Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 7088837, 11 pages
https://doi.org/10.1155/2021/7088837

Research Article

Research on Volleyball Video Intelligent Description Technology Combining the Long-Term and Short-Term Memory Network and Attention Mechanism

Yuhua Gao,[1] **Yong Mo**,[2] **Heng Zhang**,[3] **Ruiyin Huang**,[1] **and Zilong Chen**[1]

[1]*Guangzhou Sport University, Guangzhou, Guangdong 510500, China*
[2]*Guangdong Baiyun University, Guangzhou, Guangdong 510450, China*
[3]*Yingshan County No. 1 Middle School, Hubei, Yingshan 438700, China*

Correspondence should be addressed to Yuhua Gao; 11071@gzsport.edu.cn

Received 27 August 2021; Revised 25 September 2021; Accepted 27 September 2021; Published 14 October 2021

Academic Editor: Bai Yuan Ding

With the development of computer technology, video description, which combines the key technologies in the field of natural language processing and computer vision, has attracted more and more researchers' attention. Among them, how to objectively and efficiently describe high-speed and detailed sports videos is the key to the development of the video description field. In view of the problems of sentence errors and loss of visual information in the generation of the video description text due to the lack of language learning information in the existing video description methods, a multihead model combining the long-term and short-term memory network and attention mechanism is proposed for the intelligent description of the volleyball video. Through the introduction of the attention mechanism, the model pays much attention to the significant areas in the video when generating sentences. Through the comparative experiment with different models, the results show that the model with the attention mechanism can effectively solve the loss of visual information. Compared with the LSTM and base model, the multihead model proposed in this paper, which combines the long-term and short-term memory network and attention mechanism, has higher scores in all evaluation indexes and significantly improved the quality of the intelligent text description of the volleyball video.

1. Introduction

With the continuous development of big data, computer computing power, and machine learning model, video description technology has set off a research upsurge again. Video description technology is an interdisciplinary research problem. It is an exploration of the expansion of deep learning technology to the field of multidata after making outstanding achievements in the fields of natural language processing, speech recognition, and computer vision [1]. It can be widely used in video retrieval, intelligent security, human-computer interaction, virtual reality, and helping the blind understand films and videos. It has high application value and practical significance. Among all kinds of multimedia data, video has become an important carrier of information dissemination in today's society because of its large amount of information and rich content [2]. With the rapid development of video sensors, we can easily collect a large number of complex video data, and how to use natural language to describe the stored information has become an urgent problem to be solved. The task of using natural language to describe a video is very simple for normal people, but it is a very difficult task for computers. It requires that the proposed method can span the semantic gap from low-level pixel features to high-level language. The existence of the semantic gap brings great difficulties for the computer to automatically describe the video. The existing video description is usually carried out by manually labeling video data. This method is inefficient and often subjective, and it is easy to ignore many details [3]. Therefore, it is of great practical significance to find an efficient and objective way to describe the video to help people retrieve the video more quickly and conveniently.

With the rapid development of deep learning, researchers began to apply this technology to video description. Current research studies generally use the convolutional neural network (CNN) structure as the encoder to extract visual information and the long short-term memory (LSTM) network structure as the decoder to predict the description sentence [4]. Although these methods avoid the subjectivity of manual annotation to a certain extent, due to the lack of depth of language learning information and less grammatical supervision when generating description sentences, the predicted description sentences will have sentence errors such as missing predicates and loss of visual information. At the same time, sports video occupies an important position in the field of video description because of its huge audience. In particular, volleyball videos often present high-speed and detailed characteristics, which increase the difficulty of understanding the intelligent description of visual targets of video sensors [5]. Therefore, a video sensor processing method combining the long-term and short-term memory network and attention mechanism is proposed for the intelligent description of the volleyball video. The introduction of the attention mechanism can make the model pay much attention to the significant areas in the image/video when generating sentences, quickly identify the target, and effectively solve the loss of visual information.

Aiming at the problems of the lack of visual information, syntax error, and strong subjectivity in video description methods in existing video sensors, this paper proposes a method combining the long-term and short-term memory network and attention mechanism to describe the volleyball video. In the first section, the research background and significance of video description are briefly described. The second section briefly describes the research status of the video description of video sensors, discusses the problems to be solved in the current video description methods, and makes a general introduction to the research work and research methods of this paper. The third section first introduces the long-term and short-term memory network and attention mechanism and then gives the application in volleyball video description combined with the long-term and short-term memory network and attention mechanism model. In the fourth section, the datasets for training and testing are selected, and the evaluation indexes of the model recognition effect are determined. Then, a series of control experiments are set up to test the effectiveness of the attention mechanism model combined with the long-term and short-term memory network in the field of video description. The fifth section briefly summarizes the main conclusions of the article.

2. The Related Works

Thanks to the research and development of sensor technology, embedded technology, machine translation, image description, and the expansion of annotated video datasets in recent years, the video description task in video sensors has also attracted extensive attention of researchers, and the research of video description methods has also made great progress [6].

Early video description methods mainly generated sentences based on predefined templates. The sentences describing the video were first divided into several parts, each part should be aligned with the visual content, and then the words detected from the vision were filled into the predefined templates. Kojima et al. selected the most appropriate verbs and objects by detecting the human posture; then, the content displayed by the action semantics is corresponding to the features extracted from the video image, and finally filled the detected syntactic components into common case templates [7]. Rohrbach et al. first generated a rich semantic representation of visual content. They simulated the relationship between different components of the visual input by learning a conditional random field (CRF). Finally, they expressed the generation of natural language as a machine translation problem [8]. Thomason et al. obtained the confidence of the target, action, and scene in the video through the visual recognition system and estimated the most likely subject, verb, object, and place with the factor graph model (FGM) [9]. However, these methods rely too much on predefined templates and detected visual elements and can only simply describe the video, lacking the ability to express semantics.

With the development of the convolutional neural network in the image classification task, three-dimensional convolutional neural network in the video analysis task, and cyclic neural network in the machine translation task, many researchers apply the deep neural network to the video description task. Donahue et al. proposed the long-term recurrent convolutional network (LRCN) model, which can directly generate word sequences through the cyclic neural network without considering the syntax problem of generating description statements [10]. S. Venugopalan et al. proposed a video description model based on LSTM, but this method only considers the characteristics of video frames and ignores the dynamics and continuity of the video [11]. S. Venugopalan et al. proposed a two-stage video description framework, which is composed of a multichannel video encoder and a language decoder that generates sentences. The encoded features are combined by using the fusion layer, and the obtained features are input into the language decoder into a series of words [12]. C. Zhang and Tian et al. proposed a long-term and short-term memory network with visual semantic embedding, which can explore the embedding of learning LSTM and visual semantics [13]. The method proposed by Yao et al. considers the local action features of the video when generating the video description, uses the three-bit convolutional neural network to extract the features of the video clip as the local action features of the video, uses the two-dimensional convolutional neural network to extract the appearance features of the video, and combines the temporal attention (TA) to explore the global time structure of the video [14, 15]. These video description methods only consider visual features and ignore the rich semantic information in the video. Semantic concepts are highly related to the visual content and are widely used in visual recognition tasks.

To sum up, although the research on video description methods has made good achievements, there is still much room for improvement in video feature extraction, video

timing features, and video multilingual text description. In view of this, this paper describes the volleyball video by combining the long-term and short-term memory network and attention mechanism. By paying attention to the significant areas in the video, the model can quickly identify the target, effectively solve the lack of visual information, and has a good video description effect when the visual sensor processes the volleyball video.

3. Volleyball Video Description Model Based on the Long-Term and Short-Term Memory Network and Attention Mechanism

3.1. Long-Term and Short-Term Memory Network. As an improved structure of the ordinary cyclic neural network, long-term and short-term memory network (LSTM) can deal with variable input and output sequences and can effectively avoid the problem of gradient disappearance [16]. The LSTM unit outputs the hidden state h_t of step t by relying on the input x_t of the current step t and the hidden state h_{t-1} of time $t-1$ of the previous step. In the LSTM unit, the flow of input information of the current step and historical memory information is controlled by the input gating and forgetting gating unit. The calculation method is as follows:

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right), \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right), \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right), \\ g_t &= f\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot \Phi\left(c_t\right). \end{aligned} \tag{1}$$

In the formula, σ is the sigmoid activation function, Φ is the tanh activation function, $\odot$ represents the point multiplication operation of the vector, and the weight matrix W_{ij} and offset vector b_j are trainable parameters.

Aiming at the problem of automatic generation of video description, based on the LSTM cyclic neural network, by predicting the feature sequence $(x_1, x_2, \ldots, x_n)$ of a given input video, the conditional probability of the output word sequence $(y_1, y_2, \ldots, y_m)$ is

$$p(y_1, y_2, \ldots, y_m \mid x_1, x_2, \ldots, x_n). \tag{2}$$

The LSTM model is based on the encoder-decoder framework, and its structure is shown in Figure 1. In the encoding phase, the LSTM layer of the encoder uses the input sequence $X(x_1, x_2, \ldots, x_n)$ to calculate the intermediate hidden state $(h_1, h_2, \ldots, h_n)$ [17]. In the decoding stage, the conditional probability is predicted through the LSTM layer and softmax output layer of the encoder. By linking the probability of each step, the conditional probability of the given input sequence X and the output sequence Y is obtained as follows:

$$p(y_1, \ldots, y_m \mid x_1, \ldots, x_n) = \prod_{t=1}^{m} p\left(y_t \mid h_{n+t-1}, y_{t-1}\right). \tag{3}$$

In the model training stage, the parameters of the model are updated by maximizing the log likelihood probability, i.e.,

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{m} \log p\left(y_t \mid h_{n+t-1}, y_{t-1}; \theta\right). \tag{4}$$

θ represents the parameters of the model, and the optimization method adopts the random gradient descent method.

3.2. Attention Mechanism. The encoder-decoder framework combined with the attention mechanism can learn automatic alignment and translation in the training process of the model. When generating new target short language words, it can find the location of relevant source words, and then the decoder combines the content vector obtained from these locations and the generated target words to predict the target words to be generated [18]. The biggest difference between this method combining the attention mechanism and the basic encoder-decoder method is that it does not need to encode the whole input sentence into a single fixed-length vector, but encodes the input sentence into a vector sequence and dynamically selects a subset of the vector sequence to form a new content vector at each step of the decoding process to generate words at the target end [19]. The calculation method of the dynamic content vector combined with the attention mechanism is shown in Figure 2.

For step i of the decoding process, the content vector c_i is weighted by the hidden state sequence $(h_1, h_2, \ldots, h_T)$ output by the encoder and the attention weight a_{ij}:

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j. \tag{5}$$

The calculation method of attention weight a_{ij} for hidden state h_j is as follows:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T} \exp\left(e_{ik}\right)}. \tag{6}$$

e_{ij} here is calculated by a feedforward neural network model for automatic alignment:

$$e_{ij} = a\left(\mathbf{s}_{i-1}, \mathbf{h}_j\right). \tag{7}$$

In the formula, $\mathbf{s}_{i-1}$ is the hidden state at the time of decoder $t-1$, and the parameters of the $a(q,k)$ model and other parameters of the translation model are updated through the training process.

3.3. Volleyball Video Description Model Combining the Long-Term and Short-Term Memory Network and Attention Mechanism. In the task of volleyball video intelligent description, convolutional neural network is usually used to extract image features, and LSTM is used to extract the content vector. The representation ability of the content vector obtained by this method is limited. The attention mechanism can selectively focus on the subset of the video frame sequence to produce the word description of the object or action in the subset of the corresponding frame sequence. Different from the traditional model, the video

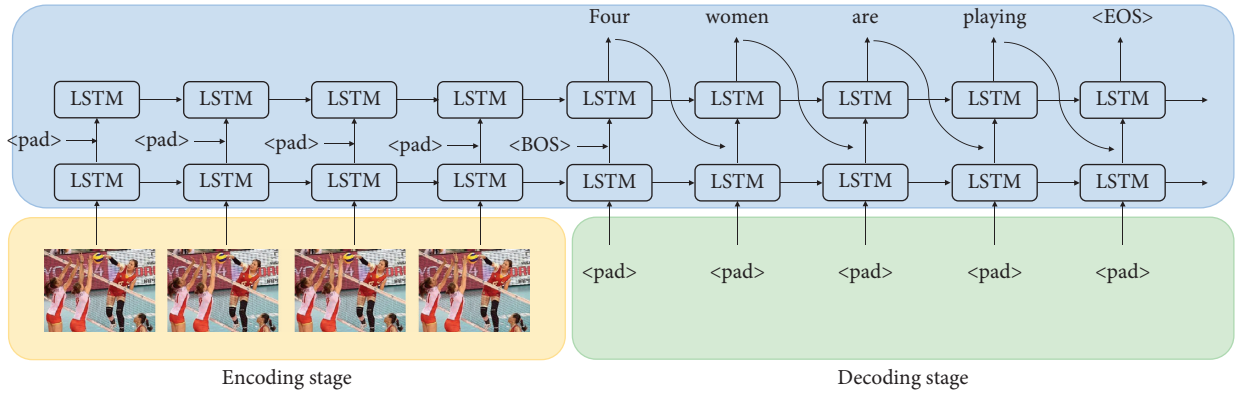

Figure 1: Network video description model based on long-term and short-term memory.

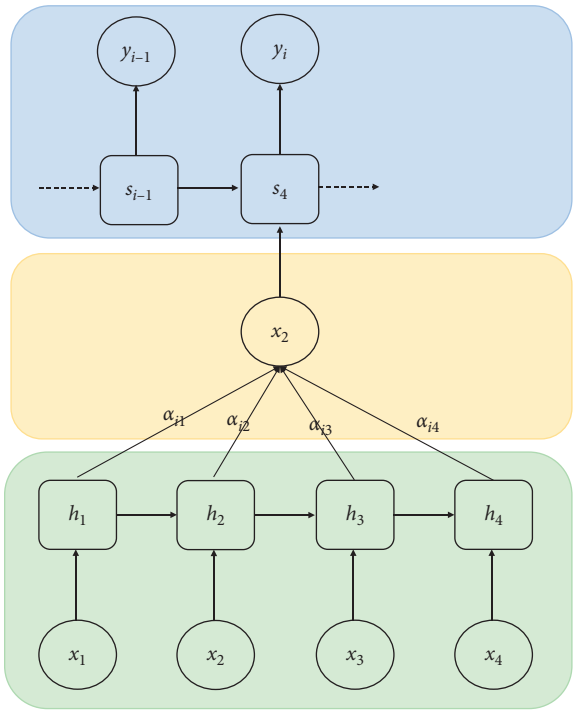

Figure 2: Calculation diagram of the attention mechanism.

intelligent description model combined with the long-term and short-term memory network and attention mechanism can dynamically adjust the context vector output by the encoder and realize the function of automatic soft alignment by replacing the convolutional layer and cyclic neural unit layer with the self-attention layer [20]. Its frame is shown in Figure 3.

As can be seen from Figure 3, the video intelligent description model combined with the long-term and short-term memory network and attention mechanism is based on the encoder-decoder framework, which is mainly composed of the encoder, decoder, feature extraction layer at the bottom, embedding layer, linear layer, and softmax layer at the top.

The visual feature extraction layer uses f_{2dCNN} to represent the visual feature extraction function; then, the sequential multiframe input of a given video is

$$I = (I_1, I_2, \ldots, I_T), \quad I_t \in \mathbb{R}^{h \times w \times c}. \tag{8}$$

In the formula, h, w, and c are the height, width, and number of channels of the image, and T is the sequence length. The visual features are extracted for each frame, respectively:

$$x_t = f_{2dCNN}(I_t). \tag{9}$$

The visual feature sequence of consecutive frames can be obtained:

$$X = (x_1, x_2, \ldots, x_T), x_t \in R^{d_{\text{feat}}}. \tag{10}$$

In the formula, d_{feat} is the characteristic dimension. After the visual feature extraction layer, the linear embedding layer is introduced to map the high-dimensional features to the vector of appropriate dimensions for the calculation of the encoder. The calculation method of the embedded layer is

$$x_t^{\text{emb}} = W_{\text{img}} x_t + b_{\text{img}}, \tag{11}$$

and $X^{\text{emb}} \in \mathbb{R}^{T \times d_{\text{model}}}$ is obtained, where d_{model} is the vector dimension of the query, key, and value in the process of calculating self-attention weight. The calculation method of the frame position information coding layer is as follows:

$$X^{\text{enc}} = X^{\text{emb}} + W_{\text{PE}}. \tag{12}$$

Here is the encoded sequence position information, which can be obtained by artificially setting rules and fixed conversion functions. The position information constructor in this paper is

$$\begin{aligned} W_{\text{PE}}(t, 2i) &= \sin\left(t/10000^{2i/d_{\text{modd}}}\right), \\ W_{\text{PE}}(t, 2i+1) &= \cos\left(t/10000^{2i/d_{\text{modd}}}\right). \end{aligned} \tag{13}$$

The trigonometric functions here have different frequencies for features in the same position and different dimensions; for features in different positions of the same dimension, their phases are different. The reason for using the trigonometric function is that the characteristics of the relative position can be described by linear transformation, so it can express the information of the relative position to a

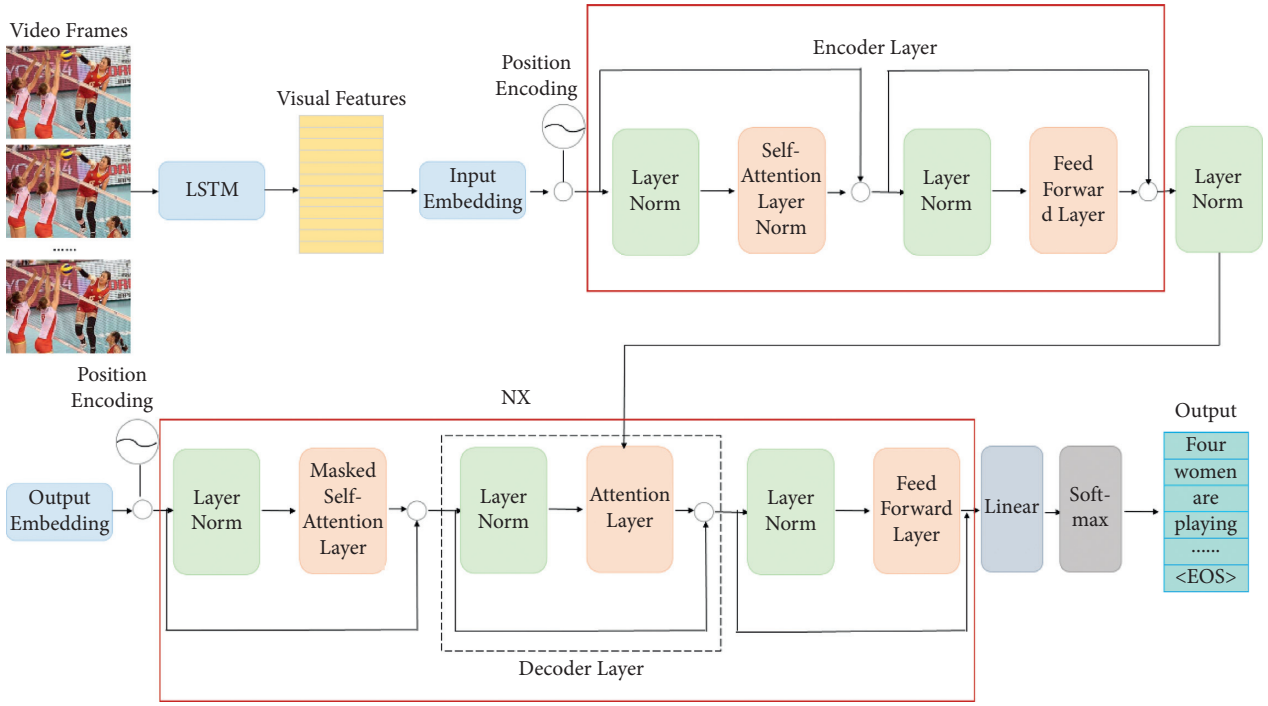

Figure 3: Video intelligent description model combining the long short-term memory network and attention mechanism.

certain extent, and trigonometric functions with different frequencies introduce diversified expression of position information.

In the model, the self-attention module adopts the multihead attention mechanism. Compared with the dot-product attention, the feature expression ability of this mechanism is more diverse, and its calculation process is [21]

$$\begin{aligned} \text{head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \\ \text{MultiHead}(Q,K,V) &= \text{Concat}_{i=1,\ldots,h}\left(\text{head}_i\right)W^O. \end{aligned} \tag{14}$$

In the formula, h is the number of "heads" in multiple heads, and $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_q}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are trainable parameters. The self-attention module mainly includes normalization, self-attention layer, and residual connection. The forward calculation process of the self-attention module of layer l can be expressed as follows:

$$\begin{aligned} Q^{(l)} &= \text{LayerNorm}\left(X^{(l-1)}\right)W_{Q_e}^{(l)}, \\ K^{(l)} &= \text{LayerNorm}\left(X^{(l-1)}\right)W_{K_e}^{(l)}, \\ V^{(l)} &= \text{LayerNorm}\left(X^{(l-1)}\right)W_{V_e}^{(l)}, \\ f_{\text{self-att}}\left(X^{(l-1)}\right) &= X^{(l-1)} + \text{MultiHead}\left(Q^{(l)}, K^{(l)}, V^{(l)}\right), \end{aligned} \tag{15}$$

where $W_{Q_e}^{(l)} \in \mathbb{R}^{d_{\text{model}} \times d_q}, W_{K_e}^{(l)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_{V_e}^{(l)} \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are trainable parameters that transform the output of the previous layer into a triple of query, key, and value. LayerNorm represents the normalization function. Layer normalization normalizes the characteristic scale. Combined with the scaling operation in the point product attention mechanism, the numerical value of the whole calculation process is more stable, and the convergence speed is faster during training.

4. Research on the Video Description Effect Combining the Long-Term and Short-Term Memory Network and Attention Mechanism

4.1. Datasets and Evaluation Indicators. The experiment selects two commonly used datasets for video description generation to verify the effectiveness of the model; they are MSVD and MSR-VTT datasets.

Microsoft Research Video Description (MSVD): this dataset contains 1970 video clips. Each video clip describes a single activity, with a duration of 10 s to 25 s and an average length of about 9 s [22]. This paper selects the first 1200 video clips of the dataset as the training set, the next 100 clips as the verification set, and the remaining 670 clips as the test set.

Microsoft Research Video to Text (MSR-VTT): the dataset contains 10000 video clips and 20 video types [23]. Using the public dataset division method, 6513 video clips are selected as the training set, 497 clips as the verification set, and 2990 clips as the test set.

In order to objectively represent the quality of text description generated by the algorithm, this paper selects four different objective evaluation methods to test the performance of the algorithm, which are BLEU@4, ROUGE-L, METEOR, and CIDEr. To measure the proximity between the generated description text and the manual description text, the ROUGE-

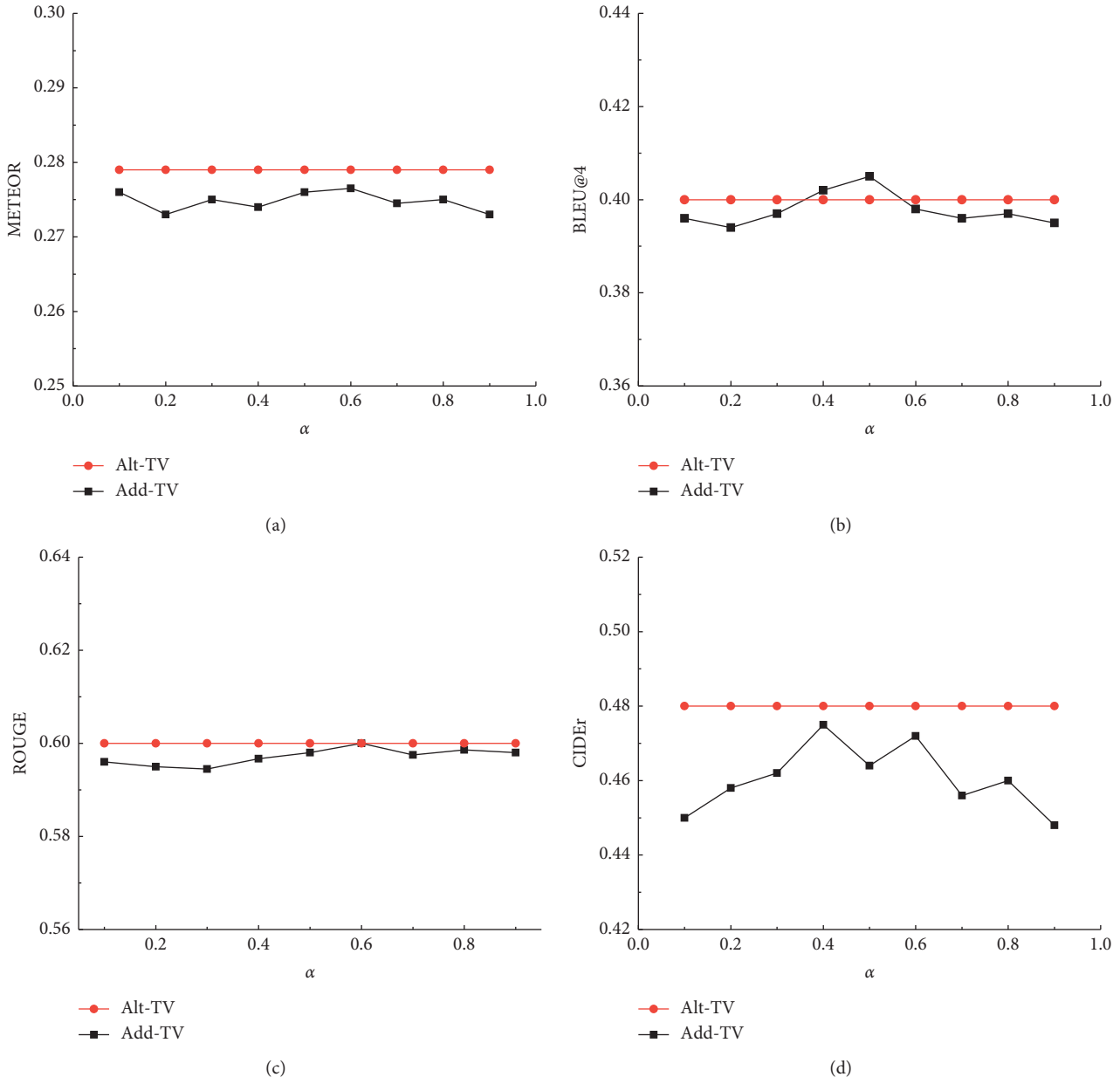

Figure 4: Different parameters in fusion and additive fusion.

L index tends to calculate the recall rate, the METEOR index is applicable to the field of machine translation, and CIDEr is used to evaluate the quality of automatic image description [24–27].

4.2. Exploration of Parameter α of the Additive Fusion Module. In order to verify the effectiveness of the attention fusion module, for the MSR-VTT dataset, α with different parameters is selected to compare the performance of the additive fusion module and the attention fusion module. The test results are shown in Figure 4.

Figure 4 shows the comparison between the attention module and the additive fusion module under different parameters. The evaluation results show that when the parameters α are adjusted to about 0.4, but its METEOR and CIDEr scores are still lower than those of the attention fusion module. Therefore, compared with the fixed weight ratio, the dynamic attention weight introduced by the attention fusion module is more flexible in fusing multimodal features and can generate higher quality text descriptions.

4.3. Comparison with the LSTM Model. In order to verify the performance of the video description model combining the long-term and short-term memory network and attention mechanism, this paper implements a mainstream video description model based on LSTM. Except that the structures of the encoder and decoder are different, the evaluation indexes are compared on MSVD and MSR-VVT

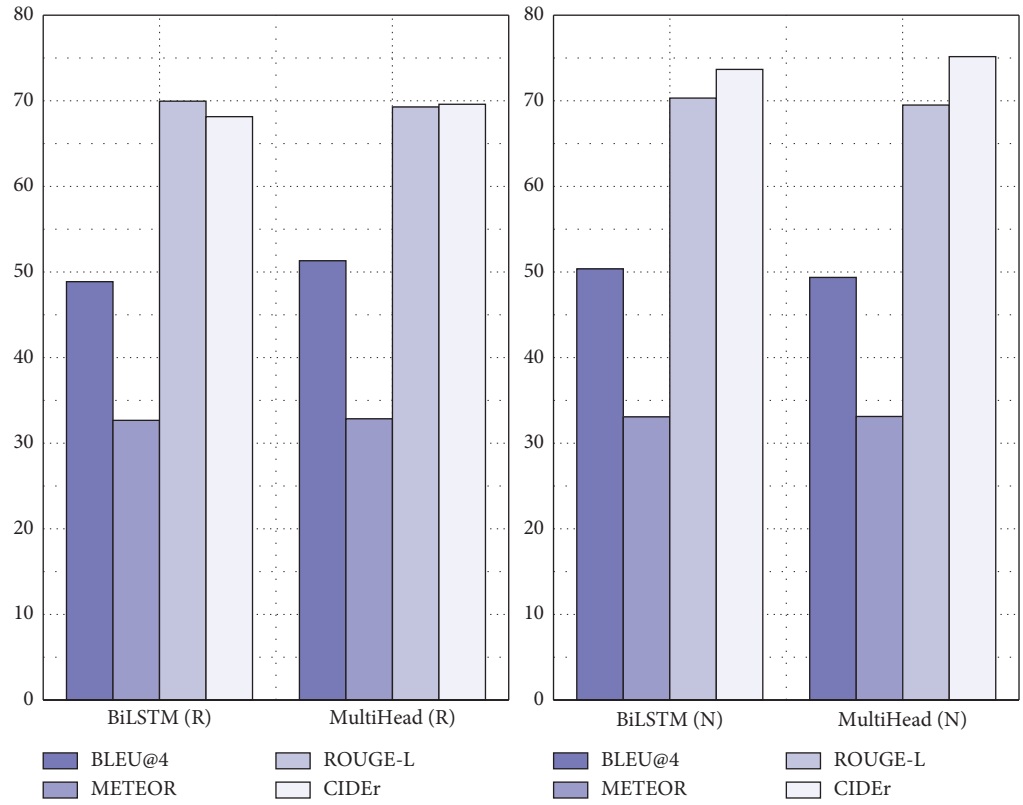

Figure 5: Test results of BiLSTM and multihead models on the MSVD dataset.

datasets when other parameters are set close to the same parameters. In this paper, the model with the attention mechanism is recorded as the multihead model, and the model without the attention mechanism is recorded as BiLSTM. The visual extraction layer is recorded as R using ResNet-152 and N using NASNet. The evaluation results are shown in Figures 5 and 6. The horizontal and vertical dimensions of the graph are the algorithm models used, and the vertical coordinates are the scores of different models.

As can be seen from Figure 5, the METEOR and CIDEr scores of the multihead model on the MSVD dataset are higher than those of the BiLSTM model. These two indicators can also better reflect the quality of text description generation, indicating that the quality of text description generation has been significantly improved after the introduction of the attention mechanism.

As can be seen from Figure 6, in addition to the ROUGE-L score, the other three indicators of the multihead model on the MSR-VTT dataset are higher than those of the BiLSTM model. This is because the introduction of the attention mechanism can make the structure of the visual feature sequence and word sequence more flexible, and better video and sentence content representation can be obtained.

In addition, in the experiment, NASNet, as a visual feature extraction, has greatly improved on the MSVD dataset compared with ResNet-152 and only slightly decreased on the MSR-VTT dataset, which shows that the NASNet pretraining model has a strong generalization ability.

4.4. Comparison of Different Parameters of Cluster Search. In order to explore the impact of different parameters on the multihead model, this paper explores the impact of different beam widths k and length penalty coefficients α_t on the text quality generated by the model on the MSVD test set. First, control the length penalty coefficient $\alpha_t = 1.0$ to remain unchanged, and change the beam width k to 1, 3, 5, 10, and 20, respectively. The evaluation results are shown in Figure 7.

The evaluation results in Figure 7 show the impact of different beam widths on the quality of the generated text. The results show that the generated text can obtain higher evaluation scores with the increase of beam width, but when the beam width increases to more than 5, the gain on scores is relatively small, and the CIDEr score will decrease slightly, which will bring greater search cost. Therefore, a beam width of 5 was used in subsequent experiments.

The evaluation results in Figure 8 show the impact of different length penalty coefficients on the quality of the generated text. The results show that when the length penalty coefficient is not set or the length penalty coefficient is small, the average sentence length generated is short, which is due to the tendency to output shorter candidate sequences during technical search. BLEU@4 Scores are used to calculate accuracy. The smaller the length penalty coefficient, the higher the score, but this has little effect on other scores.. When the generated sentence is short, the accuracy will be improved because there are fewer 4 tuples in the generated sentence.

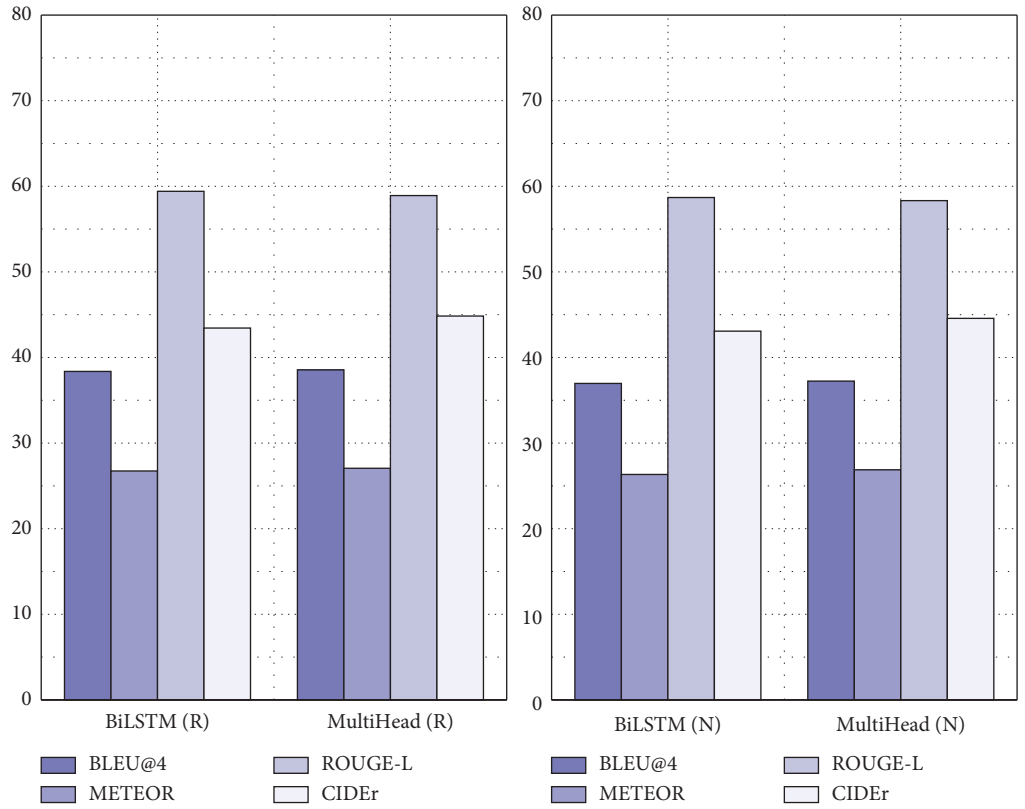

Figure 6: Test results of BiLSTM and multihead models on the MSR-VTT dataset.

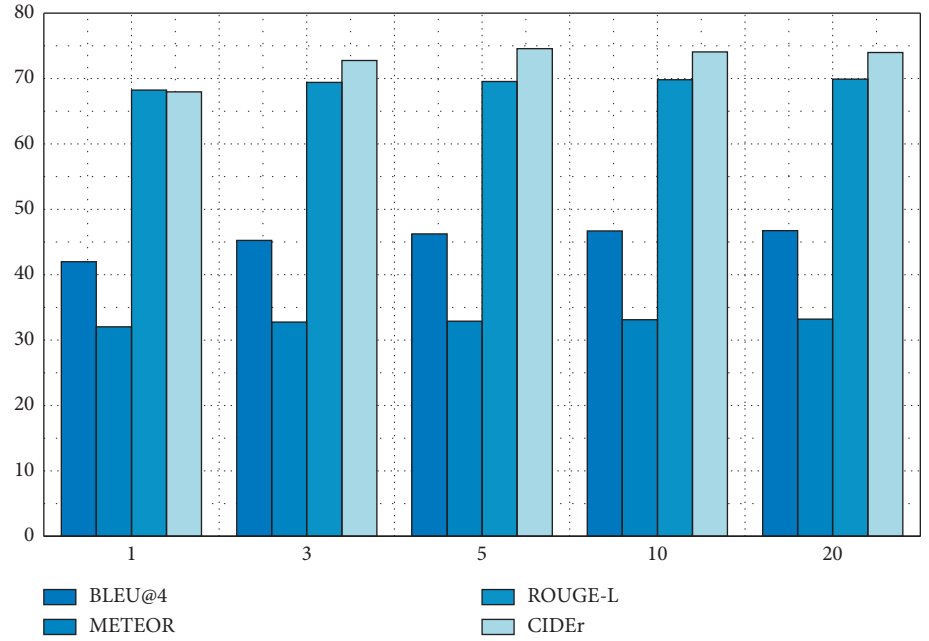

Figure 7: Effects of different beam widths on the quality of text generated by the model when the length penalty coefficient is fixed.

4.5. Comparison with the Baseline Model. This section is to verify the effectiveness of the multihead model combined with the long-term and short-term memory network and attention mechanism model and compare it with baseline model BaseModel on MSVD and MSR-VTT datasets, respectively. The test results are shown in Figures 9 and 10, respectively.

As can be seen from Figure 9, the static visual features extracted by NASNet on the MSVD dataset are greatly improved compared with ResNet-152. The index scores of

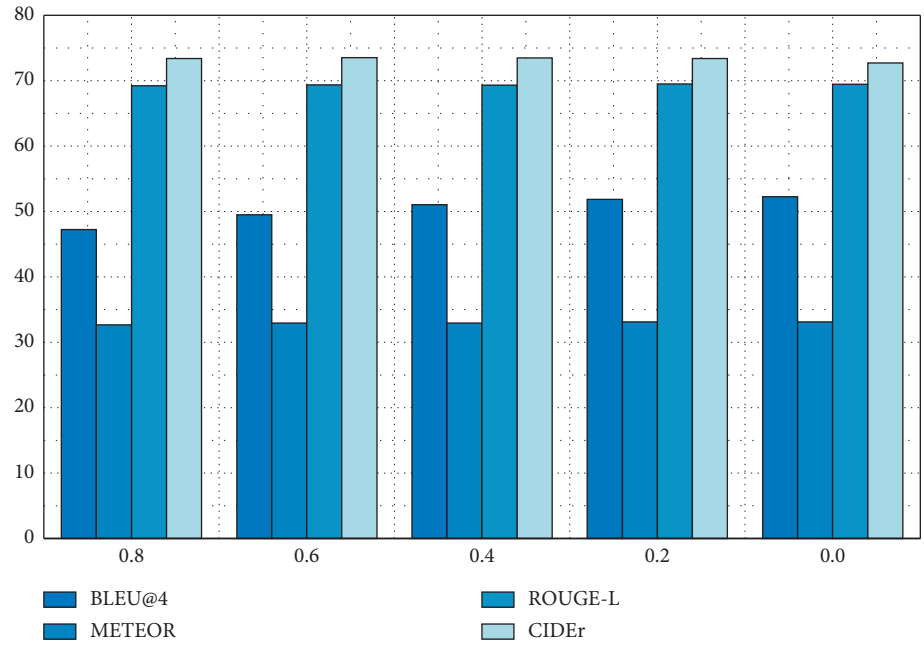

Figure 8: Effects of different length penalty coefficients on the quality of text generated by the model when the beam width is fixed.

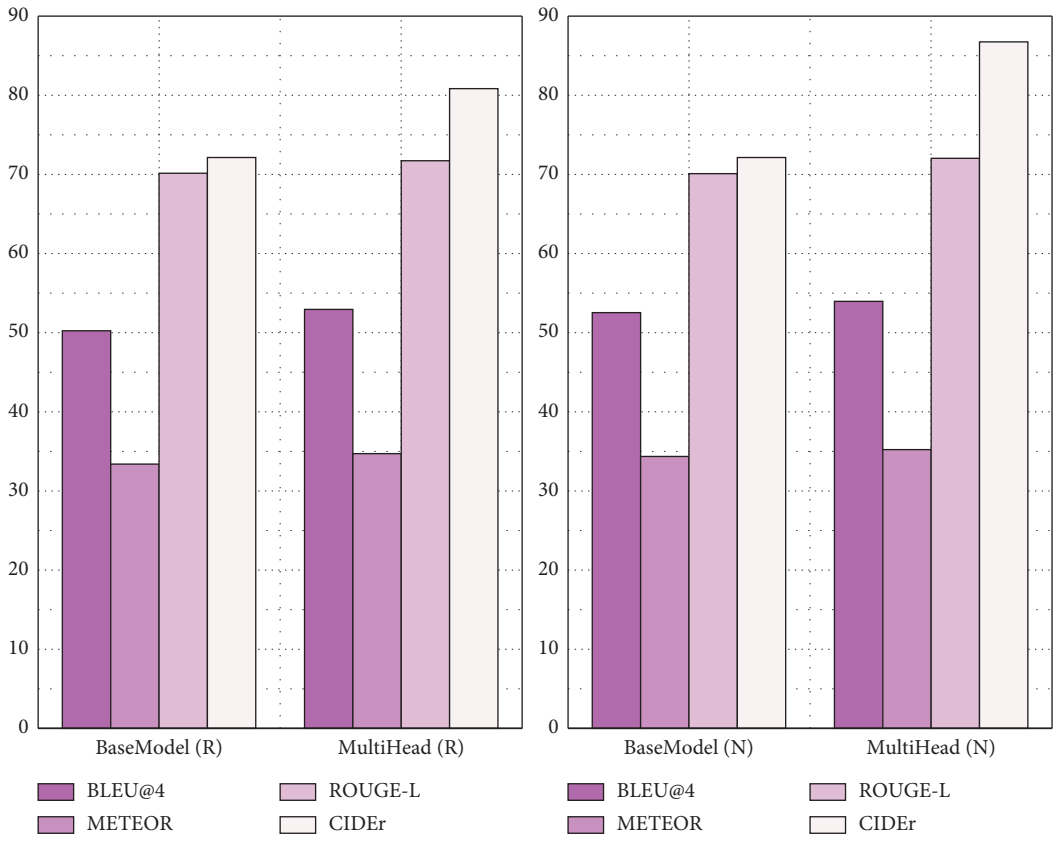

Figure 9: Test results of BaseModel and multihead models on the MSVD dataset.

the multihead model are better than those of BaseModel, which shows that the method proposed in this paper has a significant improvement in the quality of text description compared with the baseline model.

As can be seen from Figure 10, NASNet and ResNet-152 have the same performance on the MSR-VTT dataset. The index scores of the multihead model are significantly higher than those of BaseModel, which shows that the method

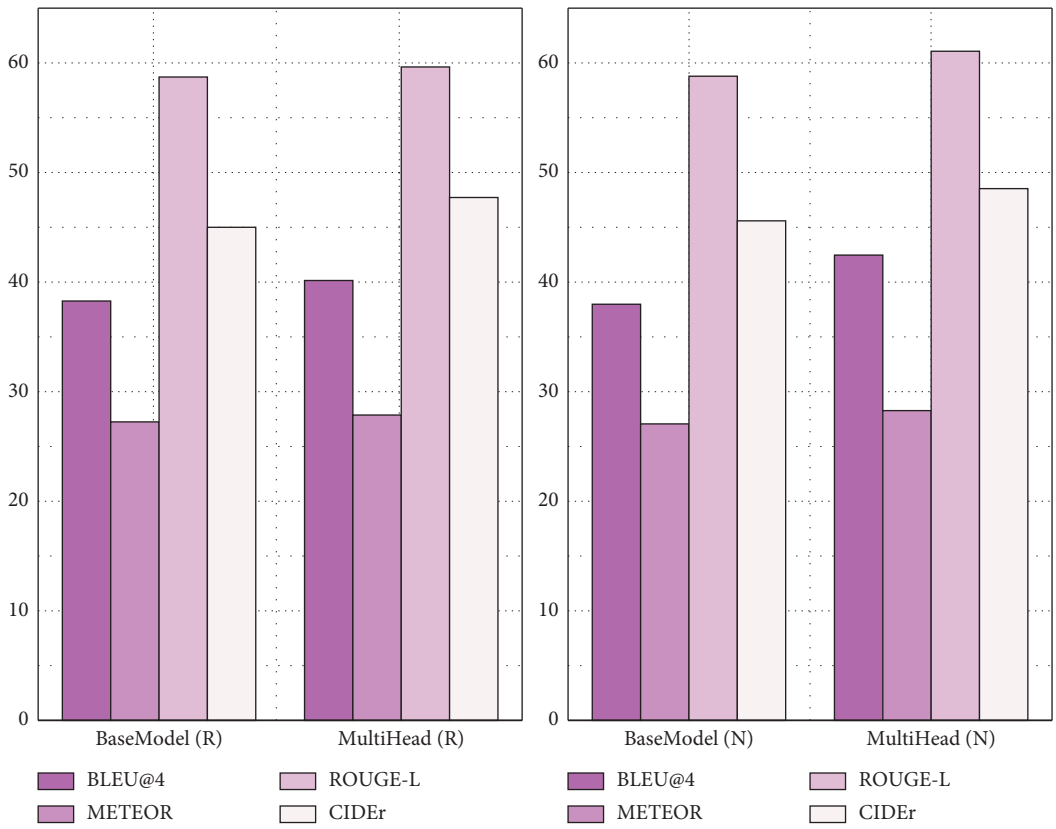

Figure 10: Test results of BaseModel and multihead models on the MSR-VTT dataset.

proposed in this paper has certain generalization ability and significantly improves the quality of text description compared with the baseline model.

5. Conclusion

In this paper, a video sensor processing method combining the long-term and short-term memory network and attention mechanism is proposed for the intelligent description of the volleyball video. The introduction of the attention mechanism can make the model pay much attention to the important areas in the image/video when generating sentences, quickly identify targets, and effectively solve the problem of visual information loss. Through the comparative experiments of different models, the results show that the dynamic attention weight introduced by the attention fusion module is more flexible than the fixed weight and can generate higher quality text description. Compared with LSTM and base model, the multihead model proposed in this paper combines the long-term and short-term memory network and attention mechanism, scores higher in various evaluation indicators, and significantly improves the quality of the intelligent text description of the volleyball video. The model has a strong generalization ability and good performance in the intelligent description of the volleyball video.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Guangdong Province Natural Sciences Fund (no. 2014A030310430), Research on the Technical and Tactical Rhythm Model and Its Forming Mechanism of Elite Athletes of "Big Three Ballgame" in China.

References

[1] Z. Chang and D. Zhao, "Review of video description methods based on deep learning," *Journal of Tianjin University of Technology*, vol. 36, no. 6, pp. 20–26, 2020.

[2] J. Ye, *Video Description Generation Based on Visual Semantic Enhancement*, Zhejiang University of Technology and Industry, Hangzhou, China, 2019.

[3] J. Xu, *Key Technologies of Surveillance Video Moving Target Detection and Pedestrian Structured Description Based on Deep Learning*, Shandong University, Jinan, China, 2019.

[4] Y. Li, H. Zongbo, and L. hang, "Review of convolutional neural networks," *Computer Applications*, vol. 36, no. 9, pp. 2508–2515, 2016.

[5] G. Zhu, *Research on Sports Video Content Analysis Method Based on Team Member Behavior Information*, Harbin Institute Of Technology, Harbin, China, 2011.

[6] C. Wang, Y. Liu, and P. Wang, "Agricultural machinery movement navigation system based on binocular vision detection technology," *Electrotehnica, Electronica, Automatica*, vol. 62, no. 2, 2020.

[7] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[8] M. Rohrbach, W. Qiu, and I. Titov, "Translating video content to natural language descriptions," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, December 2013.

[9] J. Thomason, S. Venugopalan, and S. Guadarrama, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland, December 2014.

[10] J. Donahue, L. A. Hendricks, and S. Guadarrama, *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*, Elsevier, Amsterdam, Netherland, 2015.

[11] S. Venugopalan, H. Xu, and J. Donahue, "Translating videos to natural language using deep recurrent neural networks," *Computer Science*, vol. 3, 2014.

[12] S. Venugopalan, M. Rohrbach, and J. Donahue, "Sequence to sequence—video to text," in *Proceedings of the IEEE 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4534–4542, Santiago, Chile, September 2015.

[13] C. Zhang and Y. Tian, "Automatic video captioning via multichannel sequential encoding," in *Proceedings of the European Conference on Computer Vision*, Springer International Publishing, Amsterdam, The Netherlands, October 2016.

[14] Y. Pan, T. Mei, and T. Yao, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[15] L. Yao, A. Torabi, and K. Cho, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4507–4515, Santiago, Chile, December 2015.

[16] K. Gre Ff, R. K. Srivastava, and J. Koutník, "LSTM: a search space odyssey," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[17] X. Sun, Y. Su, and Y. Zhao, "Mongolian Chinese neural machine translation based on encoder decoder reconstruction framework," *Computer applications and software*, vol. 37, no. 4, pp. 155150–155163, 2020.

[18] H. Wang, J. Shi, and Z. Zhang, "Semantic relation extraction of LSTM based on attention mechanism," *Computer application research*, vol. 35, no. 5, pp. 143–146, 2018.

[19] M. Li, *Analysis and Comparison of Image Salient Region Extraction Algorithms Based on Attention Mechanism*, Beijing Jiaotong University, Beijing, China, 2020.

[20] M. Mukhiddinov, R. G. Jeong, and J. Cho, "Saliency cuts: salient region extraction based on local adaptive thresholding for image information recognition of the visually impaired," *International Arab Journal of Information Technology*, vol. 17, no. 5, pp. 713–720.

[21] X. Xiong and P. Yan, "Chinese classification method integrating multi head self attention mechanism," *Electronic measurement technology*, vol. 43, no. 10, pp. 130–135, 2020.

[22] S. Guadarrama, N. Krishnamoorthy, and G. Malkarnenkar, "YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Columbus, OH, USA, June 2014.

[23] J. Xu, M. Tao, and T. Yao, "MSR-VTT: a large video description dataset for bridging video and language," in *Proceedings of the Conference on Computer Vision And Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2016.

[24] S. Papineni, *Blue; A Method for Automatic Evaluation ofMachine Translation*, in *Proceedings of the Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, College Park, MA, USA, June 2002.

[25] J. P. Ng and V. Abrecht, "Better summarization evaluation with word embeddings for ROUGE," 2015, https://arxiv.org/abs/1508.06034.

[26] B. Satanjeev, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," *ACL*, vol. 7, pp. 228–231, 2005.

[27] CIDEr, "Consensus-based image description evaluation," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, June 2015.