samples distributed in different domains, the smaller the distance between samples, the more difficult it is to distinguish, and the larger the distance between samples, the greater the separability. Euclidean distance can be regarded as Minkowski distance of second order, and its expression is as follows:

$$\delta\left(x_k^{(i)}, x_l^{(j)}\right) = \left(x_k^{(i)} - x_l^{(j)}\right)^T \left(x_k^{(i)} - x_l^{(j)}\right) \tag{7}$$

where $x_k^{(i)}, x_l^{(j)}$ represents the element of the feature vector.

The correctness of classification has become the most intuitive main index to examine feature discrimination (Thornton & Houser, 2005). The features that can minimize the errors of the classifier are obviously the features with the best discrimination. Therefore, the features of video shots can be analyzed directly from the classification results.

Realization of Short Video Automatic Description in English Teaching

English teaching videos carry rich visual information and massive data. How to express video effectively with natural language has always been a hot topic in the field of multimedia and vision. Deep learning (DL) has developed rapidly in recent years and is widely used in various fields. Automatically combining simple features into more complex features and using these combined features to solve problems is one of the core problems solved by DL. DL originated from the study of neural networks. Unlike neural networks, DL emphasizes the depth of model structure and feature learning. The depth of DL model is mostly multi-layer. Compared with image description, the amount of data processed by English teaching video description is far greater than image description. The image description model is simpler than the video description model, and the image description model does not need to deal with time structure. At the high level, a paragraph usually uses another loop layer to output the paragraph state, which is then used as the new initial state of the sentence generator.

In the task of natural language description generation of English teaching videos, video features are the first step and the key link, which plays a vital role in the output of the subsequent natural language model. The input of the traditional convective neural network (CNN) model is two-dimensional images, so in the task of processing related videos, we can only select appropriate video representative frames according to certain rules, use the CNN model to extract the image features of these representative frames, and then combine them to represent the video features. Each feature space gives full play to its basic kernel feature mapping ability. Because these mappings are different, heterogeneous data with different feature components can be processed by corresponding kernel functions. It can be seen that the high time and space complexity is an important bottleneck that hinders the development and application of multi-core learning algorithms.

English teaching situations cannot be simply constructed by relying on learners' imagination through pictures, verbal descriptions, and other means. We should carefully design scenes that can be immersed in cultural products, cultural customs, and cultural concepts. Through appropriate interactive means, they can actively participate in practical activities and stimulate their associative thinking with subjective consciousness. Only in this way can they truly understand and comprehend the value and significance of language use under the cultural background. It is necessary to build an immersive scene in ordinary classroom teaching through real equipment. Unlike static images, video is not fixed in the dimension of time, so it is difficult to express a longer video as a unified feature like the above methods. This task involves video image analysis, natural language processing, and other related technologies. The ultimate goal is to make the machine have the function of understanding video. Compared with using multiple sentences to describe the video content manually, this method only describes the characters, scenes, and behaviors in the video with a single sentence, which ideally encapsulates the most critical information.

Coding-decoding architecture is often used to solve some application problems of sequence-sequence. A sequence-sequence problem is simply to transform an input sequence x into another output sequence y. Sequence-sequence problems can be applied in machine translation, question answering system, image description, and other fields. Generally, the decoder uses CNN, because CNN can preserve the time dependence between sequences, which is beneficial to generate sequences with context.

According to the intermediate semantic representation vector C of sentence X and the history information $y_1, y_2, \cdots, y_{t-1}$ that has been generated before, the decoder generates the sequence element y_t to be generated at time t:

$$y_1 = g\left(C, y_0\right) \tag{8}$$

$$y_2 = g\left(C, y_1\right) \tag{9}$$

$$y_t = g\left(C, y_{t-1}\right) \tag{10}$$

In this paper, a "visual gate" is designed so that the model can achieve the balance between the visual part and the non-visual part during the generation of each word. In order to measure the contribution of visual information to the generation of current words, this paper introduces a balanced scalar $\beta_t \left[0,1\right]$ when generating words, and the calculation process is as follows:

$$\beta_t = \Phi\left(w_{bh}^T h_{t-1} + w_{bv}^T z_t\right) \tag{11}$$

Here, Φ represents the element-level calculation of the nonlinear function sigmoid, and $w_{bh} \in R^h, w_{bv} \in R^{d1}$ represents the weight parameters to be learned.

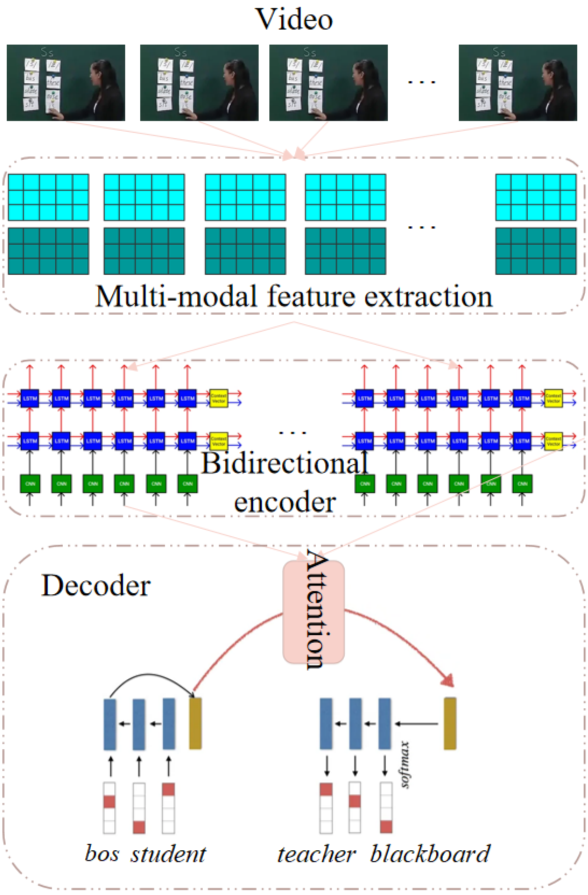
The potential problem in the task of video description generation is that the video feature sequence may be out of sync with the generated word sequence, that is, the order of objects and behaviors appearing in the video may be different from their positions in the description statement. In this paper, a new generation model of short video natural language description based on attention mechanism is designed. Figure 3 shows the whole model framework.

This model adds several bidirectional encoders. The feature fusion method is no longer set in the feature extraction module but postponed until after the feature coding. By inputting various modal features into an independent encoder for calculation, the underlying semantic information contained in various features can be more effectively learned, as can the correlation degree between each sequence and the output. The higher the correlation degree, the higher the attention weight, which makes the accuracy of the output description statement higher.

The hidden state vector h_t^d generated in each time step t in the decoding stage is related to $h_{t\text{-}1}^d$ generated in the previous step, word w_{t-1} output in the previous step and weighted sum c_t of hidden states output in the encoding stage, as shown in formula (12):

$$h_t^d = S\left(h_{t-1}^d, w_{t-1}, c_t\right) \tag{12}$$

Figure 3. Generation model of short video natural language description based on attention mechanism



The function of the Softmax layer is to convert the output of CNN into the output probability of each word in a word list. Actually, the number of objects to be classified is greater than 2, and the relationships between objects are mutually exclusive. The formula of the Softmax layer is shown in (13):

$$f\left(z_j\right)=\frac{\exp\left(z_j\right)}{\sum_{i=1}^{n}\exp\left(z_i\right)} \tag{13}$$

z_i represents a possibility of classification, and there are 1 to n, which are mutually exclusive.

CNN, which is used for coding purposes, can regard the output of the last unit as the compression of the input sequence $\left(x_1,\cdots,x_t\right)$, thus obtaining a vector representation v with a fixed dimension. Then, the intermediate representation v obtained by compressing the input sequence is used to initialize the initial state of the standard CNN language model, thus obtaining the conditional probability corresponding to the output sequence $\left(y_1,\cdots,y_t\right)$:

$$p\left(y_1,\cdots,y_t\left|x_1,\cdots,x_t\right.\right)=\prod_{t=1}^{T}p\left(y_t\left|v,y_1,\cdots,y_{t-1}\right.\right) \tag{14}$$

In the above formula, the probability distribution of each moment in the dictionary is obtained by the Softmax function. Words in video sequences and sentences come from different data domains and have different dependencies, so different models are needed to model them. This ensures the effectiveness of video information compression.

By calculating the similarity between semantic annotation and text description, we only add the most similar semantic annotation and text description to the parallel corpus. We measure their similarity S by counting the number of common words in text description and semantic representation:

$$S=\frac{\left|y_i\cap Lemma\left(z_i\right)\right|}{\left|y_i\right|} \tag{15}$$

where $Lemma\left(z_i\right)$ gets the stems of all the words in the whole sentence. It should be pointed out that the method we adopted here is only a rough approximation of semantic overlap because the text description may use completely different words to paraphrase the same meaning.

ANALYSIS AND DISCUSSION OF RESULTS

The online learning platform realizes data sharing with the unified information portal through data integration technology. In order to unify with the standard data format, users with more than 20 records were screened in the experiment. The data source was selected from 5500 randomly selected users and 992 English courses. In the selected data set, the progress of English video courses that users have learned is counted. According to the percentage of course learning progress, the corresponding grade is mapped to represent the user’s true feedback on the course. The experimental data selected in this paper are collected from the online education platform. The main learning behavior data are shown in Table 1. Each data set contains the information of the learner’s username (instead of name),

Table 1. Learner’s learning behavior data

Data type	Minimum	Maximum	Average value	Standard deviation
Landing times	107.65	3053.501	10.564	50.05
Number of registered courses	43.86	218.564	31.392	52.076
Number of courses passed	83.221	1111.437	31.567	81.428
Number of video views	98.034	710.026	13.171	3.492
Posting number	89.966	970.319	20.09	84.174
Number of replies	15.047	452.332	7.224	105.445
Number of test questions	56.405	2140.187	7.458	51.178
Reputation value	6.473	1982.306	30.375	32.346
Total grade of all courses	59.169	3083.17	24.932	86.131

registration date, last login date, login times, reputation, and so on. Analyzing the massive data of students' learning behavior accumulated by online education platforms and studying the online learning behavior of campus learners play an important role in understanding the online learning rules of campus students, improving the quality of instructors' short video construction and enhancing the plaform's teaching organization and management ability.

Due to the difference of each learner's learning behavior, there is not only a big numerical difference among learners in the same kind of behavior data, but also a bigger numerical difference among learning behaviors. In order to facilitate follow-up research and analysis, this paper normalized the above learning behavior data and linearly transformed the original data through min-max standardization, so that the results were mapped to [0,1]. During data analysis, different features often have different dimensions, which can lead to inaccurate weight estimation of different features during data processing and modeling. Therefore, in order to eliminate the impact of dimensions on data analysis, normalization processing is required. Normalization processing can eliminate dimensional differences between various features, making the importance of different features more equal, improving the comparability between features and the accuracy of the model. In addition, normalization processing can also improve the convergence speed and accuracy of the algorithm and improve the robustness and generalization ability of the model. Min-max standardization is one of the simplest and most commonly used methods in normalization processing, which maps data to the range of [0,1] through linear transformation. In addition, there are other normalization methods such as z-score standardization and log conversion. Different methods are applicable to different data types and data distribution situations. However, min-max standardization has the advantages of simple computation, easy understanding, and implementation, so it is widely used in many situations.

By studying the characteristics of learning behavior data, the weights of all learning behavior data in the final division results are set, and according to the weight relationship, the learning behavior data of learners are converted into the scores of learners in short video learning. Finally, the types of learners are classified by clustering the scores. Through the comparative analysis of data visualization, two clustering methods are used to divide the distribution of people in the two-dimensional coordinate system of "reputation value-course achievement," as shown in Figure 4 and Figure 5.

Figure 4. conventional clustering division method

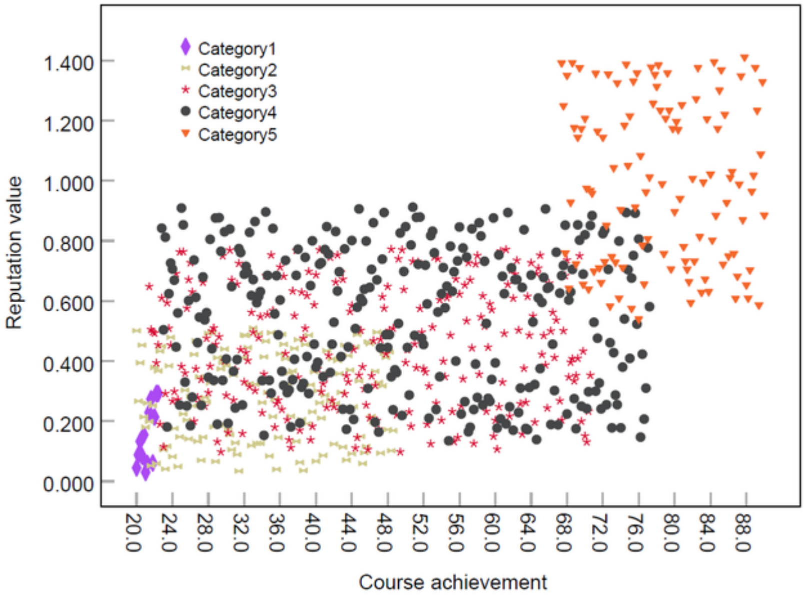
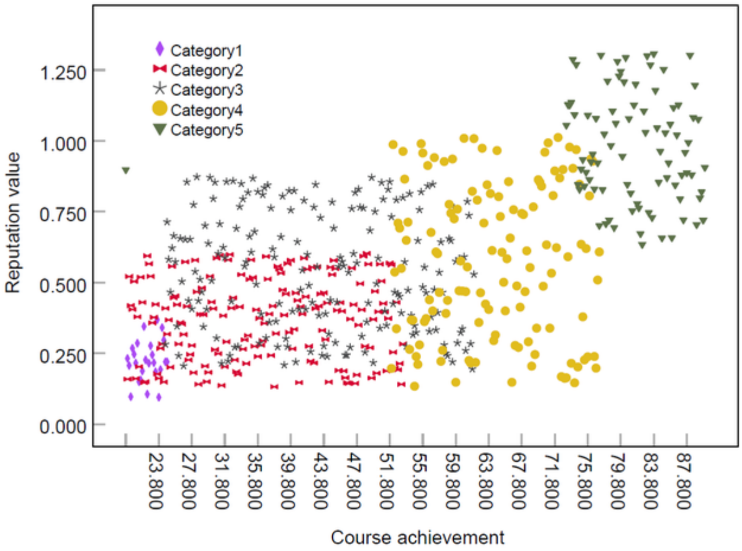


Figure 5. Integral clustering partition method



On the whole, the division of learners' groups is good. From another angle, it is proved that the classification method of learners' types using the integral clustering method has stronger individual differentiation than the conventional clustering method. Learners in each class have similar learning behaviors, while learners' learning behaviors between classes are obviously different, which is in line with the evaluation criteria for testing the validity of clustering results—the closer the cluster is, the better the separation between clusters.

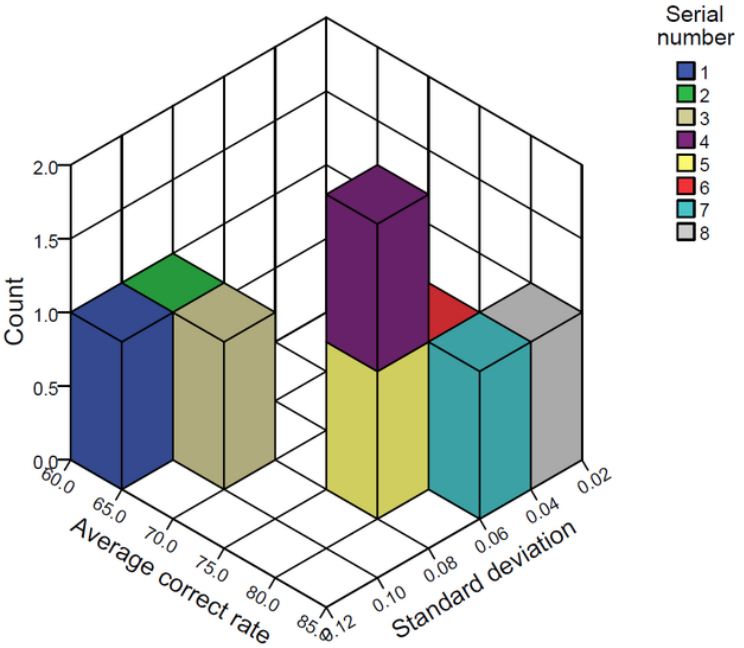
We choose support vector machines (SVM) as the classifier of shot database. We input the extracted visual features into SVM network for training and testing. According to Table 2 and Figure 6, in terms of the classification effect of these eight single features for English teaching categories, the boundary direction histogram has the best effect, followed by RGB improved color moment and RGB color histogram.

The effect of texture in the classification experiment of English teaching video shots is not ideal because English teaching videos do not have their own distinctive features in texture features.

Table 2. Single visual feature and average accuracy and standard deviation of english teaching categories

Serial number	Characteristic	Average correct rate	Standard deviation
1	Dynamic texture	60.834	0.106
2	Tamura texture	64.168	0.085
3	Gabor texture	67.23	0.08
4	Edge dynamics	77.307	0.071
5	Color dynamics	77.407	0.067
6	RGB color histogram	79.41	0.052
7	Improved RGB color moment	80.989	0.04
8	Edge histogram	83.243	0.026

Figure 6. Trend of average accuracy and standard deviation



However, the order of RGB improved color moment is much higher than that of color histogram and edge direction histogram, and its performance is not superior. Therefore, edge direction histogram and RGB color histogram are the best choices for the classification of English teaching video shots. At the same time, according to the results of standard deviation, we can see that the stability of these two visual features is also good. The visual features with strong discrimination selected by mutual information have also achieved good classification results in the experiment of SVM classifier.

Each described sentence is set as a description sentence that is finally output by the system. It is a natural language. The evaluation of the system can be converted into the evaluation of natural language. With reference to the indicators used to evaluate the quality of machine translation results in the field of natural language, these indicators can reflect the accuracy of translation results through a large number of experiments. This paper uses bilingual evaluation understudy (BLEU) and consensus-based image description evaluation (CIDEr) to evaluate the performance of the video description system.

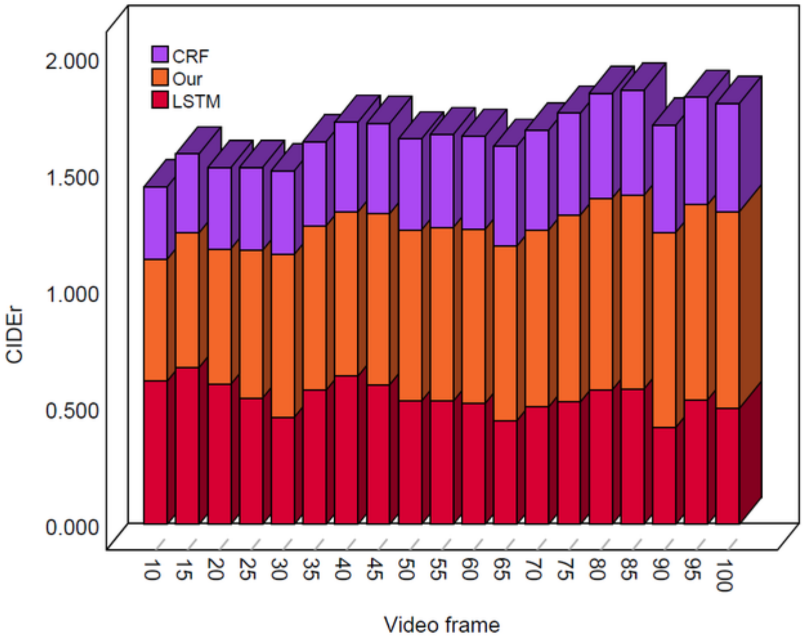
In order to verify the effectiveness of CNN's video feature compression, on the premise of fixing the model structure and super parameters, we changed the number of input video frames and designed the experiment of inputting 10 frames and 100 frames, respectively. The specific experimental results are shown in Table 3 and Figure 7.

Unlike Long Short-Term Memory networks(LSTM) and other models, CRF can take long-term context information into consideration. It more considers the linear weighted combination of local features of the whole sentence. From the perspective of data size, when the data size is small, the experimental effect of CRF is slightly better than that of BILSTM. When the data scale is large, the effect of Bidirectional Long Short Term Memory Network(BILSTM) should exceed that of CRF. If the task to be identified does not rely too much on long-term information, then recurrent neural networks(RNN) and other models will only add additional complexity. With the increase of the number of video frames, the performance indicators indicated by other parameters, except LSTM, are monotonically increasing with the increase of the number of video frames, and the increase of this method is the most obvious. With the increase of the number of video frames, this method increases

Table 3. Influence of video sequence length on model performance

Video frame	CRF	Our	LSTM
10	0.311	0.523	0.614
15	0.339	0.578	0.672
20	0.349	0.581	0.6
25	0.356	0.634	0.541
30	0.359	0.699	0.459
35	0.362	0.705	0.575
40	0.385	0.705	0.635
45	0.387	0.734	0.599
50	0.391	0.734	0.528
55	0.399	0.745	0.528
60	0.4	0.747	0.518
65	0.429	0.749	0.444
70	0.429	0.759	0.503
75	0.437	0.802	0.525
80	0.45	0.823	0.576
85	0.451	0.831	0.58
90	0.46	0.837	0.415
95	0.461	0.837	0.534
100	0.465	0.844	0.497

Figure 7. Trend diagram of model performance



by 9.22% and 6.37%, respectively, which shows that the coding neural network has played a role in compressing video information.

Figure 8 shows the index values of this experimental model and the above two advanced models. As can be seen from the table, our method has better performance than other methods.

Both CRF and LSTM describe video data based on the relationship between different components of video data input. The size of the target object in the video data set is relatively small, so CNN's results on these three indicators are not as good as other models. Our model integrates multi-layer networks and adds time attention mechanism and space attention mechanism.

In Figure 9, BLEU is used to evaluate the performance of the model. When epoch is zero, the model has already pre-trained the model parameters by supervised learning. As can be seen from the figure, after many times of epoch training, the accuracy of the model is gradually improved. This shows that learning positioning strategy is beneficial to video description.

In the process of model training, reinforcement learning algorithm plays a role in tuning in order to prove the role of reinforcement learning in improving video description effect. Then the strategy gradient algorithm is used to learn the region selection strategy step by step. It can be seen from the results that the regional positioning network and time attention mechanism of this model play an important role in improving the performance of the model. In addition, it is beneficial to the experimental results to fuse more modal features. Therefore, if all modal features are fused and input into the language generation model as a whole, the most accurate description statement will be obtained.

Figure 8. Comparison of training results of different model data

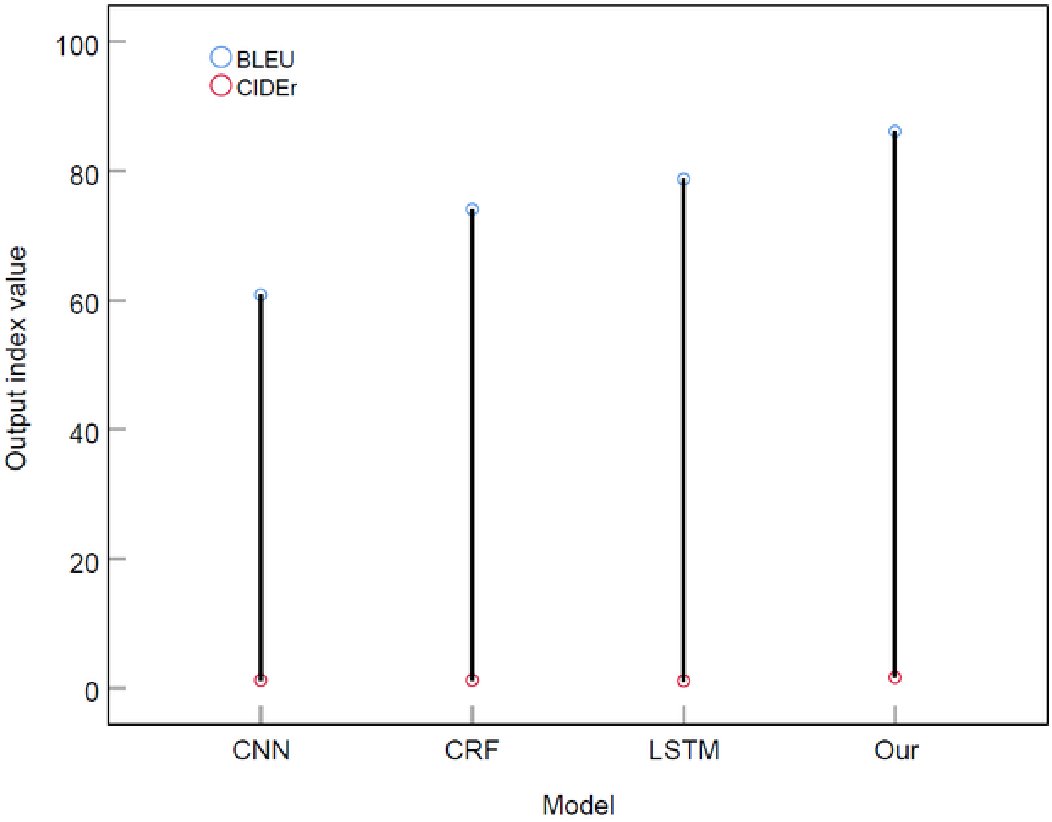
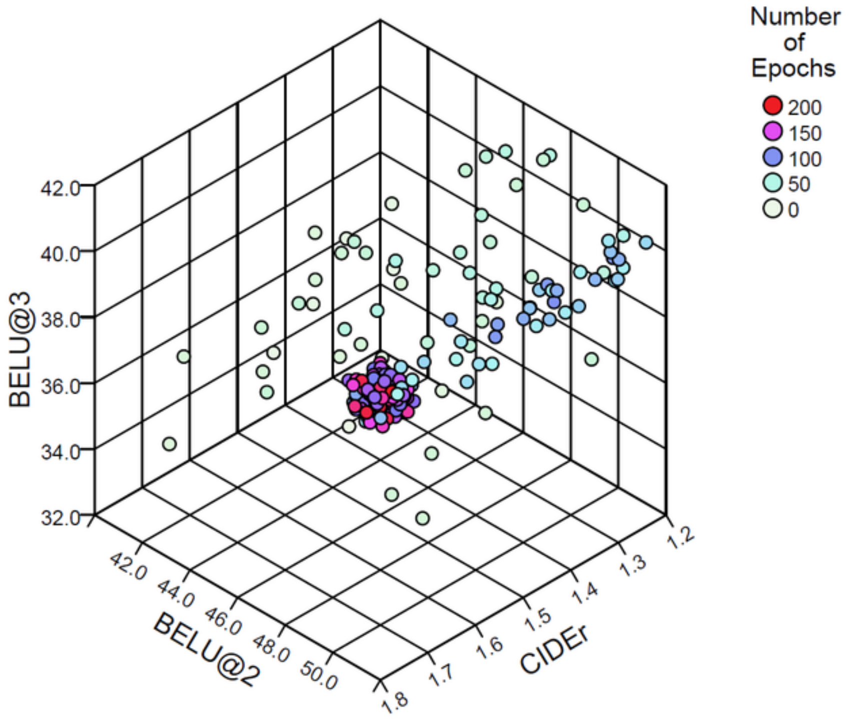


Figure 9. Reinforcement learning effect diagram



Short video clips based on attention mechanism provide a real communication scene for English learners with their rich and vivid language and the authenticity of language presentation. They create an optimal environment for learners to acquire language. However, as far as the present situation is concerned, there are still some aspects that require attention in college English teaching.

It is necessary to consider the preferences of students, which does not mean to cater to students on purpose. It means to choose as few literary films as possible. This kind of movie language is more difficult and may also have the local voice of the 18th or 19th century. The plot of the story is relatively complicated, and the content is relatively not close to the students' actual life. If students are not interested in such films, they will not achieve the purpose and effect of teaching. For teachers, English short videos can diversify teaching content. Therefore, in actual teaching, many teachers make excessive use of films to teach, so that the whole English class eventually became a film appreciation class and the teacher becomes a film projectionist. This not only ignores the main position of students in teaching, but also deviates from the original teaching objectives of the curriculum.

The challenges that were addressed in this study include the negative learning mood among students in English classroom teaching, the need to create an efficient and special teaching space that meets students' learning needs. Additionally, the study addresses the challenge of effectively applying short video description technology in college English teaching. Regarding the contributions of this study, the results showed that the proposed generation model of short video natural language description based on attention mechanism significantly improved the performance indicators. The fusion of more modal features led to better experimental results, indicating the effectiveness of this method. Furthermore, this study's findings contribute to new knowledge in the field of applying short video description technology in college English teaching. Specific recommendations can be provided for educational practice, such as implementing micro-class teaching methods and using video features to improve learning outcomes. In the analysis of clustering results, we obtained classification results of learning types by comparing different clustering methods, which has certain reference significance in the international community. At the same time, by training the natural language description generation model based on attention mechanism, we propose a new short video description generation method, which has certain promotion and application value in the field of video analysis and English education.

CONCLUSION

In contrast to traditional teaching methods, micro-courses are primarily defined as video recordings of teachers highlighting important or challenging material from classroom instruction in accordance with certain requirements for practical advice or teaching standards. Since a video consists of a collection of shots that are typically trimmed during reprocessing, investigating the shot structure of a video is crucial to understanding video content analysis. In this study, short video description technology is applied to college English teaching, and a generation model of short video natural language description based on attention mechanism is established. Learners in each class have similar learning behaviors, while learners' learning behaviors between classes are obviously different, which is in line with the evaluation criteria to test the validity of clustering results. The performance indicators shown by the other parameters, with the exception of LSTM, rise monotonically as the number of video frames increases. This method's increase is the most noticeable. This approach increases by 9.22% and 6.37%, respectively, with an increase in the number of video frames, indicating that the coding neural network has contributed to the compression of video data. This study provides an effective teaching method based on short video description technology for college English education and has certain reference significance for future English education. Limitations of this study include the use of a specific model for short video natural language description based on attention mechanism, which may not be suitable for all situations. Future research could explore the development of alternative models and investigate the application of short video description technology in other areas of education.

DATA AVAILABILITY

The figures and tables used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This paper was supported by Research Project on the Reform of Vocational Education and Adult Education in Jilin Province, Research on Adult Education Reform in the Context of Merging "Internet +" and Lifelong Learning Concept (No. PX-10222183).

ACKNOWLEDGMENT

The authors would like to show sincere thanks to those techniques who have contributed to this research.

REFERENCES

Aguirre, N., Grall-Mas, E., Cymberknop, L. J., & Armentano, R. L. (2021). Blood pressure morphology assessment from photoplethysmogram and demographic information using deep learning with attention mechanism. *Sensors (Basel)*, *21*(6), 2167. doi:10.3390/s21062167 PMID:33808925

Baltova, I. (1999). Multisensory language teaching in a multidimensional curriculum: The use of authentic bimodal video in core French. *Canadian Modern Language Review*, *56*(1), 31–48. doi:10.3138/cmlr.56.1.31

Bera, A., Wharton, Z., Liu, Y., Bessis, N., & Behera, A. (2021). Attend and guide (AG-net): A key points-driven attention-based deep network for image recognition. *IEEE Transactions on Image Processing*, *30*, 3691–3704. doi:10.1109/TIP.2021.3064256 PMID:33705316

Chen, L., Weng, T., Xing, J., Pan, Z., Yuan, Z., Xing, X., & Zhang, P. (2020). A new deep learning network for automatic bridge detection from SAR images based on balanced and attention mechanism. *Remote Sensing (Basel)*, *12*(3), 441. doi:10.3390/rs12030441

Chen, S., Wu, S., Wang, L., & Yu, Z. (2021). Self-attention and adversary learning deep hashing network for cross-modal retrieval. *Computers & Electrical Engineering*, *93*, 107262. Advance online publication. doi:10.1016/j.compeleceng.2021.107262

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., & Deng, L. (2017). Semantic compositional networks for visual captioning. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1141–1150). IEEE. doi:10.1109/CVPR.2017.127

Hang, L. (2018). Deep learning for natural language processing: Advantages and challenges. *National Science Review*, *5*(1), 24–26. doi:10.1093/nsr/nwx110 PMID:34691827

Hari, R., Caprez, R., Dolmans, D., Huwendiek, S., Robbiani, S., & Stalmeijer, R. E. (2024). Describing ultrasound skills teaching by near-peer and faculty tutors using cognitive apprenticeship. *Teaching and Learning in Medicine*, *36*(1), 33–42. doi:10.1080/10401334.2022.2140430 PMID:36322510

Huang, Z., Epps, J., Joachim, D., & Sethu, V. (2020). Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 435–448. doi:10.1109/JSTSP.2019.2949419

Hyunho, K., Minsu, P., Ingoo, L., & Hojung, N. (2022). BayeshERG: A robust, reliable and interpretable deep learning model for predicting hERG channel blockers. *Briefings in Bioinformatics*, *23*(4), bbac211. doi:10.1093/bib/bbac211 PMID:35709752

Jararweh, Y., Al-Ayyoub, M., & Benkhelifa, E. (2019). Advanced Arabic natural language processing (ANLP) and its applications: Introduction to the special issue. *Information Processing & Management*, *56*(2), 259–261. doi:10.1016/j.ipm.2018.09.003

Kabooha, R., & Elyas, T. (2018). The effects of YouTube in multimedia instruction for vocabulary learning: Perceptions of EFL students and teachers. *English Language Teaching*, *11*(2), 72–81. doi:10.5539/elt.v11n2p72

Li, L., Jia, W., Zheng, J., & Jian, W. (2018). Biomedical event extraction based on GRU integrating attention mechanism. *BMC Bioinformatics*, *19*(S9, Suppl. 9), 177–184. doi:10.1186/s12859-018-2275-2 PMID:30367569

Li, M. (2021). An immersive context teaching method for college English based on artificial intelligence and machine learning in virtual reality technology. *Mobile Information Systems*, *2021*, 2637439. doi:10.1155/2021/2637439

Li, X. (2019). Characteristics and rules of college English education based on cognitive process simulation. *Cognitive Systems Research*, *57*, 11–19. doi:10.1016/j.cogsys.2018.09.014

Li, X., Zhang, W., & Ding, Q. (2019). Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Processing*, *161*, 136–154. doi:10.1016/j.sigpro.2019.03.019

Li, Y., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2018). Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(7), 11336–11343. doi:10.1609/aaai.v34i07.6795

Luo, J. W., & Chong, J. J. R. (2020). Review of natural language processing in radiology. *Neuroimaging Clinics of North America*, *30*(4), 447–458. doi:10.1016/j.nic.2020.08.001 PMID:33038995

Taskin, Z., & Al, U. (2019). Natural language processing applications in library and information science. *Online Information Review*, *43*(4), 676–690. doi:10.1108/OIR-07-2018-0217

Thornton, P., & Houser, C. (2005). Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning*, *21*(3), 217–228. doi:10.1111/j.1365-2729.2005.00129.x

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*(1), 1–37. doi:10.1017/S0261444809990267

Waluyo, B., & Apridayani, A. (2021). Teachers' beliefs and classroom practices on the use of video in English language teaching. *Studies in English Language and Education*, *8*(2), 726–744. doi:10.24815/siele.v8i2.19214

Wu, L., Rao, Y., Yu, H., Wang, Y., & Ambreen, N. (2019). A multi-semantics classification method based on deep learning for incredible messages on social media. *Chinese Journal of Electronics*, *28*(4), 754–763. doi:10.1049/cje.2019.05.002

Xu, F., & Tan, S. (2021). Deep learning with multiple scale attention and direction regularization for asset price prediction. *Expert Systems with Applications*, *186*, 115796. doi:10.1016/j.eswa.2021.115796

Yan, S., Xie, Y., Wu, F., Smith, J. S., Lu, W., & Zhang, B. (2020). Image captioning via hierarchical attention mechanism and policy gradient optimization. *Signal Processing*, *167*, 107329. doi:10.1016/j.sigpro.2019.107329

Zeng, W., Qiu, Y., Huang, Y., Sun, Q., & Luo, Z. (2022). Multivariety and multimanufacturer drug identification based on near-infrared spectroscopy and recurrent neural network. *Journal of Innovative Optical Health Sciences*, *15*(4), 2250022. doi:10.1142/S1793545822500225

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. [TIST]. *ACM Transactions on Intelligent Systems and Technology*, *11*(3), 1–41. doi:10.1145/3374217

Xiaoyan Shi was born in JiLin, China, in 1982. From 2000 to 2004, she studied in Chang Chun College of Engineering and received her bachelor's degree in 2004. From 2009 to 2012, she studied in JiLin University and received her Master's degree. Currently, she works in JiLin Provincial Economic Management Cadre College. She has published several papers, Her research interests include English Teaching, Career Planning and Food Culture.