Application of a Short Video Caption Generation Algorithm in International Chinese Education and Teaching

Qianhui Dai, Zhengzhou Railway Vocational and Technical College, China*

ABSTRACT

With the continuous development of speech recognition technology, automatic subtitle generation has gradually attracted people's attention. However, the level of short videos is uneven, and cultural teaching is also one-sided, irregular, and lacking systematicness. Since the media era, it is possible to apply the short video subtitle generation algorithm to international Chinese education and teaching. However, Chinese teachers should pay attention to some possible problems in self-media videos and adopt appropriate teaching strategies. This paper mainly discusses the development of international Chinese education and teaching under the new media environment, and discusses the characteristics, advantages, and disadvantages of international Chinese education and teaching under the new media environment, as well as the existing problems. Short video subtitle generation algorithm provides a new way for international Chinese education and teaching, enhances the vitality of education, and expands educational channels.

KEYWORDS

Attention Mechanism, Caption Generation Algorithm, Deep Learning, International Chinese Education, Short Video Content Understanding, Teaching Strategies

INTRODUCTION

The short video industry has developed rapidly in China and has become increasingly mature. With the development of modern We Media and the emergence of We Media short videos, it has also had a certain impact on international Chinese language education and teaching. Video entertainment is becoming more and more popular in China today. Its language expression and cultural factors are very beneficial for foreign students to learn Chinese (Tian & Hu, 2021). Movies, documentaries and other film and television works have appeared more and more in Teaching Chinese as a Foreign Language classes. For foreign students, film and television works give consideration to language training in listening, reading (subtitles), speaking, and other aspects. Especially in the teaching of Chinese culture as a foreign language, film and television works can make them feel the abstract Chinese culture...
more intuitively and the cultural habits hidden behind the social image (Chen et al., 2020). There is no doubt that audiovisual media has great potential as an auxiliary means of teaching, but there are still many problems in its application in the cultural classroom, such as the inappropriate selection of film and television works, the long-playing time, and the lack of timeliness of content.

Multimedia-assisted teaching of Chinese as a foreign language is the general trend. Nowadays, the teaching of Chinese as a foreign language shows a trend of diversification. More and more teachers of Chinese as a foreign language are trying new teaching methods to help foreign students learn more language and cultural knowledge in a relaxed and pleasant teaching environment (Qian et al., 2021). In this situation, online short video has gradually entered the public’s view with its unique interest, and it is also expected to enter the classroom of teaching Chinese as a foreign language (Cao et al., 2019). Based on the above two points, in this paper the author preliminarily discusses how to use the entertainment form of short online video as an auxiliary means of cultural teaching, and tries to integrate this emerging thing into the cultural teaching of Chinese as a foreign language, bringing vitality and vitality to the teaching of Chinese as a foreign language. What is the relationship between language teaching and culture teaching? How should the content of culture teaching be defined? These are the theoretical basis for the short video subtitle generation algorithm to integrate into international Chinese language education and cultural teaching. Traditional subtitles need manual editing, which is inefficient. For many languages, subtitle editing needs to master many languages, which is unrealistic for most people. In addition, after manually editing subtitles, it is also necessary to align the voice and subtitles in time, which requires the editors to determine the start time and end time of the audio corresponding to the subtitles (Singh et al., 2022). In addition, the short video service may require various subtitle effects, and it is difficult to add these animations for manual editing. In addition, most video editing software on the market can only add animation after opening members, which is costly.

Video caption generation technology mainly includes two important challenges, namely video understanding and language generation. First of all, video understanding not only includes the description of the scene in the current video, but also needs to explore its deep logic. Secondly, on the premise that the generated content is correct, it is also necessary to consider the accuracy of the language sequence before and after. How to take these two aspects into account is the key to the current video caption generation method. The current video caption generation methods can be summarized into the following two types: Time-series based method and multimodal method (Li et al., 2019). The problem of video caption generation the author investigated in this research is difficult to study, mainly because it is a cross modal task, which requires sufficient understanding of the existence of different modes, and how to transform each mode. In the process of mode conversion, the inevitable problem is that there is strong heterogeneity between the two modes, which makes the generated statements not real enough. Secondly, it is also necessary to analyze the objects in the video, as well as the relationship between each object, and extract key information from them to generate statement annotation.

In this paper, the author proposes a new video subtitle generation and update model based on video timing attention fusion mechanism. This model combines the abstract features of low-level video content related semantics and high-level space embedded features, which can more effectively generate words and form the final subtitle sentence. Then, the author carried out a series of comparative experiments on the first large-scale Chinese continuous sign language translation and recognition video dataset, and the self-built video audit dataset, AuditV. The research results of this paper can be applied to many scenes, intelligentize people’s daily life, and help people greatly save various human costs and time costs.
APPLICATION ANALYSIS AND SHORTCOMINGS OF SHORT VIDEO CAPTION GENERATION ALGORITHM IN INTERNATIONAL CHINESE EDUCATION AND TEACHING

There are two types of subtitle production: One is manual production, where the producer listens to the voice in the video, writes the words he hears word by word, and then assigns a corresponding time period to a certain sentence to produce text in subtitle format; the other is automatic generation. Producers can generate subtitle text in the required format by inputting the voice in the video through specific software. For the former, the workload is too large and the time consumed is too long. In contrast, the latter has become a trend and has developed in related fields. Video caption generation aims to enable the computer to independently describe or summarize the information in the video (Hong et al., 2018). It is necessary to enable the model to automatically understand and analyze the content in the video. For most people, it is relatively simple and easy to describe the content of a video with a sentence or summarize the category of a video with a word tag. However, for computers, it is a great challenge to understand and analyze the video content and generate descriptive caption sentences or general caption words. Because each video is composed of many video frames and pictures, each video frame contains complex and rich information, such as character actions and surrounding environment. At the same time, for a video, in addition to considering each frame of the spatial dimension, the information of the time dimension is also very important. The front and back frames of the video are related, for example, the front and back action frames of the deaf sign language video contain the linguistic relationship; there is also temporal information before and after the video of actions or related scenes. In addition, the general words and descriptive sentences of subtitles generated by computer automatic recognition of video content should also be accurate and conform to human oral habits, that is, conform to the grammar in natural language.

Video caption generation is a very difficult task, which needs to combine the knowledge of vision and language. The former is perceptual intelligence, that is, extracting high-dimensional abstract features of video through understanding the pixel level of video picture frames; the latter is cognitive intelligence (Haijiao et al., 2019). In addition to the video visual features obtained from the former, it also needs to combine existing corpus knowledge to describe or summarize the video cognitively.

At present, attention mechanism has been widely used in machine translation, image classification, and other fields, and it has a good effect in the task of single mode. The essential reason is that the internal composition of a single modal space is relatively similar, which is conducive to calculation and fitting. However, if attention is to be derived from multimodal features, it is difficult for traditional attention to fit an ideal effect because the spatial composition between multimodes is different. Therefore, when conducting multimodal attention analysis, it is necessary to first transform the internal space of each mode, and then match and fit through collaborative analysis. Figure 1 shows the pattern fusion algorithm based on the attention mechanism.

In this study, the author implemented the attention based on the soft attention mechanism. The goal of the soft attention mechanism is to generate a spatial hotspot map that highlights the image regions associated with each generated word. An example may be a processed feature matrix $F$, which may be a specific feature or a combination of multimodal features. The soft attention mechanism guides the generation of attention weights for each element in $F$ through the hidden state of the decoder LSTM(Long Short-Term Memory), and the basic attention score for each element is calculated by Equation 1:

$$e^i_q = \phi(W^T_q F^i + U_q h_{i-1} + b_q)$$ (1)
where $w$, $W_q$, and $b_q$ are the parameters to be learned and $F^i$ is the $i$-th element in $F$. Then, the basic score of each element is transferred to a sequence Softmax layer to obtain the corresponding attention weight, as Equation 2 shows:

$$
\alpha^i_t = \frac{\exp \left( e^i_t \right)}{\sum_{j=1}^{M} \exp \left( e^j_t \right)}
$$

(2)

where $M$ represents the number of elements in $F$. The final context weighted by attention is, as Equation 3 shows:

$$
c_t = \sum_{j=1}^{M} \alpha^i_t F^j
$$

(3)

Assuming that the video contains $n$ modal features, first apply a neural network feedback layer to $V$ to reduce the dimension, as Equation 4 shows:

$$
R^i = \phi \left( W^i_p v^i + b^i_p \right), i = 1, 2, \cdots, n
$$

(4)

where $W^i_p$ and $b^i_p$ respectively represent the parameters to be learned for the $i$-th mode, $\phi$ represents the hyperbolic tangent activation function tanh, and $R = \left[ R^1, R^2, \cdots, R^n \right]$ represents the multimodal features of the same dimension after dimension reduction. Then, the attention weight of multimodal features $R$ is calculated by soft attention mechanism. In this paper, the intermodal attention mechanism at time $t$ is defined as Equation 5:
Sphinx system is a speech recognition system developed by CMU in 1987, and there are currently four versions. Sphinx4 is a modular and pluggable framework. Sphinx4 consists of four modules, including front-end processing, decoder, knowledge base (including acoustic model, language model, and dictionary), and applications. Front-end processing is responsible for collecting, labeling, and processing input data. The knowledge base provides the data needed by the decoder, including acoustic models, dictionaries, and language models. The application layer mainly aims at some operations for practical applications of speech recognition. The decoder mainly reads speech feature data and combines it with data in the knowledge base to search for cases and find the most likely word sequence. After the front-end processing of the speech signal, the feature values are extracted, and then the decoder uses the model in the knowledge base for pattern matching. During the entire implementation process, MFCC (Mel frequency cepstrum coefficients) feature extraction method, N-gram language model training technology, HMM (Hidden Markov Model) acoustic model training technology, and HMM related forward backward algorithm and Viterbi algorithm are used. Figure 2 shows the overall framework of sphinx speech recognition technology.

Equation 6 shows the common windowing formula for short video processing:

\[
W(n, a) = (1 - a) - a \cos \left( \frac{2 \pi n}{N - 1} \right), 0 \leq n \leq N - 1 \quad S'(n) = S(n) * W(n)
\]  

The logarithmic energy output of each filter bank, that is, subband energy, has to be calculated. Equation 7 shows the calculation formula:

\[
s(m) = \ln \left( \sum_{k=0}^{N-1} \left| X_a(k) \right|^2 H_m(k) \right), 0 \leq m \leq M
\]  

**Figure 2. Basic framework of sphinx speech recognition**
where $s(m)$ is subband energy, $X_s(k)$ is the result of discrete Fourier transform calculation, $M$ is the number of Mel filters, $N$ is the energy of audio frames, and $H_m(k)$ is the frequency response calculation method of triangular filters.

MFCC coefficients are calculated by DCT (Discrete Cosine Transform) transformation. Equation 8 shows the transformation formula:

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos \left( \frac{\pi n(m - 0.5)}{M} \right), \quad n = 1, 2, \ldots, L$$

where $s(m)$ is the logarithmic energy calculated by Equation 2, $L$ is the order of MFCC coefficient, and is the number of triangular filters.

An attention layer has to be applied within the mode to the temporal features and audio features to explore the temporal characteristics of the mode, as Equation 9 shows:

$$c^\text{tem}_t = \text{Intrattention}(h_{t-1}, w^\text{tem}), \quad c^\text{aud}_t = \text{Intrattention}(h_{t-1}, w^\text{aud})$$

where $c^\text{tem}_t$ and $c^\text{aud}_t$ respectively represent the time context and audio context at time $t$, and $h_{t-1}$ represents the hidden state of decoder at time $t-1$. Next, action features $v^\text{mot}_t$ and audio context $c^\text{tem}_t$ have to be input into an inter modal attention layer, as Equation 10 shows:

$$c^\text{mt}_t = \text{Intrattention}(h_{t-1}, \{v^\text{mot}_t, c^\text{tem}_t\})$$

where $c^\text{mt}_t$ represents the visual context at time $t$.

The visual context sequence is input into the coded LSTM after passing through a multimodal layer, as Equation 11 shows:

$$c^\text{mot}_t = \phi \left( w^T \left[ W^\text{ma} \cdot c^\text{ma}_t, W^\text{aud} \cdot c^\text{aud}_t \right] \right)$$

where $w^T$, $W^\text{ma}$ and $W^\text{aud}$ are parameters that need to be learned. $\phi$ is the Sigmoid bilinear activation functions.

The low-level attention layer plans the relationship between different low-level modes through hierarchical attention. Among them, the attention layer used includes intramodal attention and intermodal attention. Through the cross use of the two kinds of attention, the fusion of action mode, time mode and audio mode are solved.

**Stimulate Interest in Learning Chinese Culture**

At present, many We Media videos are related to Chinese culture. Many take the Chinese natural scenery as the video background, and integrate some human parts. They express and transmit Chinese culture by establishing an idyllic and idyllic lifestyle (Xu et al., 2019). This kind of video will attract viewers to some extent and make them resonate. For international Chinese language education learners, this is also a way to indirectly understand Chinese culture and stimulate learning interest. Short videos usually last about 10 minutes, while longer videos may last up to 20 minutes. Different from language input, in this short time of more than 10 minutes, the content presented in the video is
diverse, and any shot switching is transmitting new information. The natural scenery, Chinese food, special clothes, and similar features in the video show the distinctive side of Chinese culture. This rich and colorful display mode stimulates the viewer from visual and auditory aspects, making the viewer have the idea of continuing to watch. Besides, this way of video telling can also reduce the difficulty of understanding to a certain extent, so as to stimulate the interest of learners. Similarly, for the actual classroom, it is difficult to continue to attract students’ attention if teachers only give a single explanation (Gao et al., 2017). The reasonable use of We Media Video can create an active and relaxed classroom atmosphere, so as to better attract students and improve their interest in learning.

**Deepen Understanding of Chinese Cultural Elements**

Chinese cultural elements are symbols of Chinese culture. Learning to understand these elements can help learners feel Chinese culture better. For online video, people also gradually like works that can reflect Chinese traditional elements and have Chinese cultural heritage. One after another batch of online We Media began to produce short videos with Chinese cultural elements, and videos with themes such as Han clothing, Chinese food, drama cross dressing, and domestic tourism emerged endlessly. These videos demonstrate traditional Chinese cultural elements in a dynamic way, and unconsciously deepen the viewers’ understanding of this (Ma, 2018). The teaching of international Chinese education and culture is not achieved overnight, and learners need to deepen their understanding of the target language culture through imperceptible influence. Sometimes, a single picture display or documentary video playback can show Chinese culture to a certain extent, but it is often difficult to combine with current popular culture. On the contrary, short videos that combine formal content with modernity can achieve this goal well. By playing and watching videos with Chinese cultural elements, learners can subconsciously understand the connotation of elements and taste Chinese culture.

**Cultivate Awareness of Cross-Cultural Communication**

Intercultural communication refers to the communication between people from different countries and nationalities. The cultivation of cross-cultural awareness is very important in the teaching and learning of international Chinese education. Setting up stalls to bake cold noodles, sell ice powder, and bake skewers on the streets of the United States, watching how foreigners spend the Spring Festival in Chinatown during the Spring Festival, and so on. These videos show the possibility of combining the study of Chinese culture with the cultivation of cross-cultural communication (He & Chiang, 2016). Through such videos, teachers can not only enable students to understand relevant cultural knowledge, but also see them as a window to understand the views of people from other countries and nations on Chinese culture, so as to learn to communicate from a cross-cultural perspective. These videos contain rich intercultural communication connotations, such as the values, communication methods, customs, and habits embodied in them. On the one hand, they can understand cross-cultural knowledge; on the other hand, they can also generate awareness of cross-cultural communication. Using these videos as teaching materials in class can help teachers better explain Chinese culture and help students form a distinctive and unique impression of Chinese culture (Jin et al., 2020). Teachers can use videos to guide, assist with certain explanations or set up thematic discussions or invite students to exchange speeches, so as to cultivate students’ awareness of cross-cultural communication and make students feel the importance of cross-cultural communication.

**Shortcomings of Short Video Caption Generation Algorithm Applied to International Chinese Education and Culture Teaching**

Short video subtitle generation algorithm applied to international Chinese education and teaching has the following shortcomings.
One Sidedness

We Media videos related to Chinese culture on the We Media platform are often introduced by diversity video, whether it is Chinese food, Chinese clothing or interviews, which can be as short as two or three minutes or as long as 20 or 30 minutes. However, even long videos can hardly explain one aspect of Chinese culture clearly, and can only serve as a simple introduction. In addition, short videos are actually the works of We Media people (Wang et al., 2019). To some extent, they reflect the author’s own views and opinions. Whether such views and opinions are correct, scientific, and comprehensive also needs to be determined. In addition, due to the lack of strict identity requirements for the authors of We Media, some short videos lack professionalism, which is expressed as a statement of one mind, and it is difficult to put forward comprehensive and correct opinions on a problem.

Lack of Systematicness

There is no logical connection between many We Media videos and videos. This is because the fundamental purpose of We Media video production is to attract people to watch, that is, to attract viewers. Therefore, most video content is displayed in a way that attracts people’s eyes. If only the form is considered, the content value may be affected to some extent. Students have a better understanding of Chinese culture after watching it, but it can only be limited to this video itself. Thus, if people watch many such videos, can they achieve a profound understanding? The answer is also no, because each video content is fragmented and lacks systematicness.

Inadequate Pertinence

Although many videos related to Chinese language teaching are published on our media, the number of fans and viewers is not high. Short videos involve a wide range of aspects, but the content is fragmented. Some focus on grammar explanations, some share daily communication terms, and some introduce Chinese culture. This kind of video can play a role in spreading Chinese culture and making foreign students understand Chinese traditional customs to a certain extent (Fang et al., 2016). However, such videos often do not have a clear audience, that is, the students’ Chinese level, age, and nationality, and cultural popular science videos cannot focus on a problem, and can only introduce Chinese culture in a general way, it is not made for the characteristics of foreign students learning Chinese. Among the various kinds of We Media, it is not easy to have We Media specialized in international Chinese language education. Short videos of other cultures are more difficult to be targeted. If a short video with only a few minutes is not designed according to students’ characteristics and learning conditions, it is difficult to achieve real teaching effect.

SHORT VIDEO CAPTION GENERATION DEPTH MODEL ALGORITHM

Short videos are numerous and well produced, with the following characteristics:

1. **Rich Content**: The videos produced by foreign short video bloggers in China are rich in content, focusing on all aspects of contemporary life in China, including film and television works, food, transportation, science and technology, life customs, hot words on the Internet, tourism, current events, as well as some traditional Chinese culture, such as traditional Chinese medicine culture. These contents are close to the contemporary life of Chinese people and depict a new and real China to foreigners; at the same time, these contents are also the hot topics in which Chinese learners are most interested. The use of short videos of foreigners in China in classroom teaching can help Chinese learners further understand and understand contemporary Chinese society and culture (Yu et al., 2017).

2. **Interesting**: In addition to the richness of content, the interest of video presentation is also an important factor to ensure that foreign bloggers in China attract and retain fans. In order to
improve the interest and attraction, they often adopt various forms of expression according to the characteristics of different topics, such as single talk show, double or multiple interviews, personal experience, and role play. At the same time, in the process of video shooting and production, they will also consciously use some humorous language, dramatic body and expression, as well as post production with personal characteristics to create a relaxed atmosphere of variety, so that fans can always maintain a happy mood and deepen their impression in the watching process.

3. **Short and Pithy:** Chinese courses at the primary stage often do not include special cultural courses, and most of them are simply introduced in the comprehensive courses. Due to the limited time of a class, multimedia resources such as film and television works and variety shows often have the disadvantages of too long time and too scattered cultural distribution, and are not suitable for classroom teaching (Jin et al., 2020). If only fragments are intercepted, it will affect the continuity of the content and increase the workload of Chinese teachers. Short videos are generally short in duration and concentrated in cultural points, which just make up for the shortcomings of other multimedia resources.

4. **Bilingual Subtitles:** Students at the primary stage are new to Chinese and have an urgent desire to understand Chinese culture. However, their Chinese level is not high enough to support them to receive information on social media independently. Film, TV series, and other film and television works also have the problems of too hard to express and too fast, which limit the channels for students at the primary stage to understand Chinese culture. The short video bloggers of foreigners in China have a natural bilingual identity. They will match their videos with accurate and authentic bilingual subtitles. Chinese learners can not only understand the video content well, but also learn some Chinese expressions (Yang et al., 2022).

5. **They Keep Pace With the Times:** There has always been a tendency of “emphasizing the past and ignoring the present” in the cultural teaching of foreign students. It is understandable that when people think of cultural teaching, they think of traditional cultures such as Taijiquan and Peking Opera. At the same time, they should also see that many Chinese learners are eager to understand China’s current popular culture, but they have no way or means to obtain information, which is also an important reason why many Chinese learners always have a backward stereotype of China. In today’s era of rapid information updates, in order to always maintain popularity, foreign short video bloggers in China must keep pace with the times when choosing video topics, and pay attention to social hot spots or current popular culture. Short video bloggers of foreigners in China are also happy to share some new changes in China, such as our convenient and fast mobile payment, high-quality and fast high-speed rail, and excellent film and television works, which will help Chinese learners in nontarget language environments understand the current China, break their inherent impression of China, and thus have a new and positive understanding of China.

The feature extraction method the author used in this experiment is MFCC. The ability of the human ear to perceive sound waves at different frequencies varies. Experiments have found that, under 1000 Hz, the perception ability is linearly related to the frequency, while, above 1000 Hz, the perception ability is logarithmically related to the frequency. Figure 3 shows the visual presentation of model features in this apper. The useful information of speech is mostly concentrated in low-frequency components, while high-frequency components are susceptible to noise interference. MFCC converts frequency to Mel domain, emphasizing low-frequency speech information, making it easy to extract feature information that is conducive to recognition, and, more importantly, shielding against noise interference.

Video caption generation is a technology that uses computer models to generate text annotations for specific objects and scenes in video. It involves the understanding of objects, people, scenes, events, temporal relationships, and many other aspects. The creation of short videos consists of content planning, video shooting, post processing, and other links. As a part of post processing,
subtitle generation is usually manually added by the author. It takes a long time and energy to generate subtitles manually. At present, some subtitle generation software for video has been developed and applied. According to the voice in the video, text subtitles are exported through speech recognition technology. On this basis, it has powerful functions such as manual modification and translation, and is highly practical. However, the above software mainly relies on audio information in videos, and words cannot be extracted without clear audio information. In real life, ordinary users want to record their life through short videos. If they shoot while explaining, they will inevitably feel embarrassed to talk to themselves (Gong et al., 2018). If they record later, they will return to the problem of time-consuming and laborious. According to the survey, the automatic generation of subtitles for short videos without audio background has pain points and technical gaps. With the growing demand for video captioning, the development of domestic speech recognition technology is relatively perfect, and some domestic short videos begin to support automatic captioning. In addition, many video editing software programs also provide subtitle production. Most video editing software programs do not support automatic subtitle at present. In addition, domestic educational software has also begun to rise. As a language, English education is also essential. These educational platforms also have an increasingly strong demand for automatic subtitle services, providing opportunities and development prospects for the emergence of automatic subtitle generation systems.

To test the antinoise performance of linear prediction cepstrum coefficient feature parameters and MFCC feature parameters as well as their effect on the system recognition rate, noise with various signal-to-noise ratios must be added to the original pure speech signal. Figure 4 shows the effect diagram of a language recognition system based on the above two feature parameters under different noise conditions. Experimental data show that speech recognition systems have high recognition rates for pure speech signals. With the addition of noise with different signal-to-noise ratios, the recognition rate of a speech recognition system decreases rapidly: The higher the signal-to-noise ratio, the lower the recognition rate of the speech recognition system. Whether it is a pure speech signal or a speech...
signal with noise added, the recognition rate using MFCC feature parameters in speech recognition systems is higher than that using linear prediction cepstrum coefficient feature parameters.

In this paper, the author proposes an international Chinese education and teaching system based on short video subtitle generation algorithm. This system provides users with automatic subtitle services, which not only supports Chinese subtitles, but also supports English subtitles, saving users much time and improving their sense of experience (Chiang, 2020). The main functional modules of the system include automatic speech recognition module, audio event detection module, task scheduling module, Intelligent Terminal Network digital standardization module, self-service test module, subtitle module, and API Gateway module. The author not only describes the functional requirements of the system in detail, but also introduces the nonfunctional requirements in depth. Secondly, the author makes a comprehensive description of the overall architecture and hierarchical structure of the system. Then, the researcher describes the feature extraction algorithm MFCC and speech recognition model LSTM+CTC (Connectionist Temporal Classification) of automatic speech recognition. After that, the author introduces and details the algorithms involved in each module. The author uses:

- For the task scheduling module, the topology structure related algorithms.
- For the API Gateway module, the smooth weighted load balancing algorithm.
- For the self-service test module, the A-Res weighted reservoir sampling algorithm and the editing distance algorithm.
- For the subtitle module, the AC automata multimode matching algorithm (Cao & Li, 2018).

The project includes the adoption of a large number of traditional algorithms, which shows that traditional algorithms are also widely used. For speech recognition and speech event detection, machine learning depth learning can achieve very good recognition effect, in this field. The author uses the traditional C/S architecture, Linux operating system and C++ language as the server development,
Python as the client development, multicluster deployment to ensure the reliability of the system, Redis as the cache, and Hive handles big data. Thrift rpc and http are used to process requests, and API gateway is used to distribute requests to the corresponding machines, reducing the pressure on the system and ensuring the high availability of the system. Figure 5 shows the system hierarchical structure diagram.

Video caption generation belongs to the field of video understanding. The field of video understanding can be divided into three levels, from the bottom to the top.

1. **Video Semantic Segmentation Task**: This is a pixel level analysis, that is, it can identify the corresponding semantic tag of each pixel on each frame image in the video.
2. **Video Detection Task**: It is an object level understanding that can distinguish the location of the corresponding target object and identify the category to which it belongs. Its granularity is coarser than that of splitting tasks (Bin et al., 2018).
3. **Understanding the Whole Video at a Higher Granularity**: Video subtitle generation belongs to this level. The computer automatically recognizes the video and generates accurate subtitle summary words or descriptive statements that conform to natural language syntax, without specifically identifying the category labels corresponding to pixels on each frame of the video or fine-grained identifying the location and category of specific objects.

---

**Figure 5. System hierarchical structure diagram**
Selecting Appropriate Teaching Materials

Firstly, the culture introduced by the selected short video should be positive and representative. As the author mentioned above, short videos are produced on the Internet, and the video content is mixed. Therefore, international Chinese education teachers should have a sense of judgment and selection, and choose to introduce China’s excellent, symbolic, and representative cultural content. For example, some short videos about the ancient three cardinal principles and five virtues, social bad habits, and feudal ideas should not be broadcast in the classroom. It should also be noted that some videos have no problem with the overall content, but may involve some non mainstream and unrepresentative cultures (Zhang et al., 2019). For example, in the short video, some bloggers once went to China’s ethnic minority areas to introduce the life and customs of the local people, including the past living conditions of the local people. Thus, if these videos are simply played without distinguishing and introducing them, students may misunderstand them, that is, they mistakenly think that Chinese culture is the same all the time, and they may also create a stereotype.

Then, the content of the selected video, especially the Chinese introduction, should be appropriate and not too difficult, and should conform to the students’ Chinese level. At the same time, if the Chinese part of the video is too brief or the pronunciation is not standard, it is not conducive to students’ learning, and even errors will occur (Huynh-Thu & Ghanbari, 2008). On the contrary, some popular science documentaries are of great significance for foreign students to learn, but the proportion of language introductions is often too large, with many proper nouns, Chinese elements, historical figures, and stories, which will make it difficult for students to understand and remember (Lee et al., 2012). This requires to choose teaching videos that are suitable for students according to their Chinese level: For junior or young students, teachers can choose videos with more videos, such as those shot by some photographers. On the contrary, intermediate or advanced students can choose videos with more Chinese introductions according to their specific level.

In this paper, the author verifies the effectiveness of the proposed video caption generation method based on attention mode fusion through experiments on microsoft research video description corpus (MSVD). YouTube 2Text dataset (MSVD) is provided by Microsoft Research, and all videos are from the YouTube video Web site. In this research, the author also used 1200 video clips for training, 100 video clips for verification, and 670 video clips for testing.

In order to intuitively reflect the quality of the actual generated description, this paper includes two video segments in the MSVD dataset and the description of the target video segment generated by the model in the study (Figure 6). In this study, the author gradually fuse the extracted video stream features (i.e., time, motion, audio, and semantic features) through an attention-based hierarchical approach. Therefore, the author generate video subtitles based on different characteristics of multimodality. For example, the description of “a woman performing a dance” tends to rely more on movement characteristics than sound characteristics. However, the description of “men listening to music” relies more on audio features than motion features. Motion and audio features have a relatively complementary relationship. Therefore, the attention-based multimodal fusion video caption generation method proposed in this paper has high research value.

Choosing Appropriate Teaching Methods

Firstly, teachers should introduce the culture they have learned and need to know to students. In class, students should be pointed out the cultural content they want to learn today. For example, to introduce Chinese festival culture, teachers can ask students to collect relevant information about Chinese festivals before class, let students share and make a brief introduction in class, and then play videos of relevant festivals. This can not only enable students to watch videos purposefully, but also ensure that students have certain knowledge background before watching videos, which can give full
play to cultural video teaching. At the same time, teachers should also explain the possible cultural differences in the video in advance to avoid cultural conflicts.

In addition, teachers should guide students to pay attention to the important details in the video. Videos can improve students’ attention to a certain extent, but they may also make the students too focused on the video form and ignore the video content. Therefore, in the process of video playing, teachers should give students hints, explanations, and questions on key points, and remind students to pay attention to the video content in this way. After playing the video, teachers can monitor and adjust the students’ understanding process through discussion and speech, individual questions, and group cooperation. They can also check the learning effect of students watching videos by assigning homework after class. Figure 7 shows the comparison between Chinese Chinese Sign Language (CSL) data test set and other methods.

Creating a Strong Teaching Atmosphere

Firstly, because of its unique interest and timeliness, short videos reflect the Chinese contemporary culture in which students are interested. Dynamic sound and pictures can also make students feel fresh, giving them a sense of participation and cultural experience to the greatest extent. We Media videos are interesting and fresh for students. For example, the dialect culture and variety shows shown in the cases are cultural factors about which students do not know much before, but meet their interests.
This evidences that proper selection of video materials can make students more focused. Through video teaching, they can also accept cultural knowledge more and have a sense of living in China. Figure 8 shows the video subtitle generation results on dataset.

Then, Chinese teaching is not only teaching language elements and training language skills, but also teaching cultural factors, which is a consensus reached by the academic community of Chinese as a foreign language. Although there is no special culture course in the junior class, how to expand cultural factors in comprehensive courses is also an important topic. After class, students will independently search for other videos of bloggers introduced in the class and watch them. At the end of the class, they expressed their love for Chinese and their willingness to continue learning Chinese. It can be seen that We Media videos have changed students’ views on China to a certain extent and stimulated their interest in learning Chinese.

**CONCLUSION**

This paper focused on the topic of automatic generation of video subtitles for international Chinese language education and teaching courses. In the whole process of automatic subtitle generation, speech recognition is the core technical issue, so the full paper focused on speech recognition. First of all, the author proposed a new video caption generation model based on spatial embedded encoding and decoding structure. The decoder network also implements the deep stack residual gating loop unit structure, and sets the planned sampling for the word generation training. In order to make better use
of video spatiotemporal information and shorten the modal distance between generated word meaning and visual features, the author mainly studied the relationship between generated words and characters in different video frames in video data. Then, the author proposed a new video caption generation and update model based on video timing attention fusion mechanism. This model combines the abstract features of low-level video content related semantics and high-level space embedded features, which can more effectively generate words and form the final subtitle statement. Finally, in order to verify the effectiveness of the proposed model, the researcher conducted an experiment on another large Chinese continuous sign language recognition dataset, China CSL, and compared it with other public methods. The results showed that this method can achieve the best results.

In terms of temporal attention level fusion mechanism, in this study, the author only took the first layer and the top layer of the encoder network as the relevant output fusion layers. If the second layer, the third layer, and even the space embedding layer are considered together, they may produce better results. In addition to the attention operation on the video frame corresponding to the time step position, more powerful computing power in the future will allow to carry out the spatial attention operation on the spatial dimension of the video frame, fuse the attention background vector of the two dimensions, and better mine the spatiotemporal information of the video frequency. In brief, the author hopes that this paper can provide useful reference and new ideas for international Chinese education and teaching, so that international Chinese education and culture teaching can adapt to local conditions, so as to better help foreign students adapt to China’s life abroad, learn Chinese better, and understand Chinese culture better.
DATA AVAILABILITY
The Figures used to support the findings of this study are included in the paper.

CONFLICTS OF INTEREST
The author declares no conflicts of interest.

FUNDING STATEMENT
This work was not supported by any funds.

ACKNOWLEDGMENT
The author would like to show sincere thanks to those technicians who have contributed to this research.
REFERENCES


