Video Description Method based on Semantic Information Filtering and Sentence Length Modulation

Xiangqing Wang, Xiaodong Cai*, Meixin Zhou, Qingnan Huang

School of Information and Communication, Guilin University of Electronic Technology, Guilin, China
**Corresponding author e-mail: caixiaodong@guet_edu.cn*

Abstract:
In the current video description task, the spatial redundancy information in the video features is usually not effectively eliminated, and the commonly used loss function is composed of the logarithm of the probability of the correct word of the target, and the long sentences formed often bring great losses to the model. If the sentence length generated by the optimization of the log-likelihood loss function is too short, the description semantics will be incomplete and the accuracy will not be high. This paper proposes a video description method based on semantic information filtering and sentence length modulation to solve the above problems. Firstly, the model introduces a gated fusion mechanism, which removes redundant information in the semantic information of video features by screening the semantic features of the video, reduces the interference of redundant information on the generated description, and improves the accuracy of the description. Secondly, a new sentence length modulation loss function is proposed, which modulates the cross-entropy loss function with the label sentence length, which alleviates the tendency of the model to generate short sentences, and makes the semantics of the generated description close to the label, thereby improving the accuracy of the description. The experimental results on the MSVD dataset, which is widely used in this field, show that the method in this paper can significantly improve the accuracy of generating video descriptions, and all indicators are significantly better than existing models.

Keywords: *Video description; Encoder-decoder; Fusion mechanism; Sentence length modulation; Deep learning.*

1 INTRODUCTION

The task of generating corresponding natural language descriptions for a given video content determines its nature. Technology that combines the two directions of computer vision and natural language processing is needed, and it acts as a link for the connection between the two.

There are two main directions for solving the video description problem. Early adopt template-based methods [5] [7], which first define a sentence template with grammatical rules, and then align the subject, predicate, and object of the sentence template with the video content, which has been extensively studied. Due to the fixed grammatical structure of the predefined templates, it is difficult for these methods to generate flexible languages. Benefiting from the rapid development of deep neural networks, sequence learning methods [11] [16] are widely used to describe videos in flexible natural language. Most of these methods are based on the encoder-decoder framework, proposed by Venugopalan et al. [16] The S2VT model treats the video description task as a machine translation task. Yao et al. [20] introduced a temporal attention mechanism that assigns weights to the features of each frame and then fuses them based on the attention weights. Li et al. [8] and Chen et al. [2] further applied the spatial attention mechanism to each frame. In recent years, Wang et al. [18] and Hou et al. [6] proposed to utilize part-of-speech (POS) tags to facilitate video descriptions. [18] encoded the predicted POS sequences as hidden features to further guide the generation process. Hou et al. [6] mix the word probabilities of multiple components at each time step based on the inferred POS labels. However, for rich video content, these two types of methods still have many problems, such as failing to effectively filter out the spatial redundancy information in the video features, and failing to consider the impact of the length of the generated sentence on the integrity of the video semantic expression, resulting in The generated description is not accurate enough.

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 446-452, 2023.
https://doi.org/10.2991/978-94-6463-040-4_68

To solve the above problems, this paper proposes a video description method based on semantic information screening and sentence length modulation, referred to as SIF-SLM. Aiming at the problem that video features contain spatial redundant information, inspired by Srivastava et al. [14], a gated fusion mechanism is proposed to remove redundant or unimportant information in the semantic information of video features, improve the utilization of semantic information, and then improve the completeness of description. Aiming at the problem that existing models tend to generate shorter descriptions, inspired by Chen et al. [3], a sentence length modulation loss function is proposed, and the label sentence length is integrated into the original loss function to improve the integrity of the semantic description, thereby improving the accuracy of the description.

2 DESIGN OF SIF-SLM MODEL

Based on the codec framework, the overall framework of the video description model proposed in this paper is shown in Figure 1. The model is divided into two stages: encoding and decoding. The encoder part consists of a feature extraction module, an inference module, and a selection module. Firstly, the input video is processed to obtain two modalities of images containing static features and video clips with dynamic features, and then using a two-dimensional convolutional neural network (2D-CNN), regional convolutional neural network (R-CNN) and Three-dimensional convolutional neural network (3D-CNN) extracts features from image modalities and video modalities, and obtains three features as Va, Vo, and Vm, and then passes the features through three inference modules to obtain three part-of-speech features (nouns, verbs and functional words) are Vlt, Vrt, and Vft. Then the selection module includes two steps, first, calculate the score Scorer of the three part-of-speech features, measure the probability of being selected, and then use the Gumbel Softmax strategy, according to these scores, the selection module selects the next most likely to generate part-of-speech features. It is passed through a gated fusion mechanism Gate, and the feature V_t that filters out the spatial redundancy information are obtained as the input of the decoding stage. The decoder decodes V_t to obtain the description W corresponding to each video.

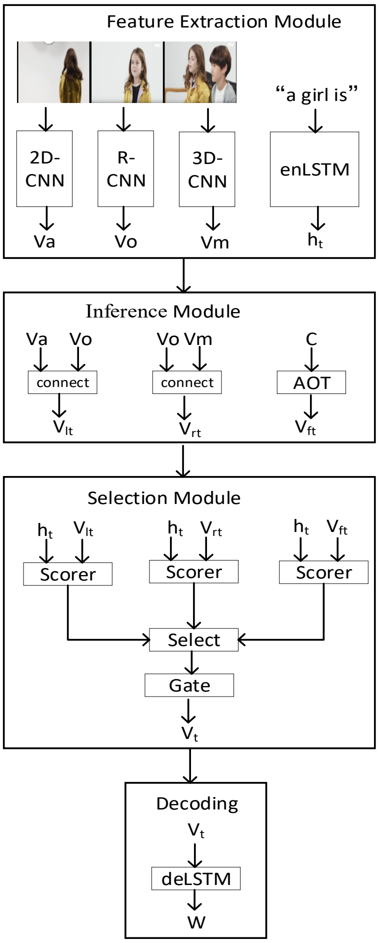


Figure 1: Overall framework of the model.

2.1 *Semantic Information Screening Using Gated Fusion Mechanism*

The video features obtained by the selection module of the coding layer have a lot of spatial redundancy information. If the features of this information are directly used for decoding, it will become an interference factor during decoding, which may lead to unclear primary and secondary descriptions, which may affect the decoding process. accuracy of the model. This study proposes a gated fusion mechanism to screen semantic information. The output of the encoding layer selection module is the splicing of two different features of the input video, which is filtered first through the gated fusion mechanism, the valid information is passed directly, and another part of the features with redundant information is screened again, and finally, the weighted and combined output of these two parts further encodes the features into a more compact representation so that the semantic information can be better integrated.

Assuming that the splicing feature of two different features of the input video is x, and the output after the gated fusion mechanism is y, inspired by Srivastava et al. [3], formulas (1) and (2) are obtained, namely:

$$h = H(x,W_H) \tag{1}$$

$$y = h \odot t + x \odot c \tag{2}$$

Among them, the transformation H whose parameter is WH is usually an affine transformation plus a nonlinear activation function, the output y is determined by the output h and the input x after the transformation H, and t and c are the weights of these two parts, so take The value is a number between 0 and 1.

To more intuitively see the composition and meaning of formula (2), it is transformed as follows:

The weight parameters t and c are obtained by nonlinear transformations T (x, WT) and C (x, WC), respectively. In this way, formula (2) becomes:

$$y = H(x, W_H) \odot T(x, W_T) + x \odot C(x, W_C) \tag{3}$$

In the above formula, the former term represents the converted part of the input information, that is, the spatial redundancy information that needs to be filtered out, and the latter term represents the retained part of the original information. Therefore, T is called the transform gate, and C is called the carry gate.

Let C=1-T, then formula (3) becomes:

$$y = H(x, W_H) \odot T(x, W_T) + x \odot (1 - T(x, W_T)) \tag{4}$$

To make t between 0 and 1, define T as:

$$T(x) = \sigma({W_T}^T x + b_T) \tag{5}$$

Because in the video feature, the proportion of the information to be retained in the input x is too large, and the proportion of the spatial redundant information to be filtered out is too small, so initialize the bias bT of T to a negative number, so that the size of the parameter t is between 0 and 0.5 The size of c is between 0.5 and 1, and the specific value will be determined according to the ratio of effective information and redundant information in the video features. In this way, useful information can be retained, redundant information can be filtered out, and more concise features can be obtained, which in turn can generate more accurate descriptions.

2.2 Design of a Loss Function Using Sentence Length Modulation

As a language generation task, video description models are usually trained by minimizing the cross-entropy loss function [9]. The cross-entropy loss function consists of the logarithm of the target correct word probability [4] [16], and long sentences tend to cause a large loss to the model since each additional word reduces the joint probability by at least an order of magnitude. In contrast, short sentences with few words have relatively small losses. Therefore, the video description model is easy to generate short sentences after being optimized by the log-likelihood loss function. Too short annotations can neither accurately describe the video nor express the content of the video in rich language.

This paper adopts the loss function proposed by Chen et al. [3], which can make the description generated by the model accurate and concise. The cross-entropy loss function is weighted according to the length of the true labels, as shown in Equation (6):

$$Loss(\hat{y}_i, s_i, X_i, \theta) = \\ -\sum_{i=0}^{bs-1} \frac{1}{L_i^{\beta}} \sum_{t=0}^{L_i-1} \log p(\hat{y}_{i,t} \mid h_{i,t-1}, c_{i,t-1}, s_i, X_i; \theta) \tag{6}$$

Among them, bs represents the batch size, Li represents the label sentence length, $\hat{y}_{i,t}$represents the t-1th word in this label, $h_{i,\, t\text{-}1}$, $c_{i,\, t\text{-}1}$, s_i are the output state, cell state, and Semantic features, Xi represents the ith video of the input, and β >= 0 is a hyperparameter used to balance the brevity and accuracy of the generated description.

If β=0, formula (6) becomes a loss function commonly used in this field, that is formula (7). In this loss function, each time a word is added, the joint probability will be reduced by at least an order of magnitude, so the long The loss of sentences is larger than that of short sentences, so after minimizing the loss, the model tends to generate shorter descriptions, which may lead to incomplete semantics of the generated sentences and affect the accuracy of sentence generation.

$$Loss(\hat{y}_i, s_i, X_i, \theta) = \\ -\sum_{i=0}^{bs-1} \sum_{t=0}^{L_i-1} \log p(\hat{y}_{i,t} \mid h_{i,t-1}, c_{i,t-1}, s_i, X_i; \theta) \tag{7}$$

If β=1, all words in the generated description are treated equally in the loss function, which may lead to redundant or repeated words in the description generation process.

During the training process, β adaptively adjusts the size and automatically adjusts the balance between the accuracy and conciseness of the generated description, so that the generated description is both concise and accurate.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Lab Bed and Setup

The experimental platform configuration in this paper includes an intel i5-1135G7-core 2.42GHz processor, 16GB memory, NVIDIA GeForce RTX 3090, and Ubuntu18.04 operating system, and uses the PyTorch deep learning framework based on the python programming language.

Model verification uses the commonly used public data set MSVD. MSVD is a public data set published by Microsoft Research in 2010. The data set consists of 1970 video clips. The data set is divided into 3 subsets, of which the training set contains 1200 videos, the validation set contains 100 videos, the test set contains 670 videos, and each video segment contains an average of 40 manually annotated sentences, and the model adopts the English annotation sentences in this dataset.

After many experiments, when the parameters are set to the following data, the experimental effect is the best. The experimental parameters are specifically set as follows: the model adopts the Adam optimization method, the initial learning rate is set to 0.0001, the hidden state dimension is set to 512, the learning rate is exponentially decayed, and the decay is performed every 10 cycles. The training batch size is 32. During testing, a beam search of size 2 was used to generate the final description.

In the evaluation indicators, this paper adopts four widely used text quality evaluation indicators officially provided by Microsoft, namely METEOR [1], BLEU(n) [12], ROUGE [10], and CIDEr [17]. BLEU(n) focuses on accuracy and is an indicator for comparing the degree of overlap between the n-grams of the generated description and the label. The higher the degree of overlap, the higher the quality of the generated description. ROUGE focuses on recall, i.e. how many n-grams in the label appear in the generated description. METEOR considers both precision and recall based on the entire corpus, and the METEOR value is calculated as the harmonic mean of precision and recalls between the corresponding best candidate translation and the reference translation. CIDEr is used for image annotation problems, which is a weighted evaluation index that pays more attention to whether keywords appear. They use different methods to evaluate the similarity between the descriptions produced by the model and the human annotations, and higher scores for each indicator indicate more similarity between the two. The higher the scores on these metrics, the higher the quality of the descriptions generated.

3.2 *Analysis of Ablation Experiment Results*

The model of RMN is from the literature [15]. In the ablation experiments in Table 1, the effectiveness of the proposed gated fusion mechanism and sentence length modulation loss function method is proved, where "RMN" corresponds to the experimental results of the literature [15], The methods proposed in this paper are based on the benchmarking paper model.

Table 1: Ablation Experiment on MSVD dataset

Models	B4	M	R	C
RMN	54.6	36.5	73.4	94.4
SIF	58.12	37.41	73.93	94.75
SLM	55.13	36.62	73.84	95.43

Firstly, the effectiveness of the gated fusion mechanism is verified. The experimental results are shown in "SIF" in Table 1. By comparing the results of RMN and SIF, each index has been significantly improved, BLEU@4 increased by 3.52%, and METEOR increased by 0.91%. The ROUGE is increased by 0.53%, and the CIDEr is increased by 0.35%. The results show that the SIF method in this paper can reduce the redundant information in the features and improve the accuracy of description generation.

Secondly, the effectiveness of the sentence length modulation loss function is verified. The experimental results are shown in "SLM" in Table 1. By comparing RMN and SLM, each index has been significantly improved, BLEU@4 increased by 0.53%, METEOR increased by 0.12%, ROUGE The increased is 0.44%, and the CIDEr is increased by 1.03%, indicating that the SLM method in this paper can make the generated description more accurate and closer to the expression of the label.

From the actual numerical analysis, the method in this paper has improved in various indicators. From the actual effect, this paper compares and analyzes the description generation results of the proposed SIF-SLM and the RMN model. The latter's 100 test samples are generated. As a result, there are about 55 samples with inaccurate descriptions. After introducing the SIF and SLM proposed in this paper, 15 and 17 samples have been significantly improved, with an increase of 27.3% and 30.1%, and the actual application effect has improved significantly.

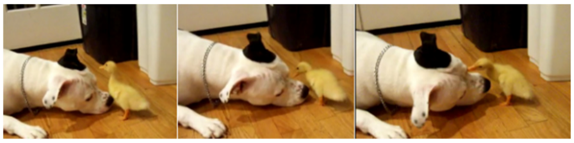
3.3 *Comparative Analysis of Experimental Results*

As shown in Table 2, the proposed SIF-SLM is compared with other state-of-the-art methods on the MSVD dataset. According to whether POS tags are utilized or not, we classify them into two groups: (1) traditional encoder-decoder-based models MAM-RNN [8] and RecNet [19], (2) POS enhanced models POS-CG [8], Mixture [19] and SGN [13]. As shown in Table 2, it can be found that: (1) the method with POS tags is better than the method without POS information, and (2) our proposed SIF-SLM is significantly better than the method with POS tags in every indicator, especially The BLEU@4 and CIDEr metrics significantly improved by 5.78% and 1.27%, reaching new levels. It is worth noting that CIDEr is especially proposed for the captioning task, which is considered to be more in line with human judgment. Our model achieves significantly better CIDEr scores than existing research results on this public dataset, indicating that our model can generate sentences that are more in line with a good human reading experience.

Table 2: Performance comparison with other models on MSVD dataset.

Models	B4	M	R	C
MAM-RNN	41.3	32.2	68.8	53.9
RecNet	52.3	34.1	69.8	80.3
POS-CG	54.3	36.4	73.9	95.2
Mixture	52.8	36.1	71.8	87.8
RMN	54.6	36.5	73.4	94.4
SGN	52.8	35.5	72.9	94.3
SIF-SLM	60.38	37.59	74.10	95.67

A visualization example of the generated description on the MSVD dataset using the method SIF-SLM in this paper is shown in Figure 2. Among them, RMN benchmarks the method of the paper [Tan 2020], SIF-SLM represents the method proposed in this paper, and GT represents the ground-truth label of the video.

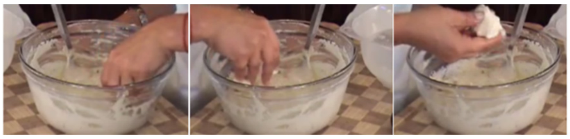


GT: a dog is playing with a toy

RMN: a dog is playing;

SIF-SLM: a dog is playing with a toy

(a)



GT: a man is mixing ingredients in a bowl

RMN: a person is mixing eggs;

SIF-SLM: a man is mixing ingredients in a bowl

(b)

Figure 2:Visual example of SIF-SLM on MSVD dataset

From the example results in Figure 2, it can be seen that the SIF-SLM model proposed in this paper is significantly improved compared to the benchmarking model. On the one hand, the description of the semantics is more complete. In Figure 2(a), the benchmarking model RMN does not express the semantic information of the toy, resulting in incomplete description semantics, and the model SIF-SLM in this paper can fully express the semantic information of the toy. On the other hand, the expression of key information is more accurate, as shown in Figure 2(b), RMN does not accurately grasp the key information, and the model in this paper accurately expresses key information such as ingredients and bowl, which improves the accuracy of the description.

4 CONCLUSIONS

In this paper, the author proposes a gated fusion mechanism to reduce redundant information in features and improve the availability of video features. A new loss function is also proposed, which can be derived from the length of manually annotated labels. The length of the generated sentences is adjusted adaptively to keep the generated description concise and accurate. In the ablation experiments and comparative experiments on the commonly used dataset MSVD, the overall performance of the model is excellent, which is better than the existing models, which verifies the effectiveness of the proposed method. In future work, the team will continue to conduct research on video description tasks, aiming to solve the problems of incomplete semantics and inaccurate descriptions in video description sentences, and further improve the performance of the model.

REFERENCES

[1] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. Ann Arbor, Michigan, USA :IEEvaluation@ACL, 2005: 65-72.

[2] Chen S, Jiang Y G. Motion guided spatial attention for video captioning[C]//Proceedings of the AAAI conference on artificial intelligence. Honolulu, Hawaii, USA :AAAI, 2019, 33(01): 8191-8198.

[3] Chen H, Lin K, Maye A, et al. A semantics-assisted video captioning model trained with scheduled sampling[J]. Frontiers in Robotics and AI, 2020: 129.

[4] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, MA, USA :IEEE, 2015: 2625-2634.

[5] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition[C]//Proceedings of the IEEE international conference on computer vision. Sydney, Australia: IEEE, 2013: 2712-2719.

[6] Hou J, Wu X, Zhao W, et al. Joint syntax representation learning and visual cue translation for video captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South) :ICCV, 2019: 8918-8927.

[7] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions[J]. International Journal of Computer Vision, 2002, 50(2): 171-184.

[8] Li X, Zhao B, Lu X. MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning[C]//IJCAI. Melbourne, Australia :IJCAI, 2017: 2208-2214.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[10] Lin C Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics. Edmonton, Canada: HTL-NAACL 2003, 2003: 150-157.

[11] Pan Y, Yao T, Li H, et al. Video captioning with transferred semantic attributes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA :IEEE, 2017: 6504-6512.

[12] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia, PA, USA :40th ACL 2002, 2002: 311-318.

[13] Ryu H, Kang S, Kang H, et al. Semantic grouping network for video captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event :AAAI, 2021, 35(3): 2514-2522.

[14] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387, 2015.

[15] Tan G, Liu D, Wang M, et al. Learning to discretely compose reasoning module networks for video captioning[J]. arXiv preprint arXiv:2007.09049, 2020.

[16] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 4534-4542.

[17] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Online:7th-W-NUT, 2015: 4566-4575.

[18] Wang B, Ma L, Zhang W, et al. Controllable video captioning with pos sequence guidance based on gated fusion network[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea (South) :ICCV, 2019: 2641-2650.

[19] Wang B, Ma L, Zhang W, et al. Reconstruction network for video captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA :IEEE, 2018: 7622-7631.

[20] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//Proceedings of the IEEE international conference on computer vision. Santiago, Chile :IEEE, 2015: 4507-4515.