Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.037503
Article

A Sentence Retrieval Generation Network Guided Video Captioning

Ou Ye[1,2], Mimi Wang[1], Zhenhua Yu[1,*], Yan Fu[1], Shun Yi[1] and Jun Deng[2]

[1]College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, 710054, China
[2]College of Safety and Engineering, Xi'an University of Science and Technology, Xi'an, 710054, China
*Corresponding Author: Zhenhua Yu. Email: zhenhua_yu@163.com
Received: 06 November 2022; Accepted: 27 February 2023

Abstract: Currently, the video captioning models based on an encoder-decoder mainly rely on a single video input source. The contents of video captioning are limited since few studies employed external corpus information to guide the generation of video captioning, which is not conducive to the accurate description and understanding of video content. To address this issue, a novel video captioning method guided by a sentence retrieval generation network (ED-SRG) is proposed in this paper. First, a ResNeXt network model, an efficient convolutional network for online video understanding (ECO) model, and a long short-term memory (LSTM) network model are integrated to construct an encoder-decoder, which is utilized to extract the 2D features, 3D features, and object features of video data respectively. These features are decoded to generate textual sentences that conform to video content for sentence retrieval. Then, a sentence-transformer network model is employed to retrieve different sentences in an external corpus that are semantically similar to the above textual sentences. The candidate sentences are screened out through similarity measurement. Finally, a novel GPT-2 network model is constructed based on GPT-2 network structure. The model introduces a designed random selector to randomly select predicted words with a high probability in the corpus, which is used to guide and generate textual sentences that are more in line with human natural language expressions. The proposed method in this paper is compared with several existing works by experiments. The results show that the indicators BLEU-4, CIDEr, ROUGE_L, and METEOR are improved by 3.1%, 1.3%, 0.3%, and 1.5% on a public dataset MSVD and 1.3%, 0.5%, 0.2%, 1.9% on a public dataset MSR-VTT respectively. It can be seen that the proposed method in this paper can generate video captioning with richer semantics than several state-of-the-art approaches.

Keywords: Video captioning; encoder-decoder; sentence retrieval; external corpus; RS GPT-2 network model

1 Introduction

Video data contains a wealth of information, which is also the most common information carrier in our daily life [1]. With the rapid development of multimedia and network technologies, the scale of video data is rapidly increasing. To facilitate people's understanding and selection of the desired video content, studies on video captioning have attracted more and more attention from academia and industry. Video captioning is to describe the contents of a given video using natural language [2], which is one of the critical tasks in computer vision and natural language processing [3–8]. In recent years, how to automatically extract and enrich the contents expressed in videos by computers has become one of the hot spots in the studies on video captioning. The studies on video captioning aim to achieve high-level semantic cognition and natural expression of visual contents, which mainly involve several research fields such as computer vision, natural language processing, and machine learning, and have broad application prospects in video retrieval, aided vision, video surveillance captioning and sign language translation [9].

Early video captioning methods mainly include two types of techniques that are based on templates [10] or based on retrieval [11]. Among them, the template-based approaches will limit the semantics of description sentences due to fixed templates. At the same time, the retrieval-based approaches rely too much on databases and manually constructed annotations, the generated textual sentences lack diversity. In recent years, various methods based on encoder-decoder structures have been widely used to address the task of video captioning. This kind of approach mainly exploits visual information by thoroughly learning high-level semantic features of videos [12] or introducing attention mechanisms [13], which utilizes decoders [14] to achieve semantic alignment between visual and text media. However, only learning high-level semantic features of videos [15] tends to ignore more detailed information in videos, which causes a problem of inaccurate semantic descriptions of videos. Furthermore, although adding attention mechanisms [16] can optimize the selection of visual features, the emotions, logic, personalization, and implicit semantics in videos are often ignored. In general, the existing studies focuses on the generation of video captioning through the encoding and decoding process of the visual feature information. However, less consideration is given to how to introduce the external corpus information to accurately predict words that conform to the context semantics to enrich the semantic content of video captioning. Therefore, at present, few relevant studies consider using external information to guide the generation of video captioning, which is not conducive to the accurate description and understanding of video content.

When videos are the only input source without any other information expansion, it can be considered to add the guidance of external information to enrich the semantic content of video captioning, which is also more in line with the human description habit. For example, as shown in Fig. 1, the title is "A man is cutting bread with a knife on a table". Even though the two objects ("knife" and "table") are not easily extracted from the visual information, it can still infer from the external information expansion that the tool used for cutting bread is a knife, and the whole process happens on a table. The relationship "cutting bread with a knife" is a commonsense relationship between a "knife" and "bread" rather than one obtained from the video content. Therefore, the combination of external information and visual information is also in line with the human habit of describing things.

In this paper, a novel video captioning method guided by a sentence retrieval generation network (ED-SRG) is constructed, which retrieves semantically related sentences in a corpus based on the generated textual sentences by an encoder-decoder, and filters out candidate sentences through similarity measurement to achieve the introduction of external corpus information. On this basis, a

novel RS GPT-2 generation model is constructed, which uses a random selector to accurately predict words that match the context semantics to guide the generation of video captioning.

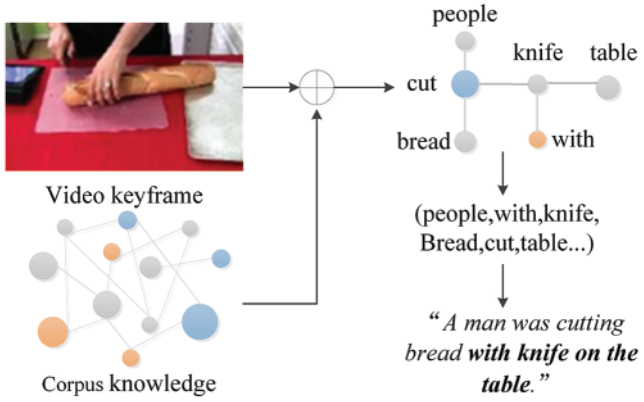


Figure 1: Example of retrieving external information based on statements to facilitate video captioning

In summary, the contributions are summarized as follows:

(1) A novel sentence retrieval generation network model is constructed in this paper. The generated textual description sentences by a mixed ResNeXt-ECO encoder and a long short-term memory (LSTM) decoder are utilized to retrieve some sentences with their semantic approximations in an external corpus as candidate sentences.

(2) A novel RS GPT-2 network model is constructed to measure the correlation between corpus words and similar sentences by designing a random selector so that the predicted words with higher correlation can be randomly selected and utilized to generate textual sentences, which are more consistent with natural human language expressions.

(3) The proposed method is verified on two public datasets (MSVD and MSR-VTT datasets) with a significant difference in size to evaluate the different effects of external corpora on the generation of video captioning and analyze the potential relationship between external corpora and generated video captioning, which provides help for in-depth studies on how the external information can guide the generation of video captioning.

2 Related Work

Early video captioning methods are based on template matching. For example, the study in [17] first detected objects, predicted attribute relationships and constructed language templates, and then utilized the language templates to generate sentences. However, this method mainly relies on the detection qualities of objects and attributes, and the effect of generating sentences is poor. To address this issue, the work in [18] shifted the focus to the core structural representation of sentences, which was represented as a "noun-verb-scene-preposition" quadruple form and used a Markov model to select the most appropriate quadruple to generate the sentences. However, this method is based on fixed templates and is constrained in terms of the semanticity of the sentences. To overcome the issue of template matching-based video captioning methods, retrieval-based video captioning methods were proposed. This kind of method retrieves videos that are visually similar to the target videos and then obtains similar video captioning as the generated textual sentences of the target video [19]. For example, the study in [20] generated video captioning by constructing an extensive database of

descriptive sentences to exploit topic retrieval. However, it relies only on the set of sentences in the database and limits the diversity of generated description sentences. In [21], to address this issue, a collection of sentences was constructed manually, from which sentences semantically similar to video contents were retrieved to generate video captioning. Nevertheless, the description sentences generated by this method are limited by manual annotation and cannot be easily extended. In general, the template-based video captioning method will restrict the semantic nature of the description sentences because of the fixed templates. The retrieval-based video captioning method relies too much on databases and manually constructed annotations. Hence, the generated description sentences lack diversity.

In recent years, the studies work on video captioning are mainly based on the encoder-decoder models [22–24] to study how to utilize better video features [25], attention mechanisms [26], and other ways to generate textual description sentences for video contents directly. For example, the study in [27] extracted frame features, spatio-temporal features, and audio features of video data based on encoder-decoders to generate textual description sentences by using multimodal information of video data. Moreover, a recurrent neural network based on a standard attention model in the encoder-decoder was proposed in [28] to encode visual and text features by integrating attention mechanisms, which can improve the accuracy of video captioning to some extent. In [29], a joint commonsense and relation reasoning method was proposed, which exploits prior knowledge for image and video captioning without relying on any detectors. Hou et al. [30] proposed a video captioning method based on grammatical expression and visual cue translation to express syntactic structures through sentence templates composed of lexical tags, which can generate video captioning with fixed grammatical structures. The work in [31] proposed a semantic-based video keyframe method that extracts the initial video keyframe using a CNN model and the feature windows. Then the keyframes were automatically marked by the image caption network. Finally, a pre-interactive LSTM network model was proposed to generate the video captioning. This method can fully extract the semantic information of a video using the video keyframes and improve the accuracy of video captioning. However, the quality of the generated video captioning depends on the accuracy of keyframe annotations. Liu et al. [32] proposed Unpaired Video Captioning with Visual Injection system (UVC-VI) to address the issue that sufficient paired data is unavailable for many targeted languages. However, the semantic information of video captioning is still limited. The study in [33] proposed a novel global-local encoder to produce rich semantic vocabulary for obtaining a descriptive granularity of video contents across frames. However, this method only enriches semantic vocabulary by encoding different visual features, and the source of semantic information is relatively single. In summary, although the above existing studies can produce rich semantic vocabulary to generate video captioning by using the visual encoding features, the video captioning methods based on an encoder-decoder still mainly rely on a single video input source and limited lexical tags. Due to less consideration of introducing the rich information from the external corpora, the generated video captioning has certain limitations in the expression of the video content and lack of semantic diversity, which is not conducive to the accurate presentation of video content.

To address this issue, several approaches have been proposed to enrich the semantic information of textual description sentences under the guidance of external data. Take image captioning [34] as an example, Aditya et al. [35] employed common sense reasoning to detect a scene description graph in images and translated this graph directly into description sentences through a template-based language model. The work in [36] combined external information graphs to enhance image information to generate image captioning better. However, both of these approaches extract explicit semantics directly from external input. The generated description sentences are too rigid and do not conform to natural human language expression habits. Hou et al. [37] constructed semantic graphs by

extracting image information and employed external information for relational reasoning to generate textual description sentences of images. However, this method only relies on image information for relation reasoning and lacks the expansion of external data. In general, the semantic diversity of image captioning can be effectively improved by introducing external data. However, video captioning is more challenging compared with image captioning. It is because the spatio-temporal features of video frames are more closely related, and the semantics are richer.

Currently, the video captioning models based on an encoder-decoder mainly rely on a single video input source. The contents of video captioning are limited since few studies employed external corpus information to guide the generation of video captioning, which is not conducive to the accurate description and understanding of video content. To address this issue, this paper considers utilizing sentence retrieval to introduce the external corpora in the video captioning task. On this basis, a random selector is designed to select prediction words with high relevance from the retrieved sentences to enrich the semantic information of video captioning. Thus, video captioning of more consistent with natural human language expressions that can be obtained.

3 The Proposed Method

The general idea of the video captioning method guided by a sentence retrieval generation network is shown in Fig. 2. This method consists of an encoder-decoder and a sentence retrieval generation network model. The encoder-decoder contains a ResNeXt-ECO encoder and a LSTM decoder, which are used to generate an initial sentence of video captioning y. The sentence transformer network model [38] in the sentence retrieval generation network is used to retrieve the first k ($k = 5$ is set in the experiments) sentences that are semantically similar in an external corpus to construct a set $z$according to y. The retrieved sentences are expressed as s, denoted as $z = \{s_i, \ldots, s_k\}$. Then the similarities between each candidate sentence in z and y are calculated. After that, by constructing a novel RS GPT-2 network model, the correlations between words in the external corpus and each word in the generated sentence y are calculated. The word with the most remarkable correlation is randomly selected as the predicted word to gradually generate a sentence of video captioning y'. The probability of a sentence of video captioning generated by this method is calculated as shown in Eq. (1):

$$p\left(y'|y\right) = \prod_{t=1}^{T} \sum_{i=1}^{topk} p_{\mu}\left(s_i|\text{corpus}\left(y\right)\right) p_{\delta}\left(y'|c_i, y, F\right) \tag{1}$$

where corpus (y) denotes a function that uses an initial sentence of video captioning y for retrieval in an external corpus, and *topk* denotes a retrieval sentence with the highest similarity, p_{μ} is the probability of retrieving a sentence s_i, p_{δ} is the probability of outputting a sentence of video captioning y', c_i denotes a word in the external corpus, F denotes the semantic features in a similar sentence s_i, and t denotes a moment in time T.

3.1 The Encoder-Decoder with the Hybrid of ResNeXt-ECO and LSTM

To complete the video captioning task, it is necessary to extract the visual features of video data for encoding firstly. Since the ResNeXt network [39] is a variant of the ResNet residual network, which can reduce the number of hyperparameters and improve the accuracy of model prediction without increasing the complexity of the model, it can be used to extract the spatial features of video data. In addition, considering the time series of video data, since an ECO (Efficient Convolutional Network for Online Video Understanding) [40] can better describe the relationships between actions in the time intervals, it can be used to extract the temporal features of videos. Currently, although some lightweight network models, such as the Mobile network model, can also be used to extract the video

features, considering the limited information capacity and the insufficient feature representation ability of lightweight network models, the problem of a heterogeneous gap in the generation process of video captioning will be magnified. In this paper, the ResNeXt and ECO network models are mixed firstly as a ResNeXt-ECO encoder to extract 2D spatial features and 3D spatio-temporal features of video data to represent the salient features of videos.

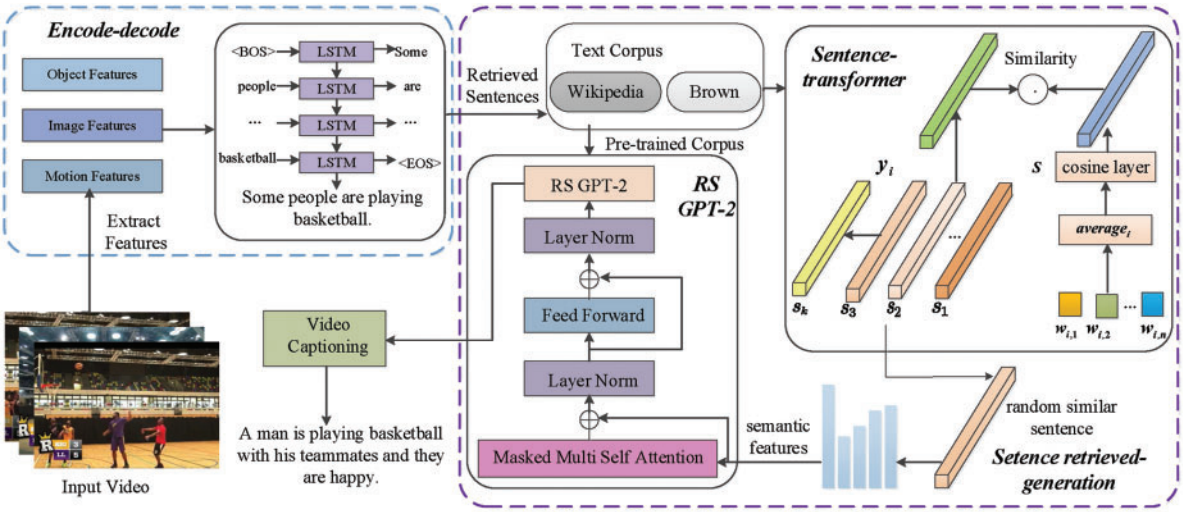


Figure 2: The framework of the video captioning method guided by a sentence retrieval generation network

To represent the salient features of video data more accurately, a residual network ResNeXt with 101 layers of network depth is utilized to extract two-dimensional features of videos. This network divides the input channels into groups, and then the residual operation is performed on these groups. In addition, there is a maximum pooling layer and an average pooling layer in this network. Finally, the outputs of all groups are merged through a 1×1 convolution. The basic structural unit of the ResNeXt network is shown in Fig. 3, where C is the cardinality, which denotes the number of all identical branches in a residual module. In this paper, C is set to 32.

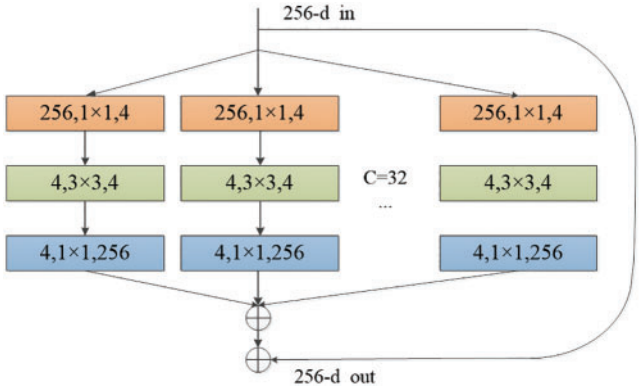


Figure 3: The basic structural unit of the ResNeXt network

First, T (set $T=16$) frames are randomly selected in key frames from the input videos, and the size of each video frame is set to 256×256. Then, the ResNeXt blocks consisting of 1×1, 3×3 and

1×1 convolution kernels are utilized for feature extraction of video frames. Finally, through the fifth convolutional layer of the ResNeXt network and an activation function Rectified Linear Unit (ReLU), a feature vector of size [2048, 1] can be obtained, which is denoted as x_s.

In addition, on the basis of using the ResNeXt network to extract two-dimensional features, an ECO network is mixed to extract the three-dimensional features of video data. The structure of the ECO network is shown in Fig. 4, which contains two 2D Net and one 3D Net component, both of that use the activation function (ReLU) to extract the video features. Moreover, there are 3 pooling layers in the ECO network model. First, a 2D Net component is employed to extract the static features of video frames. All these static features are used as the input of the 3D Net component to extract the time series features of videos. Meanwhile, the ResNeXt network is utilized to replace the original 2D Net component in the ECO network. The features are mixed and extracted by the ResNeXt network model in the parallel part of the 3D Net component. Thus, the output features of size [1024, 1] can be obtained. Then vector stitching is utilized to stitch the features of size [512, 1] from the parallel part of the 3D Net component. Finally, the spatio-temporal features of size [1536, 1] are obtained through the global pooling layer of the ECO network model, which is expressed as x_v.

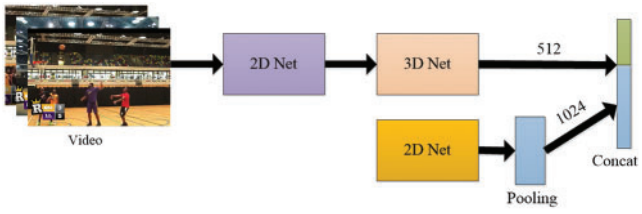


Figure 4: The structure of the ECO network

An example of the feature maps generated by 2D and 3D feature extraction modes is shown in Fig. 5, where the edge features of the video keyframes are salient in the first convolution layer. However, with the increase of the convolution layer, the feature maps are more and more abstract. Finally, the high-level features of video frames are obtained.

To obtain richer visual information, the Faster R-CNN network model is utilized as the target detector [41] to extract the object features in the videos based on the ResNeXt-ECO encoder, and the dimensions of these features are [4096, 1], denoted as x_o. Finally, the features, x_s, x_v and x_o are spliced to obtain a feature vector of size [3548, 1], as shown in Eq. (2):

$$X = \text{Concat}\,(x_s, x_v, x_o) \tag{2}$$

where Concat $(\cdot)$ represents the vector splicing function.

In the video captioning task, the decoder is mainly used to decode the encoded visual features, and each output word is gradually obtained. Thus, the entire textual description sentences are formed for video content. In short, decoding is also one of completing the semantic alignment between video features and text descriptions. In the decoding stage, a long short-term memory network (LSTM) is usually used to decode the encoded features. Therefore, an LSTM network is adopted as the decoder, which consists of a single-layer LSTM network with 256 units.

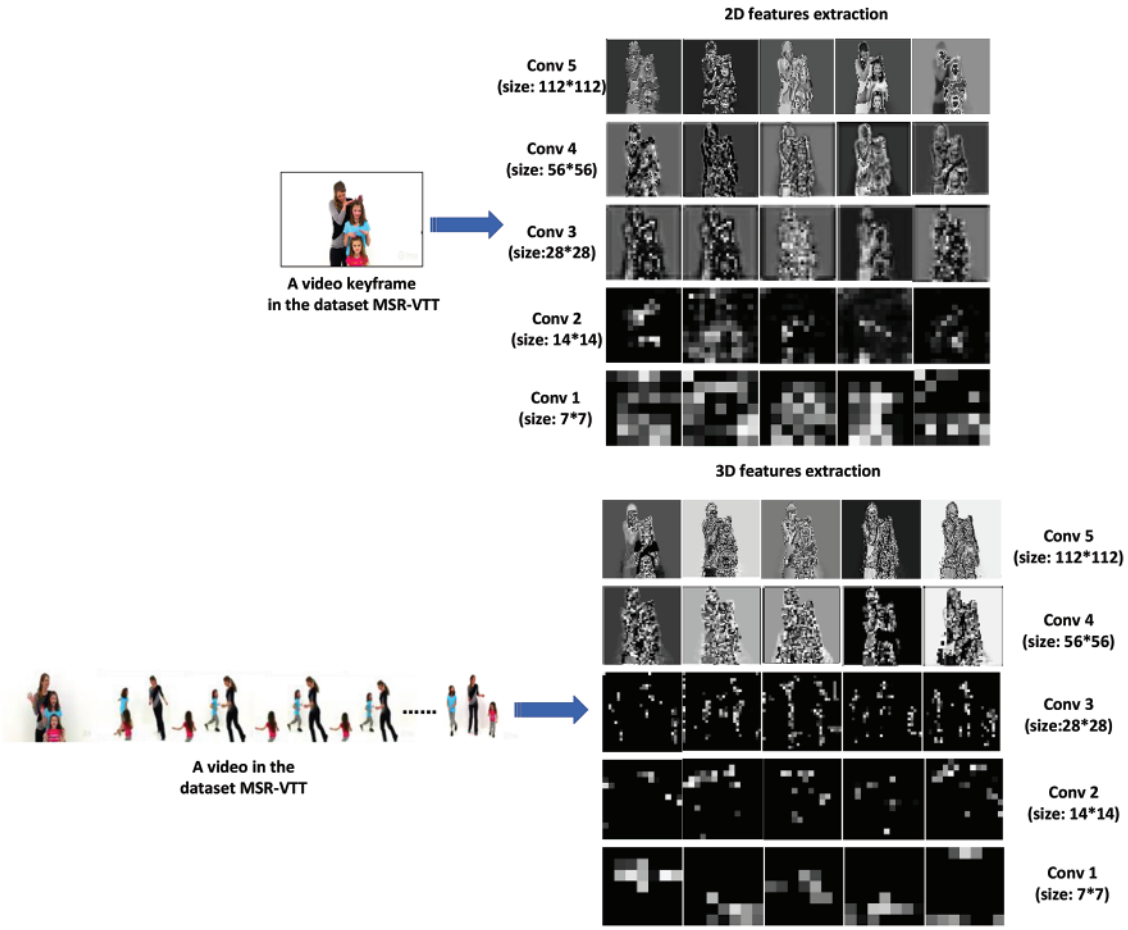


Figure 5: The features maps generated by the 2D and 3D features extraction modes

First, the video features X output by the encoder are injected into the LSTM network, and the hidden layer state is shown in Eq. (3):

$$h_t = \text{LSTM}\,(x_t, h_{t-1}, m_{t-1}) \tag{3}$$

where h_t denotes the hidden state at the t moment, $x_t \in X$, m_{t-1} denotes the memory unit at the $t-1$ moment, its output goes through the softmax layer, and the generated word on each node is $w = (w_1, w_2, \ldots, w_t)$. Therefore, a textual description sentence of the output at time t is shown in Eq. (4):

$$\log p\,(y|w_1, w_2, \ldots, w_{t-1}) = f\,(h_t, c_t) \tag{4}$$

where c_t denotes the output at the time t. Finally, in the decoding process, "BOS" and "EOS" flags are respectively used as the start and end words of the LSTM network to generate the sentence. The words generated at each moment are sorted according to the time series and connected to form a text description as an output result, denoted as y, and $y = \{< BOS >, w_1, \ldots, w_t, < EOS >\}$.

3.2 The Construction of Sentence Retrieval Generation Network

In this paper, a novel sentence retrieval generation network is constructed, which first retrieves the top k sentences from a large external corpus that are most similar to y. On this basis, this paper improves the original GPT-2 network model [42] to design a new RS GPT-2 network model. Introducing the designed random selector in the RS GPT-2 network can eliminate the influence of repeated redundant words on the accuracy of text description. By using the random selector, a word is randomly selected from the words with high probability to predict the next word one by one. By analogy, the textual description sentences of video contents are finally generated.

(1) Sentence Retrieval

In the sentence retrieval generation network, since the sentence-transformer is a network model with text similarity search and optimization, which has high retrieval efficiency and is more accurate, this network is adopted to extract textual description sentences with semantics similar to y from an external corpus. First of all, initial video captioning y is used to find the top k sentences that are semantically similar to y in the corpus to form a set z. The corpuses chosen for this paper are the English Wikipedia corpus and the Brown corpus. The probability of a retrieved sentence s_i is estimated as Eq. (5):

$$p_\mu\left(s_i|corpus\left(y\right)\right) = softmax\left(\text{sim}\left(y, s_i\right)\right) \tag{5}$$

where $p_\mu\left(\cdot\right)$ denotes the probability estimation of a retrieving sentence s_i in an external corpus using y, $softmax\left(\cdot\right)$ is the normalized exponential function, and $\text{sim}\left(\cdot\right)$ denotes the similarity measure function between sentences, which is calculated as shown in Eq. (6):

$$\text{sim}\left(y, s_i\right) = \frac{y \cdot s_i}{\|y\| \, \|s_i\|} \tag{6}$$

the closer the similarity is to 1, the more similar the sentence y and sentence s_i are.

(2) RS GPT-2 Network Model

The GPT-2 network model uses multiple masked self-attention mechanisms and a feed forward neural network to predict the next word based on the input sentences, and then the new words are added as new inputs. Through the process of continuous iteration, the final prediction sentences can be obtained. Since the retrieved sentences are closely related to the contents of external corpora, the contents of retrieved sentences can be used to predict the next word in the external corpora. However, the large number of words in the corpora will affect the training efficiency and prediction accuracy of this model. A novel RS GPT-2 network model is built to overcome this problem, whose structure is shown in Fig. 6.

The RS GPT-2 network model is based on the GPT-2 network model, and introduces a designed random selector. The external corpora are adopted to pre-train the constructed RS GPT-2 network model so that this model can predict the words at each moment for a given input sentence y, and thereby output a sentence of video captioning y'. The out probability of y' is as shown in Eq. (7):

$$p_\delta = \left(y'|c_i, y, F\right) = \prod\nolimits_{\text{corpus}} \sum\nolimits_{f=1}^{m} \left(\left(r|c_i, y\right), \left(\text{random}\left(s_i, F\right)\right)\right) \tag{7}$$

where c_i denotes the i^{th} word in the external corpus, r denotes the correlation coefficient between the words in the corpus and the video captioning y, and $\text{random}\left(\cdot\right)$ is a random function. In the designed random selector, the correlation between a predicted word c_i in the corpus and each word in an input sentence y is firstly calculated using Eq. (8), and the top m ($m = 10$ in this paper) words with higher correlation are determined; Second, to reduce the influence of redundant words in the corpus on the

prediction efficiency of the model, Eq. (9) is used to randomly select one of the words with a high correlation as the predicted word at t moment; Finally, the predicted words for each moment are output continuously as output sentences.

$$r = \frac{\sum_{m=1}^{10}(c_i - w_t)}{\sqrt{\sum_{m=1}^{10}(c_i - w_t)^2}}, w_t \epsilon y \tag{8}$$

$$c_{t-1} = \sum_{j=1}^{10} \text{random}\left(c_j | c_i, y\right) \tag{9}$$

where w_t denotes the word at t moment in a sentence y, and r has a value between 0 and 1. Generally speaking, the closer r is to 1, the stronger the correlation between the two quantities; c_{t-1} conversely, the closer it is to 0, the weaker the correlation between the two quantities; denotes a randomly selected word at the $t-1$ moment, c_j denotes the random selection of the j^{th} word. At this time, c_{t-1} is added to the previously generated word sequence, which becomes the new input for the next step of this model.

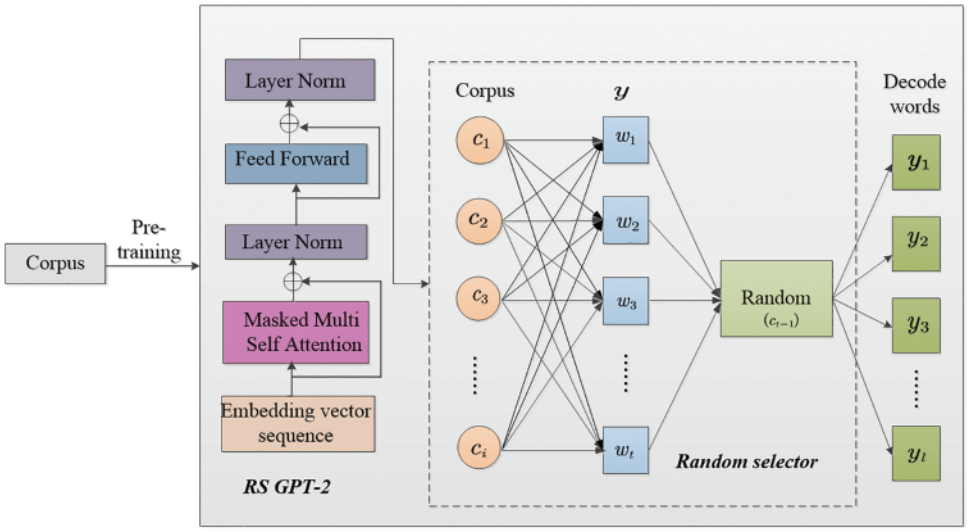


Figure 6: The structure of the RS GPT-2 network model

In addition, to enhance the robustness of the RS GPT-2 network model and avoid the overfitting problem, the q ($q \in k$) similar sentences are randomly selected from the set z to improve the RS GPT-2 network model. Assume that a selected similar sentence is s_i. Input s_i into the LSTM network unit composed of 256 units mentioned in the first section, and extract the semantic feature F in s_i, as shown in Eq. (10):

$$F = \text{LSTM}\left(W_s, s_i, \eta_s\right) \tag{10}$$

where W_s is the learned parameter matrix, and η_s denotes the parameters of the LSTM network model. Then, the output of semantic features by the LSTM network unit are input into the RS GPT-2 generation model to improve the accuracy of word prediction. Therefore, the final predicted word at t moment is expressed as y_t, as shown in Eq. (11):

$$y_t = softmax\left(h_t, c_{t-1}, F, \{y_1, \ldots, y_{t-1}\}\right) \tag{11}$$

where h_t denotes the hidden state at t moment, and $\{y_1, \ldots, y_{t-1}\}$ denotes all predicted words before the moment t. Finally, the words are predicted on the nodes at each moment are connected into a sentence

of video captioning, denoted as y', and $y' = \{y_1, y_2, \ldots, y_l\}$, where l denotes the number of words in the generated sentence.

3.3 Training

Because of the large corpus trained in this paper, it is necessary to avoid the issue of computational redundancy in the batch gradient update process. Since the mini-batch gradient descent (MBGD) optimization algorithm can make full use of the highly optimized matrix operations in the deep learning library to perform efficient gradient calculations, the convergence is relatively stable [43]. The MBGD optimization algorithm is adopted to optimize the RS GPT-2 network model, as shown in Eq. (12):

$$\theta = \theta - \eta \cdot \nabla_\theta J\left(\theta; x^{(i:\, i+n)}; y^{(i:\, i+n)}\right) \tag{12}$$

where θ is a hyper parameter, η is a learning rate, $\nabla_\theta J$ denotes the gradient vector, $x^{(i:\, i+n)}$, $y^{(i:\, i+n)}$ are a sample and its corresponding label, respectively. In addition, the proposed method in this paper performs end-to-end training to minimize the weighted loss sum of textual description sentences generated by an encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model and the textual description sentences generated by the sentence retrieval generation network. The loss function defined in this paper is shown in Eq. (13):

$$L = L_c + \beta L_g \tag{13}$$

where β is a hyperparameter that balances the two terms (set $\beta = 0.6$ in this paper), L_c is the loss based on the encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model, and L_g is the loss of video captioning based on the sentence retrieval generation network. Here the loss L_c is expressed as Eq. (14):

$$L_c = -\sum_{i=1}^{N}\sum_{q=1}^{M} \log\left(y|y_i\right) \tag{14}$$

where y_i denotes a manually annotated sentence, q denotes all the words in the sentence, N denotes the number of generated sentences, M and denotes the number of words. The loss function of the finally generated description sentences is expressed as Eq. (15):

$$L_g = -\sum_{t=1}^{T} \log \sum_{i=1}^{topk} p_\mu\left(y_t|y'\right) \tag{15}$$

where y_t denotes a generated sentence at t moment, and y' denotes the final output sentences of video captioning.

4 Experimental Evaluation

4.1 Datasets and Evaluation Criteria

Extensive experiments on two public datasets named MSVD and MSR-VTT are conducted, respectively. The MSVD dataset is an open-domain video captioning dataset, which contains 1970 videos and is equipped with 80839 English descriptions. The video contents are single-life scenes or actions (such as vegetable cutting, exercise, etc.). In this paper, according to the original division of this dataset, the training set contains 1200 videos, the cross-validation set contains 100 videos, and the test set includes 670 videos. Moreover, the MSR-VTT dataset includes 10,000 video clips of 257 commonly used topics. In this paper, according to the original division of this dataset, the training set consists of 6513 video clips, the validation set has 497 video clips, and there are 2990 videos for testing.

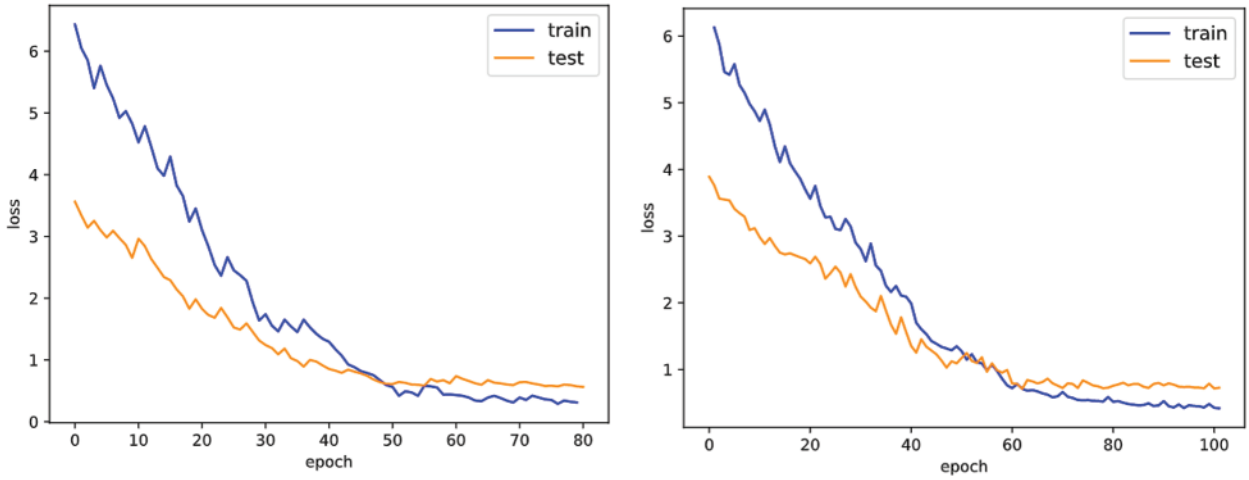
The external corpora for this paper include a Wikipedia corpus and a Brown corpus, where the Wikipedia corpus consists of 2200 articles covering topics such as life and news; the Brown corpus contains 500 texts from different sources covering issues such as news, hobbies, and music.

The experiments adopt Python3.8 programming and Pytorch1.7.1 framework for model training, and the experiments are carried out on a Linux operating system. In the experiments, the GPU is NVIDIA Titan XP, the memory size is 62 GB, the hard disk size is 100 GB, and CUDA11.0 and cuDNN8.0 are used to accelerate the calculation of the proposed method.

In addition, the widely used evaluation metrics of video captioning, namely BLEU-4 [44], Meteor [45], Rouge-L [46], and CIDEr [47] are adopted to evaluate the performances of the proposed method. Here since the consistency between the generated sentences and the actual reference sentences can be better reflected by calculating the number of n-grams overlapping between the generated sentences and one or more reference sentences, this paper also pays attention to the BLEU-4 indicator performance in the experiments.

4.2 Experimental Results and Analyses

The 2200 articles are adopted in the Wikipedia corpus for model training and 600 articles for the test. The changes in loss values in the training and test stages are shown in Fig. 7a, and the training starts to converge smoothly at about 60 times. Since normalization operations are performed, such as removing duplicate blanks and abbreviation correction on the data in the corpus, and the set learning rate is reasonable (lr = 0.001 in the experiments). Furthermore, the dropout regularization method is added to the RS GPT-2 network model, which can optimize this network model, thus avoiding the occurrence of overfitting, and the convergence is better on this dataset. In addition, the 500 texts are adopted in Brown corpus from different sources for model training and 100 texts for model testing, each text containing more than 2000 words. The changes in loss values in the training and test stages are shown in Fig. 7b, and the training starts to converge smoothly at about 50 times. Therefore, overfitting does not appear in the training and testing stages.



(a) The training and test on the Wikipedia corpus. (b) The training and test on the Brown corpus.

Figure 7: The training and test on the Wikipedia and Brown corpora

In addition, Tables 1 and 2 show the comparison results of the proposed method and other video captioning methods based on the encoder-decoder model on the MSVD and MSR-VTT datasets, respectively. It can be seen from Tables 1 and 2 that the proposed method has better performance in the four evaluation indicators on the Wikipedia corpus, and the accuracy of video captioning is higher than other comparison methods. It is considered that the proposed method in this paper can fully extract the visual information of videos by using an encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model, which can effectively guide the generation of sentences description on the video contents. The sentence transformer network can retrieve similar sentences in a corpus to help enrich sentence description. Moreover, the RS GPT-2 network model can effectively generate description sentences that are richer and in line with human expression habits, which further optimizes the description effect of the proposed method. It should be noted that, due to the limited size of both the external knowledge contained in the Brown Corpus and MSVD dataset, an indicator CIDER on the MSVD dataset is degraded. However, the overall indicators are still higher than other comparison methods.

Table 1: Comparison of experimental results of different methods in the MSVD dataset

Methods	BLEU-4	CIDEr	ROUGE L	METEOR
TDCovEDI [48]	53.2	76.4	73.9	33.8
SibNet [24]	54.2	88.2	–	34.8
STG [49]	52.2	93.0	73.9	36.9
CSA-SR [50]	52.2	83.4	72.7	35.6
The proposed method (Wiki)	**55.3**	**94.3**	**74.2**	**38.4**
The proposed method (Brown)	**54.2**	**80.3**	**73.2**	**37.9**

Table 2: Comparison of experimental results of multiple methods in the MSR-VTT dataset

Methods	BLEU-4	CIDEr	ROUGE L	METEOR
SAAT [51]	40.5	49.1	60.9	28.2
ORG [52]	43.6	50.9	62.1	28.8
CoSB [53]	41.4	46.5	61.0	27.8
RCG [37]	42.8	52.9	61.7	29.3
The proposed method (Wiki)	**44.9**	**53.4**	**62.3**	**31.2**
The proposed method (Brown)	**43.8**	**53.1**	**62.9**	**29.8**

To analyze the influence of retrieving different numbers of similar sentences on the generated video captioning, the different numbers of sentences are retrieved in the Wikipedia corpus (because the Wikipedia corpus is large and contains rich information) to compare the results in MSVD and MSR-VTT datasets. The experiment results are shown in Tables 3 and 4. It can be seen that when the number of retrieved sentences is 3, the evaluation criteria are the highest. The reason may be that since the retrieved sentences from corpora do not entirely correspond to the video contents, and noise data and video-related information will be introduced at the same time; therefore, retrieval of a large number of sentences may introduce too much noise, which is not conducive to model training. When

the number of retrieval sentences is 5, the noise is greater than the helpful information. When the number of retrieval sentences is 3, the valuable information is the most. Therefore, the number of retrieved sentences is set to 3.

Table 3: The performance of the model trained with different numbers of retrieval sentences under the MSVD dataset

Retrievals training	BLEU-4	CIDEr	ROUGE_L	METEOR
1	54.2	84.1	72.3	39.1
3	**55.3**	**94.3**	**74.2**	**38.4**
5	53.5	86.9	72.2	37.2

Table 4: The performance of the model trained with different numbers of retrieval sentences under the MSR-VTT dataset

Retrievals training	BLEU-4	CIDEr	ROUGE_L	METEOR
1	42.1	52.3	61.9	29.8
3	**44.9**	**53.4**	**62.3**	**31.2**
5	42.7	51.8	61.4	30.2

To analyze the influence of retrieved sentences from different corpora on video captioning, this paper conducts retrieval training and testing on Wikipedia and Brown corpus and performs experiments on MSVD and MSRVTT datasets. The experimental results are shown in Tables 4 and 5. The method first retrieves and trains the first three sentences on the Wikipedia corpus and the Brown corpus, respectively; then, they are tested on the Wikipedia corpus and the Brown corpus, respectively, and two video datasets simultaneously. It can be seen from Tables 5 and 6 that the evaluation criteria of generated video captioning are the highest when searching in the same corpus and testing at the same time. It is considered that retrieval and training in the same corpus will be guided by the retrieved sentences during the training process. Therefore, when testing in the same corpus, the generated video captioning and retrieved sentences have a relatively close relationship with the corpus.

Table 5: Comparison of experimental results of different corpora retrieval tests under the MSVD dataset

#	Corpus	Retrievals training	Test	BLEU-4	CIDEr	ROUGE_L	METEOR
1	Wiki	√	√	**55.3**	**94.3**	**74.2**	**38.4**
	Wiki	√	×	53.4	89.2	72.1	36.9
	Brown	×	√	49.1	80.9	71.8	37.3
2	Brown	√	√	**54.1**	**93.4**	**74.0**	**37.9**
	Brown	√	×	52.9	87.9	70.8	35.9
	Wiki	×	√	53.1	84.2	73.1	36.2

Table 6: Comparison of experimental results of different corpora retrieval tests under the MSR-VTT dataset

#	Corpus	Retrievals training	Test	BLEU-4	CIDEr	ROUGE_L	METEOR
1	Wiki	√	√	**44.9**	**53.4**	**62.3**	**31.2**
	Wiki	√	×	42.1	51.9	60.7	30.9
	Brown	×	√	43.2	52.7	61.9	30.3
2	Brown	√	√	**43.8**	**53.1**	**62.1**	**29.8**
	Brown	√	×	40.9	50.4	61.2	28.9
	Wiki	×	√	42.1	51.4	60.8	29.2

To analyze the performances of the sentence retrieval generation network model and the encoder-decoder with the hybrid of ResNeXt-ECO and LSTM network models (ResNeXt-ECO-LSTM) in the proposed method, the ResNeXt-ECO-LSTM, ED-SRG (GPT-2), and ED-SRG (RS GPT-2) network models are compared in the two datasets as mentioned above, where ED-SRG (GPT-2) denotes the sentence retrieval generation network model containing the existing GPT-2 network, and ED-SRG (RS GPT-2) represents the sentence retrieval generation network model containing a novel RS GPT-2 network. It should be noted here that the ResNeXt-ECO-LSTM network model is used for the first stage of sentence generating for the ED-SRG (GPT-2) and ED-SRG (RS GPT-2) network models, respectively, to compare and analyze the performance advantages of the ED-SRG (RS GPT-2) model in generated video captioning. The experimental results are shown in Tables 7 and 8. The generated video captioning using the ED-SRG (GPT-2) and ED-SRG (RS GPT-2) is better compared to ResNeXt-ECO-LSTM. The reason might be the introduction of external corpus information, which expands the content description of video captioning generated by the existing encoder-decoder. Furthermore, the BLEU-4, CIDEr, ROUGE_ L, and METER indicators are used to measure the performances of the ED-SRG (RS GPT-2) and ED-SRG models by matching the generated video captioning with the reference sentences. It can be seen from Tables 7 and 8 that the performance of the ED-SRG (RS GPT-2) model is higher than that of the ED-SRG model. It is considered that the existing GPT-2 network model is only trained using the external corpora and lacks the word computation related to video content. RS GPT-2 constructed in this paper can calculate the correlation between the predicted words and similar sentences in the external corpora by designing a random selector. By selecting the words with high correlation as the predicted words, the content of video captioning can be described more accurately. In Table 7, the CIDEr performance of the ResNeXt-ECO-LSTM is lower than the other two models. It is considered that because of the limited size of the MSVD dataset, the video captioning generated only by an encoder-decoder model is relatively simple. The similarities are low between the generated simple description sentences and candidate sentences. Guided by the sentence retrieval generation network model, the contents of video captioning can be effectively expanded, and the similarities between generated simple description sentences and candidate sentences can be improved; thus, the CIDEr performance can be enhanced more substantially.

In addition, different learning rates are set for training on the two corpora to analyze the impact on model training. The results are shown in Fig. 8. When the learning rate is 0.001, the model converges well, and there is no overfitting issue. When the learning rate is 0.005, the gradient oscillates back and forth between the minimum values; when the learning rate is 0.0001, the convergence process is slow.

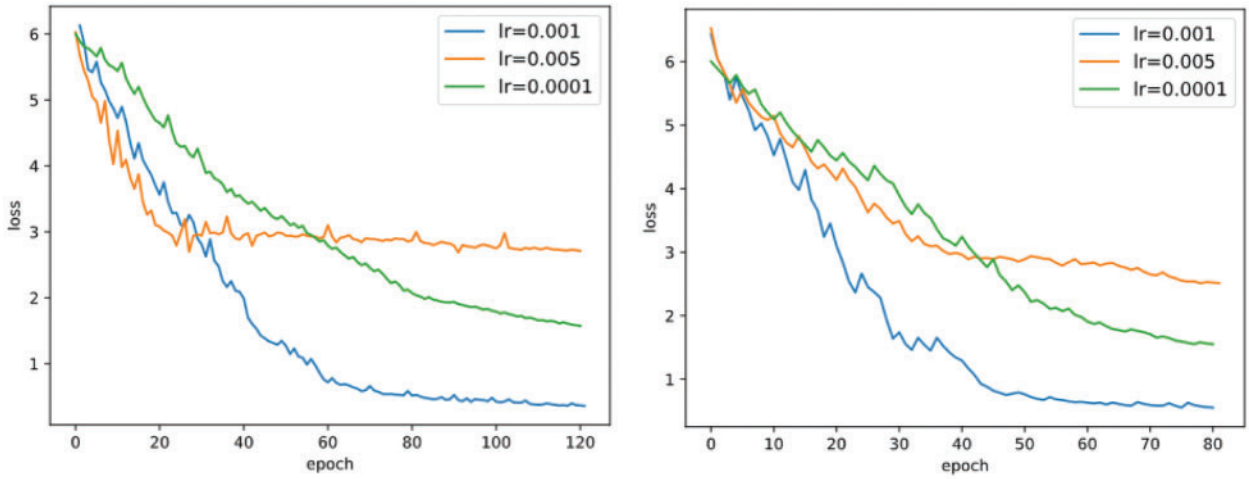
The lower the learning rate, the slower the convergence process. Therefore, the learning rates in the training and testing stage are both set to 0.001 in this paper.

Table 7: Comparison of experimental results of different network models under the MSVD dataset and Wiki corpus

Models	BLEU-4	CIDEr	ROUGE_L	METEOR
ResNeXt-ECO-LSTM	47.5	57.9	66.7	28.6
ED-RSG (GPT-2)	51.2	87.5	69.8	35.4
ED-RSG (RS GPT-2)	**55.3**	**94.3**	**74.2**	**38.4**

Table 8: Comparison of experimental results of different network models under the MSR-VTT dataset and Wiki corpus

Models	BLEU-4	CIDEr	ROUGE_L	METEOR
ResNeXt-ECO-LSTM	37.6	47.2	55.8	27.5
ED-RSG (GPT-2)	40.9	51.2	59.8	30.1
ED-RSG (RS GPT-2)	**44.9**	**53.4**	**62.3**	**31.2**



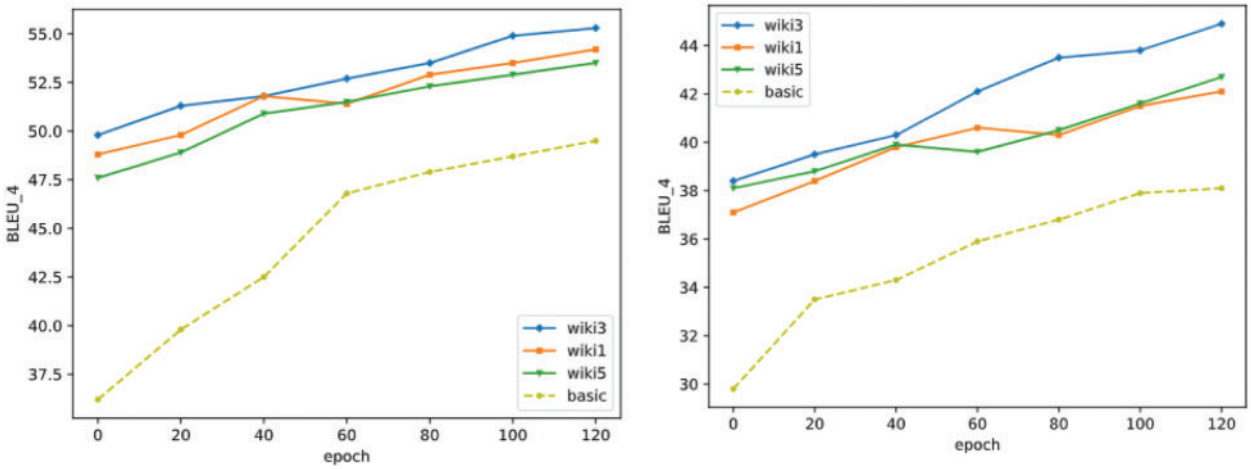
(a) The result of different learning rates based on the Wikipedia corpus

(b) The result of different learning rates based on the Brown corpus.

Figure 8: The results of different learning rates

Since the BLEU-4 evaluation metric can be used to evaluate the quality of generated sentences to video content, the BLEU-4 is adopted as a performance criterion for video captioning. In this paper, the different numbers of sentences in the Wikipedia corpus are retrieved to participate in model training. The results are shown in Fig. 9, where "basic" represents the BLEU-4 score of sentences generated based on the encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model, and

the rest respectively represent the retrieval of different numbers of sentences in the Wikipedia corpus. When the number of retrieved sentences is 3, the BLEU-4 evaluation metric of the video captioning is the highest, which is to score the description sentences according to the n-gram matching between the generated sentences and the reference sentences. When the number of retrieved sentences is 3, more information related to video content is introduced with fewer noises, and the sentences have grammatical matching.



(a) The result of MSVD set by using the BLEU-4 evaluation metric.

(b) The result of MSR-VTT set by using the BLEU-4 evaluation metric.

Figure 9: The results of video captioning with different numbers of sentences

With text descriptions of the videos on the MSVD and MSR-VTT datasets, the obtained description sentences are shown in Figs. 10 and 11. By similar sentences retrieved in the corpora, more helpful information can be obtained to guide the generation of semantically richer video captioning. In Fig. 10, "Baseline" denotes the generated video captioning based on the encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model. This encoder-decoder is utilized to search the Wikipedia corpus and select the three sentences with the highest similarity. It can be seen that they are all related to "ride a bicycle." and each gets a similarity score. Then, the retrieved sentences are trained in the RS GPT-2 generator using Wikipedia corpus, and the "Baseline" sentences can be extended. For example, information "on the beach" and "she is relaxed." can be obtained. Finally, a more semantically rich video captioning is generated.

Figure 10: The video captioning on the MSVD dataset

Figure 11: The video captioning on the MSR-VTT dataset

In addition, as shown in Fig. 11, "Baseline" represents the video captioning sentence generated based on the encoder-decoder with the hybrid of ResNeXt-ECO and LSTM model. This encoder-decoder is utilized to search the Wikipedia corpus and select the three sentences with the highest similarity. It can be seen that they are all related to "sing" and "stage" semantics, and their similarity scores are calculated. Then, the retrieved sentences are trained in the RS GPT-2 generator using the Wikipedia corpus, and the "Baseline" sentences can be extended. For example, "children", "with a microphone", and "in front of the judges" information can be obtained. Finally, a more semantically rich video captioning is generated.

Baseline (Fig. 10)	A girl is riding on a bicycle.
Retrieve sentences (Fig. 10)	A girl enjoys riding her bicycle. (Similarity score: 0.6981) The child is riding his bicycle on the beach. (Similarity score: 0.6499) Some is riding bike with her friends and they are happy. (Similarity score: 0.5622)
ED-SRG (Fig. 10)	The girl is riding her bicycle on the beach and she is relaxed.
Baseline (Fig. 11)	Three kids sing on the stage.
Retrieve sentences (Fig. 11)	The young children are singing on the stage. (Similarity score: 0.7638) Children like to sing with a microphone on the stage. (Similarity score: 0.7109) Some teenagers are singing on stage and sitting on stools. (Similarity score: 0.6324)
ED-SRG (Fig. 11)	The children enjoy singing with a microphone on stage in front of the judges.

5 Conclusion

In this paper, a novel video captioning method guided by a sentence retrieval generation network (ED-SRG) is proposed, which can efficiently retrieve similar sentences of video contents from external corpora. It takes the retrieved sentences as the guidance to generate the final video captioning by constructing a novel RS GPT-2 network. The experiment results on the MSVD and MSR-VTT datasets demonstrate the advantages of the proposed method. It is considered that video captioning should not be limited to visual information, and effective external corpora can guide and expand the generation of video captioning. However, the proposed method in this paper uses two public corpora,

and does not introduce a specialized domain corpus for specific scenarios. It has limitations for the generation of video captioning in the specialized domain. In the future, the applications of proprietary corpora on video captioning will be further explored. In addition, how to introduce more efficient matching methods for corpus will be studied to improve the efficiency of similar sentence retrieval.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grants 62273272 and 61873277, in part by the Chinese Postdoctoral Science Foundation under Grant 2020M673446, in part by the Key Research and Development Program of Shaanxi Province under Grant 2023-YBGY-243, and in part by the Youth Innovation Team of Shaanxi Universities.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, "Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap," *IEEE Access*, vol. 8, pp. 218386–218400, 2020.

[2] Y. X. Peng, J. W. Qi and X. Huang, "Current research status and prospects on multimedia content understanding," *Journal of Computer Research and Development*, vol. 56, no. 1, pp. 183–208, 2019.

[3] D. S. Guo, L. Wei and X. Z. Fang, "Capturing temporal structures for video captioning by spatio-temporal contexts and channel attention mechanism," *Neural Processing Letters*, vol. 46, no. 1, pp. 313–328, 2017.

[4] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proc. of 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 12487–12496, 2019.

[5] J. Dong, K. Gao, X. Chen and J. Cao, "Refocused attention: Long short-term rewards guided video captioning," *Neural Processing Letters*, vol. 52, no. 2, pp. 935–948, 2020.

[6] S. Chen, X. Zhong, L. Li, W. Liu, C. Gu *et al.,* "Adaptively converting auxiliary attributes and textual embedding for video captioning based on BiLSTM," *Neural Processing Letters*, vol. 52, no. 3, pp. 2353–2369, 2020.

[7] D. Liu, X. Qu, X. Liu, J. Dong, P. Zhou *et al.,* "Jointly cross-and self-modal graph attention network for query-based moment localization," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, Seattle WA, USA, pp. 4070–4078, 2020.

[8] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng *et al.,* "Context-aware biaffine localizing network for temporal sentence grounding," in *Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 11235–11244, 2021.

[9] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li *et al.,* "Task-driven dynamic fusion: Reducing ambiguity in video description," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 3713–3721, 2017.

[10] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li *et al.,* "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[11] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney *et al.,* "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Sydney, NSW, Australia, pp. 2712–2719, 2013.

[12] F. Nian, T. Li, Y. Wang, X. Wu, B. Ni *et al.,* "Learning explicit video attributes from mid-level representation for video captioning," *Computer Vision and Image Understanding*, vol. 163, no. 2, pp. 126–138, 2017.

[13] W. Li, D. Guo and X. Z. Fang, "Multimodal architecture for video captioning with memory networks and an attention mechanism," *Pattern Recognition Letters*, vol. 105, no. 12, pp. 23–29, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6000–6010, 2017.
[15] C. Sun, A. Myers, C. Vondrick, K. Murphy and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 7464–7473, 2019.
[16] C. Deng, S. Chen, D. Chen, Y. He and Q. Wu, "Sketch, ground, and refine: Top-down dense video captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 234–243, 2021.
[17] J. Song, Y. Yang, Y. Yang, Z. Huang and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. of the 2013 ACM SIGMOD Int. Conf. on Management of Data*, New York, NY, USA, pp. 785–796, 2013.
[18] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," in *Proc. of the Twenty-Seventh AAAI Conf. on Artificial Intelligence*, Bellevue, WA, USA, pp. 541–547, 2013.
[19] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li *et al.,* "Open-book video captioning with retrieve-copy-generate network," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 9837–9846, 2021.
[20] V. Ordonez, G. Kulkarni and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. of the 24th Int. Conf. on Neural Information Processing Systems*, Granada, Spain, pp. 1143–1151, 2011.
[21] A. Gupta, P. Srinivasan, J. Shi and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. of 2009 IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, Florida, USA, pp. 2012–2019, 2009.
[22] P. H. Seo, A. Nagrani, A. Arnab and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 17959–17968, 2022.
[23] T. Deb, A. Sadmanee, K. K. Bhaumik, A. A. Ali, M. A. Amin *et al.,* "Variational stacked local attention networks for diverse video captioning," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 4070–4079, 2022.
[24] S. Liu, Z. Ren and J. Yuan, "SibNet: Sibling convolutional encoder for video captioning," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 9, pp. 259–3272, 2021.
[25] W. Zhao, X. Wu and X. Zhang, "MemCap: Memorizing style knowledge for image captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, NY, USA, pp. 12984–12992, 2020.
[26] H. Ahmad, H. U. Khan, S. Ali, S. I. U. Rahman, F. Wahid *et al.,* "Effective video summarization approach based on visual attention," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1427–1442, 2022.
[27] A. Cherian, J. Wang, C. Hori and T. Marks, "Spatio-temporal ranked-attention networks for video captioning," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Snowmass Village, CO, USA, pp. 1617–1626, 2020.
[28] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 8327–8336, 2019.
[29] S. Cao, W. M. Liu, G. Y. An and Q. Q. Ruan, "Joint common sense and relation reasoning for dense relational captioning," in *Proc. of the 15th IEEE Int. Conf. on Signal Processing*, Beijing, China, pp. 156–159, 2020.
[30] J. Hou, X. Wu, W. Zhao, J. Luo and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 8918–8927, 2019.
[31] Y. Li, X. Cui and X. Jin, "Research on video captioning method based on semantic key frame," in *Proc. of the 2022 2nd Asia-Pacific Conf. on Communications Technology and Computer Science*, Shenyang, China, pp. 38–44, 2022.

[32] F. Liu, X. Wu, C. You, S. Ge, Y. Zou *et al.,* "Aligning source visual and target language domains for unpaired video captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9255–9268, 2022.
[33] L. Yan, S. Ma, Q. Wang, Y. Chen, X. Zhang *et al.,* "Video captioning using global-local representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6642–6656, 2022.
[34] R. Ramos and B. Martins, "Using neural encoder-decoder models with continuous outputs for remote sensing image captioning," *IEEE Access*, vol. 10, pp. 24852–24863, 2022.
[35] S. Aditya, Y. Yang, C. Baral, Y. Aloimonos and C. Fermuller, "Image understanding using vision and reasoning through scene description graph," *Computer Vision and Image Understanding*, vol. 173, no. 4, pp. 33–45, 2018.
[36] Y. Zhou, Y. Sun and V. Honavar, "Improving image captioning by leveraging knowledge graphs," in *Proc. of the Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 283–293, 2019.
[37] J. Hou, X. Wu, X. Zhang, Y. Qi, Y. Jia *et al.,* "Joint commonsense and relation reasoning for image and video captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, NY, USA, pp. 10973–10980, 2020.
[38] S. K. Dash, S. Acharya, P. Pakray, R. Das and A. Gelbukh, "Topic-based image caption generation," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3025–3034, 2020.
[39] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1492–1500, 2017.
[40] M. Zolfaghari, K. Singh and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 695–712, 2018.
[41] Z. Abedjan, C. G. Akcora, M. Ouzzani, P. Papotti and M. Stonebraker, "Temporal rules discovery for web data cleaning," *Proceedings of the VLDB Endowment*, vol. 9, no. 4, pp. 336–347, 2015.
[42] X. Zheng, C. Zhang and P. C. Woodland, "Adapting GPT, GPT-2 and BERT language models for speech recognition," in *Proc. of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, pp. 162–168, 2021.
[43] M. Munsif, H. Afridi, M. Ullah, S. D. Khan, F. A. Cheikh *et al.,* "A lightweight convolution neural network for automatic disasters recognition," in *Proc. of the 10th European Workshop on Visual Information Processing*, Budapest, Hungary, pp. 1–6, 2022.
[44] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, pp. 311–318, 2002.
[45] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, USA, pp. 65–72, 2005.
[46] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. of Workshop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81, 2004.
[47] R. Vedantam, C. L. Zitnick and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 4566–4575, 2015.
[48] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao *et al.,* "Temporal deformable convolutional encoder-decoder networks for video captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Honolulu, Hawaii, USA, pp. 8167–8174, 2019.
[49] J. Zhang and Y. Peng, "Video captioning with object-aware spatio-temporal correlation and aggregation," *IEEE Transactions on Image Processing*, vol. 29, no. 99, pp. 6209–6222, 2020.
[50] Z. Lei and Y. Huang, "Video captioning based on channel soft attention and semantic reconstructor," *Future Internet*, vol. 13, no. 2, pp. 1–18, 2021.
[51] Q. Zheng, C. Wang and D. Tao, "Syntax-aware action targeting for video captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 13096–13105, 2020.

[52] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang *et al.,* "Object relational graph with teacher-recommended learning for video captioning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 13278–13288, 2020.

[53] J. Vaidya, A. Subramaniam and A. Mittal, "Co-segmentation aided two-stream architecture for video captioning," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 2774–2784, 2022.