Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VMSG: a video caption network based on multimodal semantic grouping and semantic attention

Xin Yang (yangxin@nuaa.edu.cn)
Nanjing University of Aeronautics and Astronautics
Xiangchen Wang
Nanjing University of Aeronautics and Astronautics
Xiaohui Ye
Nanjing University of Aeronautics and Astronautics
Tao Li
Nanjing University of Aeronautics and Astronautics

Research Article

Keywords: Video caption, multimodal, semantic grouping, semantic attention

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1542723/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

VMSG: a video caption network based on multimodal semantic grouping and semantic attention

Xin Yang, Xiangchen Wang, Xiaohui Ye, Tao Li

(College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

(04/08/2022)

Corresponding Author: Xin Yang:
yangxin@nuaa.edu.cn

Tel: 86-13770510498 Fax: 86-25-58784479

Abstract: Network video typically contains a variety of information that is used by the video caption model to generate video tags. The process of creating video captions is divided into two steps: video information extraction and natural language generation. Existing models have the problem of redundant information in continuous frames when generating natural language, which affects the accuracy of the caption. As a result, this paper proposes a Multimodal Semantic Grouping and Semantic Attention Video Caption Model (VMSG). VMSG uses a novel semantic grouping method for decoding, which divides the video with the same semantics into a semantic group for decoding and predicting the next word, to reduce the redundant information of continuous video frames, which differs from the decoding mode of grouping by frame. Because the importance of each semantic group varies, we investigate a semantic attention mechanism to add weight to the semantic group and use a single-layer LSTM to simplify the model. Experiments show that VMSG outperforms some state-of-the-art models in terms of caption generation performance and alleviates the problem of redundant information in continuous video frames.

Keywords: Video caption; multimodal; semantic grouping; semantic attention

1 Introduction

Human intelligence manifests itself in two ways: visual perception and language expression. Video caption generation is a common application of artificial intelligence that combines visual data and natural language. Understanding a visual scene and naturally describing it is referred to as video captioning. It is a hot topic in computer vision. There are two important aspects to video captioning. The first step is to extract information about distinguishable features from the video. The second step is to extract features and convert them into natural language [1-4]. In computer vision and natural language processing, data-driven deep learning is the primary processing method. Figure 1 depicts a video caption generation example.

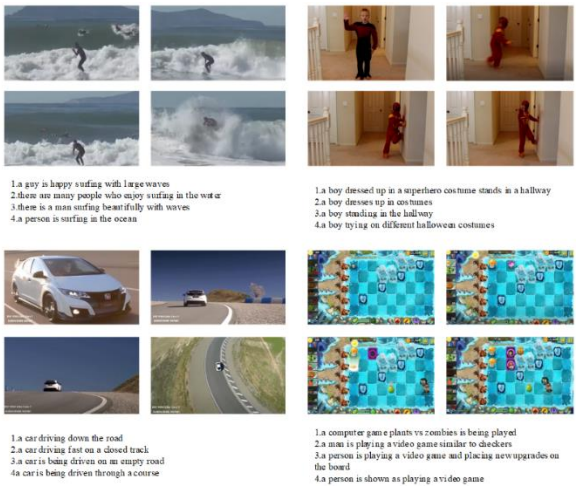

Figure 1 Example of video caption generation

The most widely used video caption algorithms are encoder-decoder frameworks based on convolutional neural networks and recurrent neural networks. The convolutional neural network encoder obtains a set of continuous frames from the input video and generates the corresponding video features. The recurrent neural network-based decoder then takes the visual coding characteristics and previously predicted words as input and generates a word each time. The static content in a single image must be understood by the video caption. In contrast, the video caption must fully comprehend the video context.

Video on the network contains information that is similar between adjacent video frames. As a result, there is more redundant information between adjacent video frames, which cannot usually provide enough unique information [5]. It is natural for humans to comprehend video by segmenting it into information units using semantics. Therefore, viewing each frame as a separate information unit is ineffective for comprehending video.

The complementarity of visual and text information is critical for the video caption algorithm. However, when encoding video, the previous methods primarily focus on the visual aspect (i.e., video frame) while paying less attention to the text aspect (i.e., partially decoded title). The video caption is made up of text that was predicted by the decoder and summarizes the visual content. As a result, the word phrase made up of partially decoded abstracts can group semantically related frames into information units to form a semantic group.

In a video scene where a boy meets a girl and talks to her, for example, the decoder has partially generated "a boy talking with". The phrase "a boy" can form a semantic group from the image of the boy standing alone, and "talking with" can form a semantic group from the video frame of the following two people talking. The next word, "girl", can be predicted by this semantic group.

To be used as information units for understanding video, semantic groups must

have the three characteristics listed below. To begin, semantic groups' meanings should be concrete and observable. Second, a semantic group should have distinct meanings from other semantic groups, allowing it to be treated as a standalone information unit with no redundancy. Third, all video frames in a semantic group should be closely related to the phrases they represent.

This paper proposes a video caption model based on multimodal semantic grouping and semantic attention (VCMSGA) to address the problem of continuous frame redundancy and insufficient feature extraction. To extract 3D and 2D features, the model employs the 3DResNet [6] neural network and the residual neural network [7]. The classification information for audio and video is then added to the multimodal framework for coding. Once the multimodal features have been obtained, they must be decoded. Unlike the previous decoding mode, which groups frames frame by frame, VMSG decodes using semantic grouping. Because the importance of different semantic groups varies, this paper investigates a semantic attention algorithm to give semantic groups more weight. For decoding and predicting the next word, videos with the same semantics are grouped together. This paper's work and contributions are summarized as follows:

1. This paper proposes a multimodal semantic grouping-based video caption model. Adjacent video frames are composed of a semantic grouping to combine redundant information when using video information.

2. Due to the disappearance of the deep network gradient and other issues, we use the 3DResNet network to extract video rather than the C3D [8] network.

3. We replace the original two-layer network simplification model with a single-layer LSTM and add a semantic attention mechanism.

4. We put the proposed method to the test using a public data set. Experiment results show that the method described in this paper outperforms more advanced algorithms. Following analysis, the model is capable of extracting video features effectively.

The remainder of the paper is organized as follows: Section 2 describes the VMSG model in detail, and Section 3 demonstrates the model's performance through detailed experiments. The final section contains the conclusion and future work.

2 The proposed method

2.1 Overall structure of the model

Our VMSG consists of four modules: video coding, phrase coding, semantic grouping, and decoding. Figure 2 depicts the overall structure of VMSG.

We use multimodal input in the video coding module to provide more information for the input, adding 2D features, 3D features, audio features, and classification features. Based on the foregoing, replacing C3D with a 3DResNet

network can sufficiently alleviate the deep network's degradation and improve the model's extraction of dynamic features.

We encode the phrase and form the phrase based on the obtained words after acquiring the multimodal features. Then, using semantic grouping, we group the phrases that correspond to the video frame and finally form the video representation. Figure 3 depicts a schematic diagram of semantic grouping. W stands for the generated word, P stands for the generated semantic sentence, f stands for the multimodal video frame, and X stands for the video. Each semantic group has a different level of importance. As a result, this paper augments the decoder with an attention mechanism that assigns weight to each semantic group.

We simplify the model by converting the double-layer LSTM into a single layer as the VMSG decoder. In addition, contrastive attention loss based on cross-entropy loss is added by VMSG.

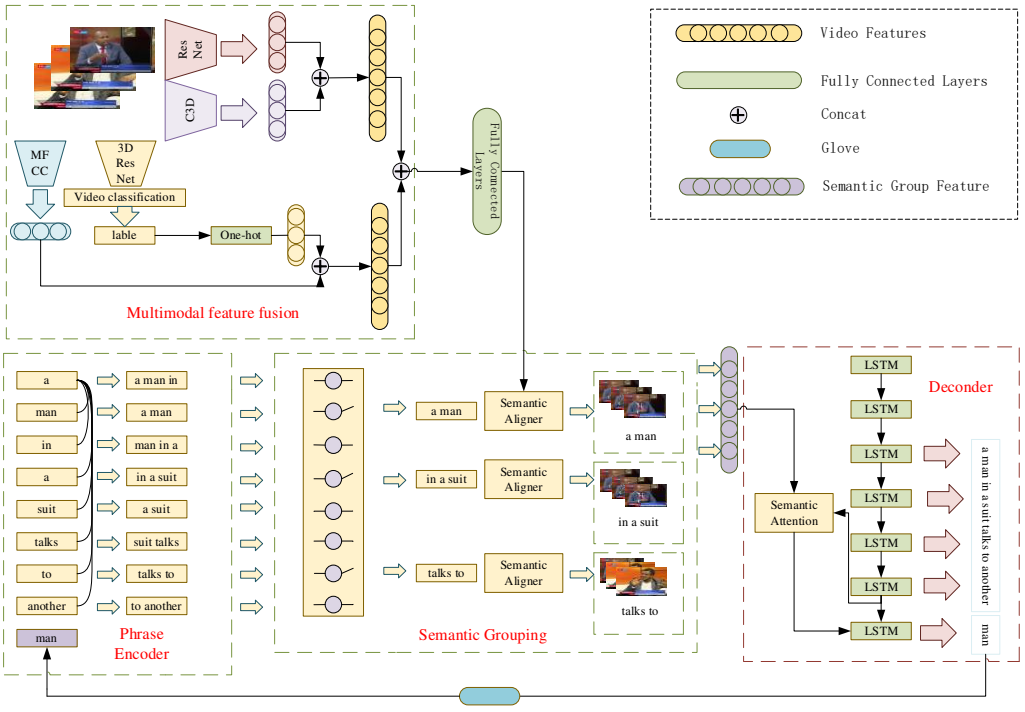

Figure 2. Over architecture of VMSG

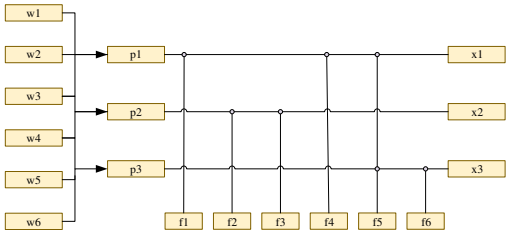

Figure 3. Semantic grouping mode

2.2 Multimodal input

VMSG builds a multimodal segmented label video caption architecture that takes a variety of modal inputs. VMSG significantly improves the types of features and

contributes to the generation of video caption. This paper's multimodal feature fusion employs the early fusion method to splice the multimodal features step by step. The multimodal input includes the following 2D features, dynamic features, video category features, and audio features.

2D features:

2D features are widely used in image detection and classification tasks. It gives specific information about objects and scenes. We employ the ResNet-152 residual network. More than 1.2 million images from 1000 different categories are used to train the model. We add a pooling layer at the end of ResNet and generate 2048 dimensional 2D features.

Dynamic features:

ResNet's ability to extract dynamic features is limited, despite its ability to generate visual features in still images. Each object's motion information can be well described by dynamic features. We improve the recording of dynamic features by extending the two-dimensional neural network to a three-dimensional convolutional neural network.

When compared to the C3D network, the 3DResNet network can alleviate the degradation of the deep network model sufficiently. Given that 3DResNet has 18 layers, 34,101 layers, and other hierarchical structures, we study and test the 3DResNet of each layer before selecting 3DResNet -101 to extract dynamic features for VMSG. To better extract dynamic features, we train 3DResNet on the Kinetics dataset.

Category features:

In the video feature ablation experiment, we discover that the video category information contains information that is useful for the generation of video captions. For example, if the object is a music video, the audio weight should be increased appropriately. The object is a motion video, and the visual weight should be increased. To extract video classification information, we use a 3DResNet network and a full connection layer, and we train 3DResNet on the Kinetics dataset. The dataset has 400 categories, including small categories in sports, film, food, and other fields, and the level of detail has been greatly improved. Furthermore, because tags can be generated independently, generalization performance has improved.

Audio features:

We use a commonly used audio feature - Mel Frequency Cepstrum Coefficient - to make good use of the original audio features (MFCC). To extract the features of uniformly sampled 1-second audio clips, PyAudioAnalysis [9] is used. The actual audio representation is made up of the average and standard deviation of 34 different audio features.

After the above processing, the 2D feature representation $\{v_i^a\}_{i=1}^N$ of the video, the dynamic feature representation $\{v_i^m\}_{i=1}^N$, the category feature $\{v_i^c\}_{i=1}^N$ and the audio feature representation $\{v_i^v\}_{i=1}^N$. Connect different video features frame by frame to form a frame representation: $v_i = [v_i^a; v_i^m; v_i^c; v_i^v]$.

2.3 Phrase coding module

Some words, such as 'is' and 'the,' have no meaning when used alone. When used alone, some words do not have a clear meaning. For example, when the words 'woman' and 'cap' are combined to form 'woman with cap,' the meaning becomes clearer. As a result, when we make a semantic grouping, we use phrases rather than individual words.

To build the model's semantic phrases, we must generate appropriate words and phrases from some generated abstracts. To accomplish this goal, we must discover the relationships between words. When the t-th word w_t of the caption is generated, there is a word representation matrix $W_t = [E[w_1] \mathrm{L}\ E[w_{t-1}]]^T \in ¡^{(t-1)\times d_w}$, where E represents a word embedding matrix. We use the phrase encoder ϕ^p to generate the phrase representation matrix $P_t = [p_{1,t} \mathrm{L}\ p_{t-1,t}]^T \in ¡^{(t-1)\times d_w}$ by using the word representation matrix W_t, which is given as

$$P_t, A_t = \phi^p(W_t), \tag{1}$$

where $A_t = [a_{1,t} \mathrm{L}\ a_{t-1,t}]^T \in ¡^{(t-1)\times(t-1)}$ is the word attention matrix. The weight of $\{w_i\}_{i=1}^{i=t-1}$ is $a_{j,t} \in ¡^{t-1}$. For the encoder ϕ^p, we use the self-attention mechanism module [10] proposed by Vaswani et al., which can well model the dependency between words in sentences.

2.4 Semantic grouping module

A phrase serves as the foundation for a semantic grouping, which is made up of phrases and all semantically related features. The number of candidate phrases equals the number of words, and many of them are very similar. As a result, phrase suppressors are used by the model to filter out these phrases. The semantic aligner will semantically align the video frame with the phrase once the model has obtained all of the available phrases.

To keep those phrases with meaning and low coupling, the model must use a

phrase filter to determine which phrases to discard based on their similarity. In this paper, we calculate similarity using the attention matrix of phrases, which is given as

$$R_t = A_t (A_t)^T \quad , \tag{2}$$

where $r_{i,j,t}$ denotes the similarity between $p_{i,t}$ and $p_{j,t}$. We set a threshold τ and if $r_{i,j,t}$ is more significant than this threshold, it will be determined that the two phrases are related. After getting two associated phrases, compare the similarity between the two phrases, and all phrases and the party with the most significant value will be discarded. If $\sum r_{i,k,t} > \sum r_{j,k,t}$, then $p_{i,t}$ will be discarded. Table 1 shows the detailed phrase filter mechanism.

Table 1

The detailed phrase filter mechanism

Algorithm 1 Phrase filter

Input: phrase $P = \{p_1, \mathrm{L}\ , p_k\}$, a word attention matrix A and a threshold τ

output: filtered phrase set $\hat{P} = \{\hat{p}_1, \mathrm{L}\ , \hat{p}_k\}$

function phrase Filtering(P,A,τ)

$\hat{P} \leftarrow P$

$R \leftarrow AA^T$

for $r_{i,j} \in \{r_{i,j} \mid r_{i,j} \in R, r_{i,j} > \tau\}$ do

if $\sum_k r_{i,k} > \sum_k r_{j,k}$ then

$\hat{P} \leftarrow P \setminus \{p_i\}$

else

$\hat{P} \leftarrow P \setminus \{p_j\}$

end if

end for

return $\hat{P}$

end function

2.5 Single layer LSTM decoding module based on semantic attention

For encoding and decoding, the traditional video caption model typically employs a two-layer LSTM. The two-layer LSTM network, however, increases the model's parameters, increasing the difficulty of training. As a result, we adopt a single-layer LSTM network. To simplify the model, the network has one embedding layer, one bidirectional layer, and two full connection layers.

Figure 4 shows the working mechanism of semantic attention. The contribution of each semantic group $s_{i,t}$ in generating the t[th] word w_t is different. For example, when generating "people" in "a man is talking to a group of people", two semantic groups, "a man is talking" and "a ground of" will be generated. For these two semantic groups, "a group of" is more important to "people".

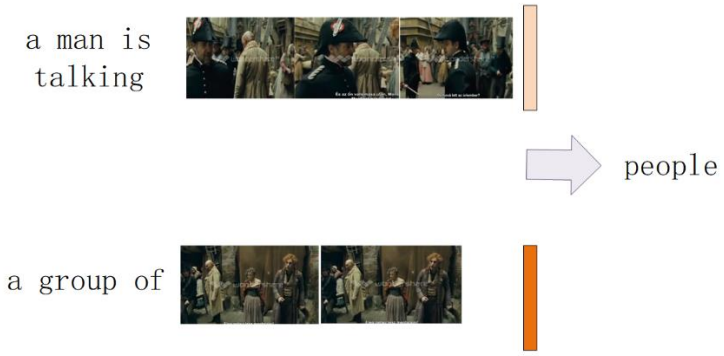

Figure 4. Working mechanism of semantic attention

The attention mechanism has three parameters Q, K, and V. Q is the semantic group $s_{i,t}$ of this paper, K and V are h_{t-1}. Considering that the expression ability of the linear model is not enough, we add a hyperbolic tangent function σ to the attention module. The final form is shown in Eq. (3), where u_s, U_d, H_d and b_d are learnable parameters and $\beta_{i,t}$ is the weight of each semantic group. The equation is given as

$$\beta_{i,t} \propto u_s^T \sigma\left(U_d h_{t-1} + H_d v_j + b_d\right) \quad . \tag{3}$$

After adding semantic attention, the decoder in this paper will assign a score to each semantic group according to the state function h_{t-1} of the previous decoder. After getting the weight of the semantic group, the weighted average gets x_t, which is given as

$$x_t = \sum_{i=1}^{M_t} \beta_{i,t} s_{i,t} \quad . \tag{4}$$

Then x_t is output to LSTM. The possible probability of the next word consists of a full connection layer and a Softmax layer, which is given as

$$h_t = LSTM\left(\left[x_t; E\left[w_t - 1\right]\right], h_{t-1}\right) \quad , \tag{5}$$

$$p\left(w_t \mid V, w_1, \mathrm{L}\ , w_{t-1}\right) = softmax\left(U_h h_t + b_h\right) \quad , \tag{6}$$

where U_h and b_h are learnable parameters. The decoder in this paper is similar to the traditional decoder. The difference is that it changes the frame into a semantic group and reduces the number of layers of LSTM.

Table 2 shows the detailed semantic attention algorithm.

Table 2

The detailed semantic attention mechanism

Algorithm 2. Semantic attention

Input: semantic group $s_{i,t}$, status function h_{t-1}

Output: weighted average semantic group x_t

1: function semantic Attention ($s_{i,t}$, h_{t-1})

2: $\beta_{i,t} \leftarrow u_s^T \sigma\left(U_d h_{t-1} + H_d v_j + b_d\right)$

3: for ($i : M_t$)do

4: $x_t \leftarrow \beta_{i,t} s_{i,t} + x_t$

5: end for

6: return x_t

7: end function

2.6 Loss function of the model

The key to training the model is to generate a distinct and coherent semantic information group. To ensure the generation of semantic groups, the phrase suppressor will filter out redundant phrases. This paper employs the typical cross-entropy loss function Lce and the comparative attention loss function Lca to train the model. When a video v and its standard translation $Y = [y_1, \mathrm{L}\ , y_T]$ is given, its loss function can be obtained as

$$L = Lce + \lambda Lca. \tag{7}$$

Cross-entropy loss is defined as the negative logarithmic probability of producing the correct title:

$$L_{ce} = \sum_{(V,Y)\in D} \sum_{t} (-\log p(y_t | V, y_{1,} \mathrm{L}\ , y_{t-1})) \tag{8}$$

To ensure that the semantic group's members have a consistent meaning, a semantic group should only contain frame information that is highly related to the semantics. In this paper, we use another group of videos with low correlation as the semantic alignment module's incorrect candidate. To ensure low correlation, we choose a control video at random from a group of videos with completely different abstracts. From equation 3, the positive correlation coefficient $\alpha_{i,j,t}^{pos}$ and the negative correlation coefficient $\alpha_{i,j,t}^{neg}$ of the input video frame f_j and the phrase $\hat{p}_j$ can be obtained.

After obtaining the positive and negative correlation coefficient, we use softmax for normalization. $p_{ca}(s_i,t) = \sum_{j=1}^{N} \alpha_{i,j,t}^{pos}$ represents the probability that there is no control frame in the semantic group. $p_{ca}(s_i,t)$ increases with the increase of positive correlation coefficient relative to negative correlation coefficient. M_t indicates the number of semantic groups. The equation is given as

$$L_{ca} = \sum_{(V,Y)\in D} \sum_{t} \sum_{i}^{M_t} (-\log p_{ca}(s_{i,t})) \tag{9}$$

3 Experimental results and analysis

This section first describes the dataset's characteristics, then provides the evaluation criteria and thoroughly analyzes the experimental results.

3.1 Dataset

We train and test our MGVC in MSR-VTT [11] throughout the experiment. In the field of video caption, MSR-VTT is a critical dataset. It specifies the video category as well as the video's audio characteristics. MSR-VTT has 10,000 online videos totaling 41.2 hours in 20 different classes. AMT staff created 20 video captions for each online video.

During the experiments, we find that some issues with the dataset's video, such as word spelling errors and audio information that cannot be used. Despite the fact that the total number of words in the video caption is 23667, 10040 of

them appear only once. Furthermore, when all of the words are compared to Wikipedia's vocabulary, we find that 836 words do not exist, owing to spelling errors. It makes the model's training and testing are complicated.

The dataset's video includes audio features that can be used to generate a video caption. However, because approximately 13% of videos lack audio information, the experiment is complicated.

More than 90% of the videos are under 30 seconds long, and 90% of the video captions are under 16 words long. As a result, we take 30 frames evenly, which allows us to better characterize the video features while keeping the data size manageable.

3.2 Evaluating criteria

We use four criteria to evaluate the model: BLEU [12], METEOR [13], ROUGE-L [14], CIDEr [15].

BLEU: BLEU (Bilingual Evaluation understudy), proposed by IBM in 2002, is an overall video caption evaluation criterion. It represents the likelihood of the occurrence of n-word lengths in the text of video caption and standard translation. The result of video caption can be expressed as $C=\{c_1,c_2,\mathrm{L}\ c_N\}$ and the data scale is N. The corresponding reference translation $S_i=\{s_{i1},s_{i2},\mathrm{L}\ s_{im}\}$, in which m represents the number of reference sets. N in the equation represents the word length, w_k represents the k$^{\text{th}}$ caption result, and the word phrase with length n in the reference translation. $h_k(c_i)$ represents the total number of occurrences of w_k in c_i and $h_k(c_{ij})$ represents the number of occurrences of w_k in standard translation s_{ij}. The corresponding statement BLEU can be calculated according to

$$CP_n(C,S)=\frac{\sum_i\sum_k \min(h_k(c_i),\max_{j\in m} h_k(s_{ij}))}{\sum_i\sum_k h_k(c_i)} \tag{10}$$

From the above equation, we can see that the lower the length of machine translation, the higher the final score, so the length penalty factor $b(C,S)$ （BP(Brevity Penalty) ）is added, which is given as

$$b(C,S)=\begin{cases}1 & \text{if } l_c>l_s\\ e^{1-\frac{l_s}{l_c}} & \text{if } l_c\le l_s\end{cases}. \tag{11}$$

BLEU is the weighted average of the accuracy of all statements, which is given

as

$$BLEU(C,S)=b(C,S)\exp(\sum_{n=1}^{N}w_n\log CP_n(C,S)) \tag{12}$$

METEOR: METEOR, which Lavir proposed in 2004, is a weighted average based on single-precision and recall rate. In comparison to BLEU, the research shows that METEOR is more similar to the results judged by people themselves. METEOR is defined specifically as follows:

$$Pen=\gamma(\frac{ch}{m})^{\theta},\quad F_{mean}=\frac{P_mR_m}{\alpha P_m+(1-\alpha)R_m},\quad P_m=\frac{|m|}{\sum_k h_k(c_i)},$$

$$R_m=\frac{|m|}{\sum_k h_k(s_{ij})},\quad METEOR=(1-Pen)F_{mean}, \tag{13}$$

where α, γ and θ are set default parameters, m is a given set of calibrations and ch is a continuous and orderly statement block.

ROUGE-L: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a common video caption criterion that is primarily based on recall rate. The letter L in ROUGE-L stands for the longest common subsequence (LCS). ROUGE-L employs the longest common subsequence of video caption C and standard caption S, which is give as

$$R_{LCS}=\frac{LCS(C,S)}{len(S)},\quad P_{LCS}=\frac{LCS(C,S)}{len(C)},\quad ROUGE-L=\frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}}. \tag{14}$$

CIDEr: CIDEr (Consensus-based Image Description Evaluation) considers each sentence to be a document and computes the cosine angle of the TF-IDF vector. The number of times a specific phrase appears in a sentence is represented by TF. The acronym IDF is used to emphasize the significance of a phrase as

$$TF(k)=\frac{h_k(c_i)}{\sum h_l(c_i)},\quad IDF(k)=\log(\frac{N}{\sum_{i=1}^{N}\min(1,\sum_{j=1}^{m}h_k(S_{ij}))}),$$

$$g_k(c_i)=TF(k)*IDF(k), \tag{15}$$

where $g_k(c_i)$ represents the TF-IDF vector of w_k. $h_k(c_i)$ indicates the number of times w_k appears in the sentence c_i. $\sum h_l(c_i)$ indicates the number of all n-length words in c_i. $\sum_{i=1}^{N}\min(1,\sum_{j=1}^{m}h_k(S_{ij}))$ indicates the number of sentences contained w_k in the reference translation of the dataset.

According to the cosine angle, the similarity between sentences is obtained as

$$CIDEr_n(c_i,S_i)=\frac{1}{M}\sum_{j=1}^{M}\frac{g^n(c_i)\cdot g^n(S_{ij})}{\|g^n(c_i)\|\times\|g^n(S_{ij})\|}. \quad (16)$$

After obtaining the results of all sentences, we calculate the $CIDEr_n(c,S)$ of all n lengths and take the average result as

$$CIDEr(c,S)=\frac{1}{N}\sum_{n=1}^{N}CIDEr_n(c,S), \quad (17)$$

where $g^n(c)$ is the transpose of all matrices of $g_k(c_i)$, $g^n(S_{ij})$ is the matrix of all w_k TF-IDF vectors in the standard translation.

CIDEr is identical to ROUGE and BLEU. It only targets the words in the sentence, not the sentence's semantic information.

3.3 Experimental setup

First, we sample each input video uniformly, and each video samples 30 frames of images and 30 video clips. The video clip is made up of frames that surround the video clip. We can extract the 2D and dynamic features of the video as well as the audio features from the one-second clip at the beginning of these video frames using these 30 frames of images and video clips as the input of ResNet and 3DResNet.

Because VMSG is a multimodal input, multimodal input will inevitably result in increased input dimension, which dramatically reduces hardware requirements. As a result, we use a full connection layer to reduce feature dimension. The initial dimensions of 2D and 3D features are 2048, while audio features and classification labels are 1.

We fed the sampled video frames into the 3DResNet network after it had been trained on the Kinetic dataset. To encode the label and feed it into LSTM, we use a single hot coding method. We initialize the word embedding matrix with a glove, set embedding_ size to 300, and train it with the entire model. Before producing the first word, we use <SOS> as the caption's beginning, then ignore it, τ and λ are set to 0.2 and 0.16, respectively.

A thesaurus is required to generate a sentence or a word. With a total of 23667 words, the model thesaurus is entirely derived from the video captions of the training and test sets in MSR-VTT. We set dropout to 0.5 during training to reduce overfitting. Adam optimizer is used to optimize the model, and the initial learning rate is set to 0.0005. The MSR-VTT dataset videos 6513, 497, and 2990 are used to train, verify, and test the model in this paper. The batch size is set to 100, and the cycle speed is set to 50. To evaluate the model, we use

Microsoft Coco's official code. The maximum caption length is set at 15 characters.

3.4 Ablation Experiment

(1) C3D and 3DResNet

When the problems of gradient disappearance and gradient explosion in the deep network are considered, 3DResNet can solve gradient disappearance and gradient explosion better than the C3D network, so the effect of extracting image features in deep 3DResNet is better than C3D. On the ActivityNet dataset, 3DResNet-18 outperforms C3D. Table 3 shows that 3DResNet- 34 outperforms C3D on various datasets.

Table 3

Accuracy of C3d and 3D resnet-34 in extracting features on each data set

Model	ASLAN	Sports1M	UCF101	HDMB51
C3D	78.3	61.1	82.3	51.6
3D ResNet-34	**78.8**	**65.6**	**85.8**	**54.9**

We use C3D, 3D ResNet-34, and 3D ResNet-101 to extract dynamic features to form multimodal features, and single-layer LSTM for decoding. Table 4 displays the results. Where C stands for C3D, R34 stands for 3D ResNet-34, and R101 stands for 3D ResNet-101. The table shows that 3D ResNet-101 has the best effect, outperforming C3D in BLEU4, METEOR, and CIDEr. In this paper, we use 3D ResNet-101 as the dynamic feature extraction model after conducting experimental comparisons. The experimental results are depicted in Figure 5. The experimental results of 3D ResNet-101 are more accurate and better predict the "person".

Table 4

Comparison of experimental criteria between C3d and 3D RESNET

Model	BLEU4	METEOR	ROUGE-L	CIDEr
Multi(C)	40.4	27.8	60.6	45.9
Multi(R34)	40.3	27.8	60.4	46.1
Multi(R101)	40.5	28.0	60.6	46.4

Multi(C): A man is drawing
Multi(R34): A person is drawing
Multi(R101): A person is drawing
Ground Truth: A person draws a cartoon character on a piece of paper

Figure 5. Experimental comparison between C3D and 3D ResNet

(2) Multimodal and semantic grouping module

We conduct ablation experiments on each module to evaluate the effectiveness of each module in the multimodal semantic grouping, and the results are shown in Table 5.

Table 5

Ablation experiment of multimodal and semantic grouping module (*Red* indicate the best performance.)

Multi	SA	PS	CA	BLEU4	METEOR	ROUGE-L	CIDEr
×	×	×	×	36.2	25.9	58.7	41.5
×	√	×	×	37.3	27.6	59.7	46.7
×	√	√	×	37.5	27.8	59.7	47.4
×	√	√	√	39.7	28.1	60.0	48.5
√	√	√	√	**40.7**	**28.2**	**60.8**	**49.0**

Note: use 2D and 3D features before using multimodal features.

Multi represents multimodal features, which enrich the model's extracted features. SA is a semantic aligner (including phrase coding) that allows video frames with similar semantics to be combined to form a semantic group. PS is a phrase filter that can generate semantically relevant words. CA indicates the contrastive attention loss, which promotes accurate alignment of semantic words and video features and improves the model's ability to form a semantic group. According to the table, the performance improved by SA is the most outstanding, while the range improved by PS is the smallest. SA and CA do a better job of combining adjacent frames into a semantic group, and PS generates semantic words that correspond to the semantic group. In comparison to forming adjacent features into a semantic group, the effect of generating semantic words in the semantic group is subtle. The reason for this is that both SA and CA promote the formation of a semantic group directly, whereas PS indirectly promotes the formation of a semantic group. Multimodal video features can significantly improve the model's performance by enriching the video information contained in the encoder. Figure 6 depicts the experimental results. VMSG experimental results are more precise and accurately generate "apply makeup" when compared to other experimental results.

Baseline: A woman is making up
SA+PS: A woman is putting on her makeup
SA+PS+CA: A woman is putting on her makeup
VMSG: A woman is showing how to apply makeup
Ground Truth: A woman applying make up to her eyes and eye brow

Figure 6. Ablation experiment of each module of semantic group

(3) Semantic attention mechanism

To investigate the effect of the semantic attention mechanism, we conduct experiments with and without the semantic attention mechanism, as well as 50 cycles of iterative training. Table 6 displays the results. VMSG-WA denotes that

the semantic attention mechanism is not added, whereas VMSG denotes that the semantic attention mechanism is added. As shown in the table, VMSG outperforms VMSG-WA in four criteria.

VMSG has improved in four criteria when compared to VMSG-WA. VMSG has improved slightly in BLEU4, METEOR, and ROUGE-L, and has increased by 2% in CIDEr. CIDEr represents the model's ability to grasp key points. The attention mechanism assigns a weight value to each semantic group. As a result, the weight value of the semantic group with high importance is high, which improves the model's ability to grasp the key points. The experimental results are depicted in Figure 7. The attention mechanism improved the focus of the VMSG model's generated caption, replacing "cooking" with "making a dish".

Table 6

Ablation experiment of semantic attention

Model	BLEU4	METEOR	ROUGE-L	CIDEr
VMSG-WA	40.0	28.0	60.7	47.9
VMSG	40.7	28.2	60.8	49.0

VMSG-WA: A woman is cooking in the kitchen
VMSG: There is a woman is making a dish in the kitchen
Ground Truth: In the kitchen, there is a women working with mixer grinder

Figure 7. Comparison after adding attention mechanism

3.5 Comparison with stat-of-the-arts

This section compares the performance of our model to that of the state-of-the-art models. Table 7 displays the results. In the MSR-VTT dataset, it can be seen that VMSG has progressed to the advanced level. VMSG took first place in METEOR and CIDEr, ahead of second place by 3% and 1%, respectively. In the other two metrics, BLEU4 and ROUGE-L, VMSG reaches the current advanced level. Therefore, on the whole, VMSG is the state-of-the-art video caption generation model at present.

METEOR represents the semantic correctness of the model-generated abstract, while CIDEr represents the ability to extract key information from the model. We believe that the semantic group can eliminate redundant information, resulting in less interference information. in addition, we include a semantic attention mechanism to increase the weight of the semantic group, highlight key points, and improve the model's ability to extract key information and extract semantically correct results.

The model's accuracy and recall rate are represented by BLEU4 and ROUGE-L,

respectively. VMSG makes use of multimodal input. We use ResNet-152 to extract 2D features and 3DResNet-101 to replace the C3D network, which can extract dynamic video features better. We also include audio features, which allow the model to extract enough information via classification features. Meanwhile, the semantic grouping and attention techniques used in this paper reduce redundant information from adjacent frames. As a result of these factors, the VMSG model's abstraction accuracy and recall rate reach an advanced level.

Table 7

Comparison of VMSG and several state-of-the-art methods on BLEU4, METEOR, ROUGE-L, and CIDEr. *Red* and *blue* indicate the best and second-best performance.

Model	Year	BLEU4	METEOR	ROUGE-L	CIDEr
SA-LSTM[2]	2018	36.3	25.5	58.3	39.9
M3 [16]	2018	38.1	26.6	-	-
RecNet[5]	2018	39.1	26.6	59.3	42.7
PickNet*[16]	2018	41.3	27.7	59.8	44.1
SibNet[17]	2019	40.9	27.5	60.2	47.5
MGSA[18]	2019	42.4	27.6	-	47.5
MARN[19]	2019	40.4	28.1	60.7	47.1
OA-BTG[20]	2019	41.4	28.2	-	46.9
Two-stream[21]	2020	39.7	27.0	-	42.1
STG-KD[22]	2020	40.5	28.2	60.9	47.1
RMN[23]	2020	39.2	27.8	59.9	46.7
BilSTM-CG[24]	2020	39.1	27.7	59.9	46.4
VideoTRM[25]	2020	38.8	27.0	-	44.7
NA-B[26]	2021	40.4	28.0	-	47.6
RAE[27]	2021	40.0	28.0	60.0	45.7
SGN(R152)[28]	2021	39.6	27.6	59.6	45.2
VMSG(our)	-	41.0	28.2	60.8	49.0

SA-LSTM: A cartoon character is walking out of a car and talking to a man
VMSG: Cartoon characters are interacting
Ground Truth: A cartoon is shown of a girl running away from a murderer

Figure 8. Experimental results of the 7743th video

SA-LSTM: A man plays a video game
VMSG: A person is playing a video game
Ground Truth: Someone playing a fun little video game

Figure 9. Experimental results of the 8235th video

Figures 8 and 9 show an example of SA-LSTM and VMSG generating titles, with VMSG outperforming SA-LSTM in terms of accuracy. As shown in Figure 8, VMSG can generate the subject who is acting in the video scene. VMSG predicts a group of cartoon characters rather than just one, and the content is more accurate as a result. Meanwhile, its ability to extract critical information has improved. In general, VMSG outperforms SA-LSTM.

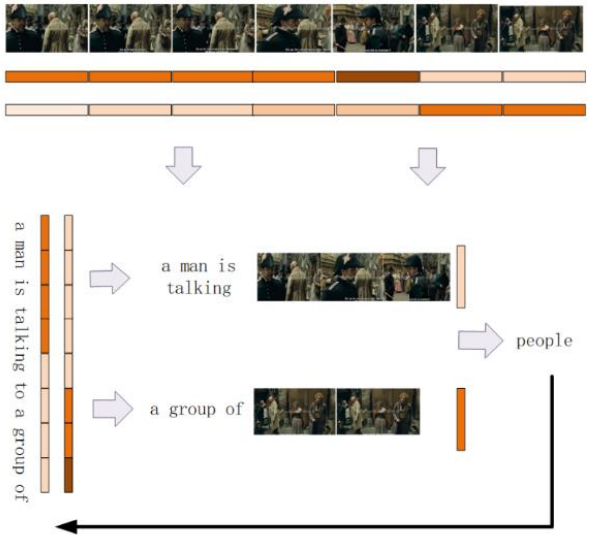

Figure 10. Formation process of semantic groups of the 9234th video

(Note: the orange shades represent the size of semantic weight)

Figure 10 constructs the phrases "a man is talking" and "a group of" from the words in the partially decoded title "a man is talking to a group of". Collecting a man's speech and a group of people resulted in the formation of one semantic group. More data from the latter semantic group predicts the following word, "people". The findings show that VMSG can form semantic phrases and associate image frames with semantic phrases.

4 Conclusion

In this paper, we propose a semantically grouped multimodal video description method that employs multimodal feature fusion based on 2D and 3D features, as well as tag and audio features. To extract dynamic features from videos, VMSG employs 3DResNet rather than C3D. We compose video frames with the same semantics

into a semantic group for decoding to solve the problem of redundant information in adjacent video frames. Because the importance of different semantic groups varies, we investigate a semantic attention algorithm that assigns weight to semantic groups. Finally, to simplify the model, we employ a single-layer LSTM. On MSR-VTT, VMSG achieves good results. Our long-term goal is to develop better multimodal models that take into account video decoding using non-autoregressive methods.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (61573182) and by the Fundamental Research Funds for the Central Universities (NS2020025).

Conflict of Interest:

Xin Yang declares that he has no conflict of interest, declares that he has no conflict of interest. Chen Zhu declares that he has no conflict of interest, declares that he has no conflict of interest. XX declares that he has no conflict of interest.

Reference:

[1] Liu D, Zha Z J, Zhang H, et al. Context-aware visual policy network for sequence-level image captioning[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 1416-1424.

[2] 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. 2018: 7622-7631.

[3] Wang J, Jiang W, Ma L, et al. Bidirectional attentive fusion with context gating for dense video captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7190-7198.

[4] ZHA Z-J, LIU D, ZHANG H, et al. Context-aware visual policy network for fine-grained image captioning[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[5] Chen Y, Wang S, Zhang W, et al. Less is more: Picking informative frames for video captioning[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 358-373.

[6] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns

and imagenet?[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

[9] GIANNAKOPOULOS T. pyaudioanalysis: An open-source python library for audio signal analysis[J]. PloS one, 2015, 10(12): e0144610.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[11] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5288-5296.

[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. 2002: 311-318.

[13] 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. 2005: 65-72.

[14] ROUGE L C Y. A package for automatic evaluation of summaries[C]//Proceedings of Workshop on Text Summarization of ACL, Spain. 2004.

[15] 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. 2015: 4566-4575.

[16] Wang J, Wang W, Huang Y, et al. M3: Multimodal memory modelling for video captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7512-7520.

[17] Liu S, Ren Z, Yuan J. Sibnet: Sibling convolutional encoder for video captioning[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(9): 3259-3272.

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

[19] Pei W, Zhang J, Wang X, et al. Memory-attended recurrent network for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8347-8356.

[20] ZHANG J, PENG Y. Object-aware aggregation with bidirectional temporal graph for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[21] ZHANG K, LI D, HUANG J, et al. Automated video behavior recognition of pigs using two-stream convolutional networks[J]. Sensors, 2020, 20(4): 1085.

[22] 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, 7.

[23] LIU D, QU X, DONG J, et al. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network[C]//Proceedings of the 28th International Conference on Computational Linguistics.

[24] CHEN S, ZHONG X, LI L, et al. Adaptively Converting Auxiliary Attributes and Textual Embedding for Video Captioning Based on BiLSTM[J]. Neural Processing Letters, 2020, 52(3): 2353-2369.

[25] PAN Y, XU J, LI Y, et al. Pre-training for Video Captioning Challenge 2020 Summary[J].

arXiv preprint arXiv:2008.00947, 2020.
[26] Yang B, Zou Y, Liu F, et al. Non-autoregressive coarse-to-fine video captioning[J]. arXiv preprint arXiv:1911.12018, 2019.
[27] ZHU M, DUAN C, YU C. Video Captioning in Compressed Video[J]. arXiv preprint arXiv:2101.00359, 2021.
[28] RYU H, KANG S, KANG H, et al. Semantic Grouping Network for Video Captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence.