Journal of Contemporary Issues in Business and Government Vol. 28, No. 04, 2022
https://cibgp.com/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

VIDEO CAPTIONING WITH SPATIAL-TEMPORAL ATTENTION MECHANISM (STAT)

T. SARITHA, Associate Professor, sarithathota006@gmail.com, Rishi UBR Women's College, Kukatpally, Hyderabad 500085

Abstract

Video captioning refers to automatic generate natural language sentences which summarize the video contents. Inspired by the visual attention mechanism of human beings, temporal attention mechanism has been widely used in video description to selectively focus on important frames. However, most existing methods based on temporal attention mechanism suffer from the problems of recognition error and detail missing, because temporal attention mechanism cannot further catch significant regions in frames. In order to address above problems, we propose the use of a novel spatial-temporal attention mechanism (STAT) within an encoder-decoder neural network for video captioning. The proposed STAT successfully takes into account both the spatial and temporal structures in a video, so it makes the decoder to automatically select the significant regions in the most relevant temporal segments for word prediction. We evaluate our STAT on two well-known benchmarks: MSVD and MSR-VTT-10K. Experimental results show that our proposed STAT achieves the state-of-the-art performance with several popular evaluation metrics: BLEU-4, METEOR and CIDEr.

I. INTRODUCTION

THE goal of video captioning is to make computer understand what is happening in a given video and establish the bridge between the video content and its meaningful natural language description[1], [2], [3], [4], [5], [6], [7], [8], [9]. It has been extensively used to many applications, for instance, it may help users of video sites to retrieve videos efficiently, or benefits visually impaired people for better understanding of the video content. However, since a video clip often involves complex interactions of actors and objects that evolve over time, it is still a challenging task to automatically generate an accurate description for the complicated sequence of events. Video captioning has become an active and flouring research topic in recent years, and the most effective method is the encoder-decoder neural networks, in which 2D or 3D convolutional neural networks (CNNs) are utilized for video content encoding and recurrent neural networks (RNNs) for decoding to a sentence. Early work with this method attempted to encode an entire video into a single feature vector, which is prone to clutter, because temporally distinct events and objects will be potentially fused incoherently. In order to address the problem, recently, some work incorporated

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

temporal attention mechanism (TAT) [10], [11], [12], [13] into the encoderdecoder network to exploit the temporal structure underlying the video. Rather than encode an entire video into a single feature vector, temporal attention mechanism can make the decoder selectively focus on a small subset of frames by attention weights. Therefore, each word is generated according to the most relevant frames.

Although temporal attention mechanism can select important frames for word prediction, it cannot further catch significant regions in these frames. When there is clutter in some frames, this drawback can result in the problems of detail missing and object misprediction. As shown in Fig.1. (a), the 'bear' is mispredicted as 'panda'; as shown in Fig.1. (b), the 'phone' is missed. When we are asked to describe the main content of a video, we have a large possibility to only mention the semantic of the regions of significance in some frames. Hence, besides temporal structure in an entire video, exploiting spatial structure in each frame is also necessary for generating more accurate and detailed description.

Fig. 1. Illustration of problems of the object misprediction ((a): misrecognizing 'bear' as 'panda') and detail missing ((b): missing 'phone'), where TAT represents temporal attention mechanism.

To take full advantage of the spatial and temporal structure in a video, a novel spatial-temporal attention mechanism (STAT) is proposed in this work, which is our major contribution. Furthermore, global features and motion features, which are both at frame-level, are widely used in previous work. In this paper, in addition to global and motion features, we expand the feature set to include the local features at object-level, which can represent accurate semantic concepts of objects in each frame. While generating a description, the proposed STAT not only selectively focuses on important frames, but also further catches significant regions in those frames, making it possible for the decoder to obtain enough input information and conduct accurate decoding. Therefore, our proposed STAT can generate more relevant video description sentences.

The rest of this paper is organized as follows. Firstly, it reviews the existing materials in Section II. Then, it presents the overview of our framework and expounds

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

on the details of STAT in Section III. After that, it gives the evaluation of STAT on different dataset and analyzes the performance in Section IV. Finally, it comes to the conclusion, together with the discussions for future work.

II. RELATED WORK

In this work, we provide relevant background on previous work on video captioning, temporal attention mechanism and exploitation of local features.

Video Captioning: Up to now, some methods have been proposed for addressing the problems of video captioning, and these approaches have proved to be significant progress. These methods could be roughly classified into three types, depending on the manner in which the sentences are generated. (1) template-based method; (2) the method based on neural network.

Template-based method [14], [15], [16], [17], [18] firstly predicts semantic concepts or words (e.g., subjects, objects and verbs) by different classification methods, then employs a pre-defined sentence template to form them into a description. This method is intuitive, but need to deal with the complex data. Meanwhile, the limitation of sentence template cannot flexibly generate meaningful sentences [5].

Temporal Attention Mechanism: The early work on encoder-decoder framework [21] tried to represent the information using a single, temporarily folded feature vector, which is likely to lead to confusion, because different events and objects in temporal sequence may fuse incoherently [12]. Therefore, a captioning model should be clever enough to exploit the temporal structure underlying the video sufficiently.

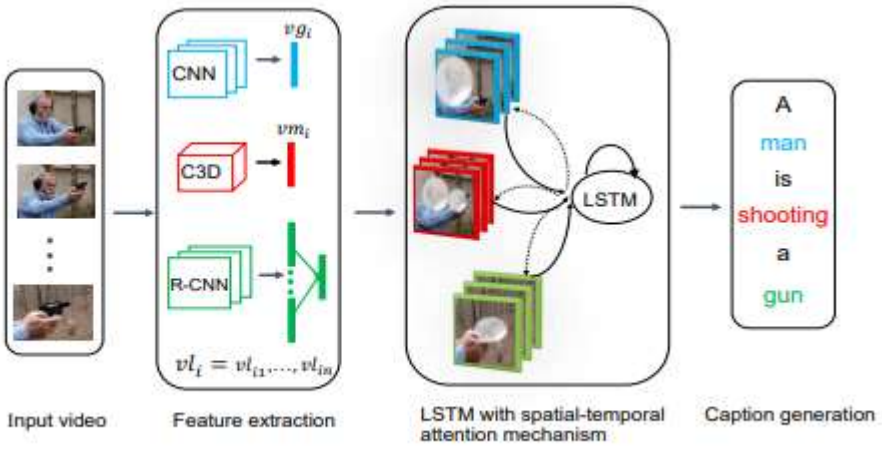

Fig. 2. Illustration of our proposed STAT video caption framework based on spatial-temporal attention mechanism.

III. EXPLOITING SPATIAL-TEMPORAL STRUCTURE IN VIDEO CAPTIONING

In this part, we deeply study the main contributions of this paper and put forward

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

an approach of using the spatial and temporal structure for video captioning.

A. Overall Framework

To begin with, we describe our overall framework, based on the popular ConvNet + LSTM architecture[32], [33], [34], [35], [36], [37], [38]. It mainly consists of three important processes as shown in Fig. 2.

First of all, we use 2-D/3-D Convolutional Neural Network (CNN) and Region-based Convolutional Neural Networks (RCNNs) to encode the video inputs to a set of fixed length vector representation. Deep convolutional neural networks (CNNs) and regions with convolutional neural networks (RCNNs) have been found to be useful in computer vision tasks such as object localization and detection. For example, 2-D CNN such as GoogleNet can represent an image as a single feature vector. 3-D CNN such as C3D can represent consecutive frames as a single feature vector. R-CNN such as Faster R-CNN can represent a region or object as a single feature vector. The feature vector is from upper or intermediate layers of a CNNs as a high-level feature for vision [12]. Most of state-of-the-art work therefore exploits a pre-trained 2-D/3D CNNs and R-CNNs as an encoder. We will follow the implementation of these work to exploit CNNs and R-CNNs as the encoder.

Secondly, we fuse three kinds of features via two-stage attention mechanism. In this phase, when the current semantic context is given, spatial attention mechanism firstly makes the decoder to select local features with more spatial attention weights, which represent the significant regions. Then, temporal attention mechanism make the decoder to select global and motion features, as well as local features representing significant regions. Finally, three types of features are fused to represent the information of key frames.

In the end, we use the language model to generate sentences with dynamic temporal features. In this paper, LSTM is used as a language model, because it can effectively capture longterm sequence information.

B. Encoder: Convolutional Neural Network

The encoder network is designed for learning a proper representation of visual information. Then the decoder network can form corresponding sentences according to the output of the encoder. In the encoder network, we extract global feature and local feature from every frame, and extract the motion feature from video clips. Hence, a video inputted to encoder network will finally be transformed into a sequence of visual information: V = {v1, ..., vk }, where each vi = {vgi , vli , vmi }, and k refers to number of frames. The vgi extracted from 2-D CNN represents one of the global features which contain the context information of a video frame. The vmi extracted via 3-D CNN represents one of the motion features which characterize punctuated actions. The vli extracted via R-CNN represents one of the local features of maintaining more accurate object information.

C. Spatial-Temporal Attention Mechanism

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

As far as primates and human beings are concerned, visual attention mechanism is a significant mechanism. Thus, we exploit visual attention mechanism of human beings to design a spatial-temporal attention mechanism (STAT) method. When human beings are asked to describe a certain content of video, they do not describe everything in a video. Instead, they tend to talk more about semantically more important spatialtemporal segments in the video. Thus, we argue that before generating each target word, the decoder firstly should exploit spatial structure to catch semantically most relevant objects on each frame. Then, it should exploit temporal structure to track their trajectories and studies the interactions among them on consecutive frames. We will elaborate the details of proposed two-stage attention mechanism in Section D and Section E.

D. Exploiting spatial Structure: A Spatial Attention Mechanism

Since a video has multiple objects, the decoder should selectively focus on the most significant regions of a video sequence. We use spatial attention mechanism to dynamic weighted sum of the top-n local features, i.e. vli = {vli1, ..., vlin}, to obtain a single spatial local feature Ψi(V L) on each frame such that

$$\Psi_i^{(t)}(VL) = \sum_{j=1}^{n} \alpha_{ij}^{(t)} vl_{ij}, \qquad (1)$$

where α (t) i j represents spatial attention weights. It is calculated at each time and Ín j=1 α (t) i j = 1. When the previous words are given, i.e. y1, ..., yt−1, unnormalized relevance scores e (t) i j will be measured by spatial attention functions:

$$e_{ij}^{(t)} = w_l^T tanh(W_e h_{t-1} + U_e vl_{ij} + z_e). \qquad (2)$$

where w T l , We, Ue, ze are the shared parameters to be learned by our model at all the time steps. The unnormalized relevance scores, which can reflect the relevance of the j-th local features in the input video. will be normalized over n local features to calculate attention weights α (t) i j :

$$\alpha_{ij}^{(t)} = exp\{e_{ij}^{(t)}\} / \sum_{j'=1}^{n} exp\{e_{ij'}^{(t)}\}. \qquad (3)$$

The spatial attention mechanism makes the decoder to selectively catch the most significant regions according to increasing the attention weights.

E. Exploiting Temporal Structure: A Temporal Attention Mechanism

With regards to a generated sentence, each word should be in line with different temporal segment of the video. Hence, while generating each sentence, the model should attend to different temporal segments at each time. For this reason, we respectively exploit temporal attention mechanism to dynamically weighted sum of the global, local and motion features such that

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

$$\varphi_t(VG) = \sum_{i=1}^{k} \beta_i^{(t)} vg_i; \quad (4)$$

$$\varphi_t(VM) = \sum_{i=1}^{k} \delta_i^{(t)} vm_i; \quad (5)$$

$$\varphi_t[\Psi(VL)] = \sum_{i=1}^{k} \gamma_i^{(t)} \Psi_i(VL), \quad (6)$$

where Ík i=1 β (t) i , Ík i=1 δ (t) i and Ík i=1 γ (t) i are all equal to 1. β (t) i , δ (t) i and γ (t) i are calculated respectively at each time step t inside the LSTM decoder. We denote by them as the temporal attention weights at time t.

The relevance scores, namely $b_i^{(t)}$, $c_i^{(t)}$ and $d_i^{(t)}$, will be measured by three temporal attention functions:

$$b_i^{(t)} = w_k^T tanh(W_b h_{t-1} + U_b vg_i + z_b); \quad (7)$$

$$c_i^{(t)} = w_e^T tanh(W_c h_{t-1} + U_c vm_i + z_c); \quad (8)$$

$$d_i^{(t)} = w_r^T tanh(W_d h_{t-1} + U_d \Psi_i(VL) + z_d), \quad (9)$$

where w_*^T, U_* and z_* are shared by all the features respectively.

After that, the relevance scores will be normalized via *softmax* function:

$$\beta_i^{(t)} = exp\{b_i^{(t)}\} / \sum_{i'=1}^{k} exp\{b_{i'}^{(t)}\}; \quad (10)$$

$$\delta_i^{(t)} = exp\{c_i^{(t)}\} / \sum_{i'=1}^{k} exp\{c_{i'}^{(t)}\}; \quad (11)$$

$$\gamma_i^{(t)} = exp\{d_i^{(t)}\} / \sum_{i'=1}^{k} exp\{d_{i'}^{(t)}\}. \quad (12)$$

When obtaining three kinds of temporal features, they are fused by sum fusion such that

$$\varphi_t(V) = \varphi_t(VG) + \varphi_t(VM) + \varphi_t[\Psi(VL)]. \quad (13)$$

Finally, each ϕt(V) will input to LSTM unit at each time step. The temporal attention mechanism makes the decoder to selectively focus on a subset of frames according to increasing the attention weights of the corresponding three different kinds of features. In this way, the decoder can utilize the temporal structure sufficiently. Consequently, these two attention mechanisms are systematically integrated into the encoder-decoder neural framework, which can focus on how to predict significant regions more precisely while focusing on more semantically relevant video frames.

F. Decoder: Long Short-Term Memory Network

The decoder network can convert visual features into a word sequence Y = {yi , y2, ..., ym}, which summarizes the video contents. Hence, each ϕt(V) will be fed into LSTM unit at each time step, which is formulated as follows:

$$i_t = \sigma(W_i E[y_{t-1}] + U_i h_{t-1} + A_i \varphi_t(V) + b_i); \quad (14)$$

$$f_t = \sigma(W_f E[y_{t-1}] + U_f h_{t-1} + A_f \varphi_t(V) + b_f); \quad (15)$$

$$o_t = \sigma(W_o E[y_{t-1}] + U_o h_{t-1} + A_o \varphi_t(V) + b_o); \quad (16)$$

$$g_t = \sigma(W_g E[y_{t-1}] + U_g h_{t-1} + A_g \varphi_t(V) + b_g); \quad (17)$$

$$c_t = c_{t-1} \odot f_t + i_t \odot g_t; \quad (18)$$

$$h_t = o_t \odot \phi(c_t), \quad (19)$$

where ϕt(V) is the encoder temporal representation. E is a word embedding matrix, and we refer to E[yt−1] as an embedding vector of word yt−1. σ is a sigmoid activation function; φ is a tanh function, yt−1 is the previous word; ht−1 is the previous hidden state; Besides, W∗, U∗, A∗ and b∗ are the shared weight matrices and bias to be learned. Finally, the probability distribution of a series of target words at each time will be obtained via a single hidden layer:

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

$$\hat{y}_t = softmax(U_y \phi(W_y[h_t, \varphi_t(V), E[y_{t-1}]]) + b_y), \quad (20)$$

where [ht , ϕt(V), E[yt−1]] denotes the concatenation of the three vectors. Wy, Uy, and by are the shared weight matrices and bias to be learned. The softmax function allows us to interpret yˆt as the probabilities of the distribution p(yt |y<t, V, θ) over words. Decoder can approximately find the sentence with the highest probability by using the beam search.

IV. EXPERIMENT

A. Dataset and Evaluation Metrics

Dataset: The extensive experiments are conducted on two famous video captioning benchmarks: MSVD [39] and MSRVTT-10K [40]. The MSVD has 1970 video clips including a series of human annotated language sentences. There are 80,839 sentences in total, with about 41 annotated sentences per clip. The words in all the sentences form a vocabulary that contains 13,010 unique words. Following [12], the dataset can be divided into a training set of 1,200 video clips, a validation set of 100 clips, and a test set consisting of the rest of 670 clips. The MSR-VTT-10K [40] consists of 10,000 video clips that is the most challenging dataset for video captioning so far. We use the official spilt with 6513 videos for training, 497 for validation and 2990 for testing.

Evaluation Metrics: There are kinds of approaches with regards to the evaluation of generated sentences have been adopted, such as BLEU [41], METEOR [42] and CIDEr [43]. BLEU has been widely used to evaluate the performance of machine translation. It is mainly based on the n-gram accuracy. The METEOR is proposed to correct a kind of problems in BLEU. It can generate an alignment according to exact token matching to judge the word correlation between candidate and reference sentences. CIDEr exploits human consensus to evaluate video descriptions. We get all the results in this paper according to the Microsoft COCO evaluation server [44].

B. Feature Extraction

Global Feature Extraction: For frame-level global features, on MSVD, we adopt 1024-dimension pool5/7 × 7_s1 layer from GoogLeNet [45] and denote them as VG = {vg1, ..., vgk }. On MSR-VTT-10K, we have tried to use two CNN features for frame- level global features, such as GoogleNet pool5/7×7_s1 and ResNet-152 [46] pool5. Finally we find that ResNet-152 pool5 gets best performance on this task. Therefore, we use pool5 feature from ResNet-152 as globall feature on MSR-VTT-10K.

Motion Feature Extraction: For motion features, we use the 4096-dimensional fc6 layer from C3D [47] pre-trained on the Sports-1M video dataset [48]. We take continuous 16 frames as the input short clips for the C3D, similar as the default setting. The C3D features are denoted as V M = {vm1, ..., vmk }.

Object Detection: During the training and testing phase, detecting kinds of local

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

features vli is an indispensable part of our STAT. In the recent years, the most advanced approaches in general class of detection objects are mainly grounded on CNN Girshick et al. [49] proposed a multilevel network called Regions with Convolutional Neural Networks (R-CNN). It aims to train depth CNN for classifying proposals of target detection. In order to accelerate the training speed, Fast RCNN and Faster R-CNN are proposed constantly. In this paper, we exploit Faster R-CNN [50] as our object detector due to its accuracy in much object detection work.

Pre-trained Faster R-CNN model is used to directly detect kinds of objects in video frames. Faster R-CNN model will obtain a series of rectangular object proposals, each with a class confidence score, via taking an image (of any size) as input. The higher the score, the more likely there is an object for a certain class. For the purpose of alleviating redundant computation complexity, we attempt to make a set of strategies to tackle it. Firstly, the number of proposals are reduced from 300 (e.g. default setting) to 100. The reason is that when we use the top-ranked 100 proposals at a testing phase, it also achieves a better performance [50]. In fact, Non-Maximum Suppress algorithm (NMS) will further reduce the average number of proposals. Furthermore, we only detect objects in 28 equally-spaced frames in each video in that there is little change in adjacent frames. Finally, we trained Faster R-CNN model detecting 80 objects on MS COCO dataset.

(1) Visual Feature Extraction: For local visual features, we represent top-n objects as 4096-dimensional features respectively, which are extracted from the fc7 layer in the Faster R-CNN network. After that, we obtain a series of local visual features vli = {vli1, ..., vlin} where vli j ∈ R 4096 in each frame;

(2) Label Feature Extraction: For local label features, we firstly obtain names of top-n objects. To make these important objects relate to semantic space, we connect them to semantic relationship using Glove [51], which maps each object name to 300 dimension semantic space. The embedded label features then are used to form our visual-semantic attention context. Finally, we obtain a set of local label features vli = {vli1, ..., vlin} where vli j ∈ R 300 on each frame.

C. Model and Training

Our video captioning framework is shown in figure 2. On MSVD, the decoder network has one LSTM layer with 1024 cells. Each word is embedded to a 512-dimensional vector when it is fed to the LSTM layer. Especially, on MSR-VTT10K, in order to alleviate computation and keep consistent with the [52], [53], [54], [55], the hidden layer size and word embedding size are all set to 512. On two both datasets, learning rate is set to 2 × 10−4 empirically.

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{t_m} logp(y_i^m | y_{<i}^m, x^m, \theta), \qquad (21)$$

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

where there are N training video-description pairs (x m,y m). θ are parameters of the video captioning model and each description y m is tm words long.

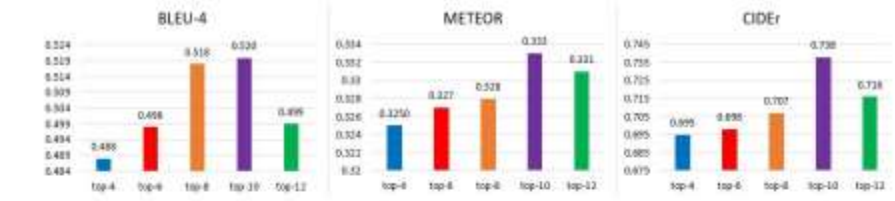

Fig. 3. Illustration of the performance with different top-n objects which are represented by local visual features on MSVD.

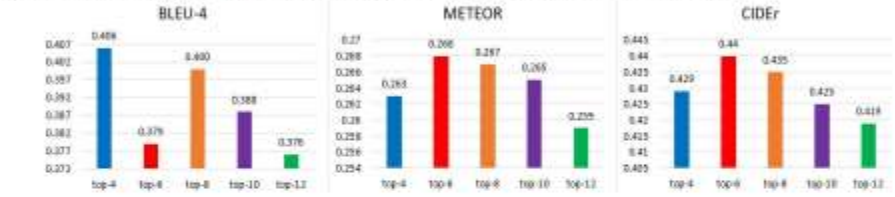

Fig. 4. Illustration of the performance with different top-n objects which are represented by local visual features on MSR-VTT-10K.

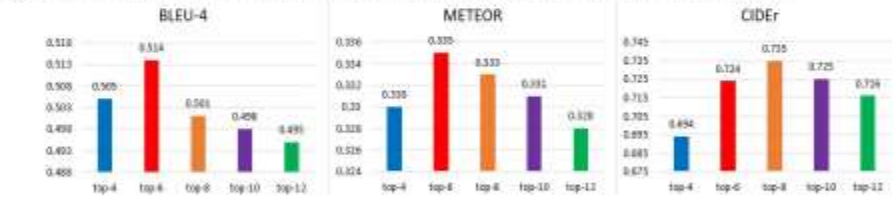

Fig. 5. Illustration of the performance with different top-n objects which are represented by local label features on MSVD.

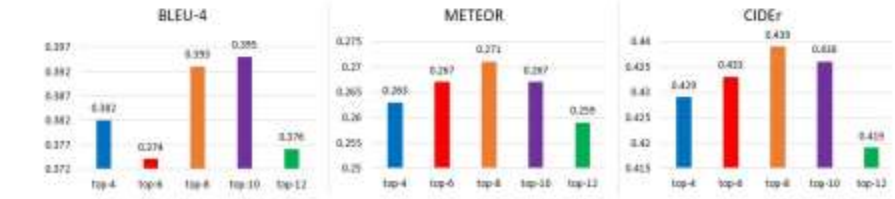

Fig. 6. Illustration of the performance with different top-n objects which are represented by local label features on MSR-VTT-10K.

D. Video Captioning Experiments and Results

in this section, we will conduct five comparative experiments as follows:

• Exploring the proper number of top-n objects when exploiting local visual and label features, respectively;

• Comparison of different local features ;

• Comparison of different kinds of features ;

• Comparing four method to process local features, i.e. 1) no exploiting; 2) mean pooling; 3) max pooling; 4) spatial attention mechanism;

• Comparing STAT with state-of-the-art method

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

TABLE I
COMPARISON EXPERIMENTS FOR DIFFERENT LOCAL FEATURES ON MSVD

		Feature Types				Evaluation Metric		
Method	Attention	Global	Motion	Local	Local Number	BLEU4	METEOR	CIDEr
STAT	Spatial & Temporal	GoogleNet	C3D	RCNN-fc7	10	**0.520**	0.333	**0.738**
STAT	Spatial & Temporal	GoogleNet	C3D	Glove	6	0.514 (-1.2% ↑)	**0.335 (0.6% ↑)**	0.724 (-0.2% ↑)

TABLE II
COMPARISON EXPERIMENTS FOR DIFFERENT LOCAL FEATURES ON MSR-VTT-10K

		Feature Types				Evaluation Metric		
Method	Attention	Global	Motion	Local	Local Number	BLEU4	METEOR	CIDEr
STAT	Spatial & Temporal	ResNet-152	C3D	RCNN-fc7	6	0.379	0.268	**0.440**
STAT	Spatial & Temporal	ResNet-152	C3D	Glove	8	**0.393 (3.7% ↑)**	**0.271 (1.1% ↑)**	0.439 (-0.2% ↑)

Evaluation on Different Number of Local Features: How many local features to select is critical to generate the correct description. For the purpose of obtaining an appropriate number of local features, in this section, we respectively use local visual and label features to conduct experiments. Then, we analyze the influence of different number of them for video captioning. Here, we chose 4, 6, 8, 10 and 12 objects in each frame to conduct experiments.

1) The Determination of The Number of Local Visual Features: Fig.3 and Fig.4 respectively show the results on each benchmark with different top-n objects. These objects are represented by local visual features. Each subfigure shows the accuracy of different n with the same metric and each color shows the accuracy of the same n with different metrics. From these experiments, we find that it realizes best performance when choosing top-10 objects in each frame on MSVD. For MSR-VTT-10K, we find that it achieves better performance when choosing top-6 objects. Hence, we finally choose that n is equal to 10 on MSVD and 6 on MSR-VTT-10K to carry out our following experiments.

2) The Determination of The Number of Local Label Features: Fig.5 and Fig.6 respectively show the results on each benchmark with different top-n objects. These objects are represented by local label features. Each subfigure shows the accuracy of different n with the same metric and each color shows the accuracy of the same n with different metrics.

According to these experiments, we obtain a finding that the number of either local visual or local label features should be moderation rather than the bigger the better. If the number of local features is larger, there may have some non-object information to impact the correction of prediction. In contrast, if the number of local features is smaller, there may miss some critical object information, which could lead to generate incomplete description.

TABLE I shows the comparison results of the exploitation of different local features on MSVD. Compared with exploiting local label features, the performance of exploiting local visual features is better on BLEU-4 and CIDEr. By contrast, as shown in TABLE II, compared with exploiting local

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

label features, the performance of exploiting local visual features on MSRVTT-10K becomes badly on BLEU-4 and METEOR, and their scores in CIDEr has little difference.

Meanwhile, compared to using local visual features, we note that when using local label features, the performance over all metrics on MSVD averagely decreases 0.27%, but averagely increases 1.53% over all metrics on MSR-VTT-10K. Especially, local label features with dimension of 300 are much smaller than local visual features with dimension of 4096, so they take up much less space and are loaded faster. Hence, from the perspectives of performance and practical operation, we finally argue that using local label features is a better selection.

1) GloFeat: Only employing the global features which are extracted from GoogleNet.

2) MotFeat: Only employing the motion features which are extracted from C3D.

3) Loc_V: Only employing the local visual features which are extracted from fc7 layer in Faster-RCNN.

4) Loc_L: Only employing the local label features which are extracted from Glove model.

5) STAT_V: Employing the global, motion and local visual features.

6) STAT_L: Employing the global, motion and local label features.

TABLE III
COMPARISON EXPERIMENTS FOR EXPLOITING DIFFERENT KINDS OF FEATURES ON MSVD

Methods	BLEU-4	METEOR	CIDEr
GlobFeat	0.489	0.326	0.671
MotFeat	0.474	0.310	0.665
Loc_V	0.432	0.305	0.628
Loc_L	0.429	0.310	0.625
STAT_V *	0.520	0.333	0.738
STAT_L *	0.514	0.335	0.724

* V represents local visual features. * L represents local label features.

TABLE IV
COMPARISON EXPERIMENTS FOR EXPLOITING DIFFERENT KINDS OF FEATURES ON MSR-VTT-10K

Method	BLEU-4	METEOR	CIDEr
GlobFeat	0.371	0.259	0.410
MotFeat	0.365	0.254	0.399
Loc_V	0.351	0.246	0.369
Loc_L	0.352	0.251	0.358
STAT_V *	0.379	0.268	0.440
STAT_L *	0.393	0.271	0.439

* V represents local visual features.* L represents local label features.

TABLE V
COMPARISON EXPERIMENTS FOR EXPLOITATION OF LOCAL FEATURES ON MSVD

		Feature Types				Evaluation Metric		
Method	Attention	Global	Motion	Local	Local Number	BLEU-4	METEOR	CIDEr
TAT-NL	Temporal	GoogleNet	C3D	-	0	0.464	0.318	0.625
TAT-L_{mean}	Temporal	GoogleNet	C3D	R-CNN fc7	10	0.435	0.307	0.597
TAT-L_{max}	Temporal	GoogleNet	C3D	R-CNN fc7	10	0.448	0.326	0.669
STAT	Spatial & Temporal	GoogleNet	C3D	R-CNN fc7	10	**0.520**	**0.333**	**0.738**

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

TABLE VI
COMPARISON EXPERIMENTS FOR EXPLOITATION OF LOCAL FEATURES ON MSR-VTT-10K

		Feature Types				Evaluation Metric		
Method	Attention	Global	Motion	Local	Local Number	BLEU-4	METEOR	CIDEr
TAT-NL	Temporal	ResNet-152	C3D	-	0	0.371	0.264	0.398
TAT-L_{mean}	Temporal	ResNet-152	C3D	Glove	8	0.343	0.243	0.319
TAT-L_{max}	Temporal	ResNet-152	C3D	Glove	8	0.374	0.256	0.406
STAT	Spatial & Temporal	ResNet-152	C3D	Glove	8	**0.393**	**0.271**	**0.438**

The comparison results are shown in the TABLE III and TABLE IV. On both benchmarks, firstly, according to results of the three kinds of single feature experiments, we found that the effectiveness of only using global features or motion features is better than only using local features. The reason may be that global features represent overall semantic concepts of corresponding video frames, and motion features characterize punctuated actions such as "jumping up" or "standing up". These two kinds of features can both roughly summarize the content of a video. Especially, global features can provide the relationship of several objects that is useful for generating video descriptions. However, only using local features cannot summarize the content of a video. By comparing STAT experiment with the others, we observer that fusion form can get the best performance. Our conjecture is that STAT can not only get overall global context information, but capture local object information and their relationship.

Evaluation on Eexploitation of Local Features: Local features are very critical for obtaining specific information in object detection. Compared to a single image, a video is composed of multiple frames, so it naturally has more objects. Hence, we conduct four experiments to explore the effectiveness of local features and how to properly process them. The specific configurations of the four experiments are shown as follows:

1) TAT-NL: Only employing the global features and motion features with temporal attention.

2) TAT-Lmean: First, employing mean pooling method to process top-n local features and get a single local representation on each frame. Then, employing temporal attention mechanism to respectively select most relevant global, local and motion features and fusing them by sum fusion.

3) TAT-Lmax: First, employing max pooling method to process top-n local features and get a single local representation on each frame. Then, employing temporal attention mechanism to respectively select most relevant global, local and motion features and fusing them by sum fusion.

4) STAT: First, employing spatial attention mechanism to dynamically assign attention weights for each local features to select significant regions with maximal attention weights. Then, employing temporal attention mechanism to selectively attend to important temporal segments with maximal attention

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

weights. Finally, these selected features are fused by sum fusion.

TABLE VII
COMPARISON WITH SEVERAL STATE-OF-THE-ART MODELS ON MSVD

Methods	BLEU-4	METEOR	CIDEr
STAT_V *	**0.520**	0.333	**0.738**
TA[12]	0.419	0.296	0.517
LSTM-E[37]	0.453	0.310	-
h-RNN[13]	0.499	0.326	0.658
HRNE [11]	0.467	**0.339**	-
M-Fusion[10]	0.499	0.318	0.634

* V represents local visual features.

TABLE VIII
COMPARISON WITH TOP-5 RESULTS FROM THE 1ST MSR-VTT-10K CHALLENGE

Method	BLEU-4	METEOR	CIDEr
STAT_L *	0.393	0.271	0.439
v2t_navigator[53]	**0.408**	**0.282**	0.448
Aalto[52]	0.398	0.269	0.457
VideoLAB[54]	0.391	0.277	0.441
ruc-uva [55]	0.387	0.269	**0.459**
Fudan-IILC	0.387	0.268	0.419

* L represents local label features.

TABLE V and TABLE VI show the results of these experiments. On both benchmarks, firstly, according to results of the TAT-NL experiment with TAT-Lmean experiment, we observer that former gets a better performance. Our conjecture is that mean pooling fuses all the objects into a single vector representation, leading to the loss of spatial structure underlying the input frame. Hence, these local features might become noises to disturb other features.

Secondly, from the comparison result of the TAT-NL experiment with TAT-Lmax experiment, we observe that using local features by max pooling will achieve better performance. We argue that compared to mean pooling, max pooling not only selects the most prominent object, but also don't impact on the spatial structure underlying the input frame. Hence, when introducing these local features, if we don't confuse them and keep spatial structure of input frame, these local features indeed improve the performance.

Finally, by comparing STAT with the TAT-Lmax, we find STAT has better performance on all evaluation metrics. Our conjecture is that the only one most prominent object is insufficient for video description task. As shown in the 5th of Fig.7 , if only 'person' has the most high class confidence score, other objects, such as 'bottle', 'bowl' et al., will be overlooked in max pooling.

According to above all analyses, we argue that it is indispensable to incorporate spatial attention mechanism into encoder-decoder network. We draw a conclusion that temporal and spatial cues are complementary. When both attention mechanisms are utilized simultaneously, it will achieve the best performance

Comparison with State-of-the-art Methods: Finally, we compare our STAT method with several state-of-the-art methods. On MSVD, here, we choose local visual features and compare STAT with five methods: HRNE [11], TA [12], hRNN [13], LSTM-E [37], and M-Fusion [10]. HRNE exploited a hierarchical recurrent encoder to model video temporal information. TA is the first work to utilize temporal attention mechanism for video captioning. h-RNN proposed an approach which use hierarchical RNN to handle the problems in video captioning. LSTM-E intended to explore visual semantic embedding and learning of

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

LSTM. M-Fusion explores a proper method to fuse different types of features with attention model. On MSR-VTT-10K, we choose local label features and compare STAT with the top-5 results from the 1st MSR-VTT challenge on the LeaderBoard1 , including v2t_navigator [53], Aalto [52], VideoLAB[54], ruc-uva [55] and Fudan-ILC 2 , which are all based on multiple types of features. For instance, v2t_navigator and VideoLAB exploit almost all the features extracted from this dataset, such as global, motion, aural and category features. ruc-uva and Aalto exploit global, motion and category features.

1) Results on MSVD: We show the results on MSVD in TABEL VII. Our STAT realize the best performance on BLEU4 and CIDEr compared with other work. The main feature of BLEU is for corpus-level comparisons where many n-gram matches exist [44]. CIDEr, a consensus-based metric, is used to reward a description which conforms to the writing habits of human beings. Our STAT therefore not only generate accurate sentences but maintain human language habits. We note HRNE has better performance than us on METEOR. We compute the relevant improvements based on them. Compared to them, in terms of METEOR, our STAT decreases 1.8%. In terms of BLEU-4, however, our STAT increases 11.3%. Hence, our STAT can generate more relevant description sentences than them.

2) Results on MSR-VTT-10K:We report the results on MSR-VTT-10K in TABLE VIII. By comparing our STAT with the top-5 results from 1st MSR-VTT-10K challenge, which represent the state-of-the-art results on this benchmarks, we note that our STAT method obtains a medium level on all evaluation metric. Compared the existing methods on this benchmark, STAT hardly have any improvements. There may be several likely reasons accounting for this limited improvements. Firstly, the human annotated sentences of the same video are more diverse on this challenging dataset. Secondly, considering the fact that 836 out of 23,667 words in MSRVTT-10K annotated sentences are misspelled (e.g., ‘basktball’ and ‘peson’) [30], If the spelling is not corrected, it is much more difficult to describing the content of a video on this dataset. In addition, the dataset is more challenging due to the diversity of its videos. This might result in misprediction when detecting objects on this dataset, so weakening the effectiveness of local features and the advantage of spatial attention mechanism. Therefore, the detection of tiny object is worth exploring. We will attempt to train an object detector based on videos to improve the accuracy of our STAT in the future. Finally, v2t_navigator [53], Aalto [52], VideoLAB[54], and ruc-uva [55] all use category features and demonstrate that its useful to improve the performance. MSR-VTT-10K has been classed as total 20 different categories (e.g., music, tv shows, sports, etc.) by the official. Each video is tagged with a category. Category features provide a strong prior information about video content. It is also useful for dynamically weighting other kinds of

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

features [53]. Besides, most of them use the aural features. Aural features are complementary to visual features, which are especially useful to distinguish different scene events. However, our goal in this work is to improve the visual encoder, which is quite different from their research.

Fig. 7. Eight representative samples from MSVD and MSR-VTT-10K, which also involve the sentences from ground truth, temporal attention mechanism (TAT) and spatial-temporal attention mechanism (STAT). In the figure, SA and TA indicate the results of spatial and temporal attention. The white aperture in each frame indicates the change in the degree of significance of corresponding regions in the spatial attention stage. The bar plot under each frame are the temporal attention weights for this frame when the corresponding word (color-coded) was generated in the stage of temporal attention.

E. Qualitative Analysis

Although we can evaluate our model through the evaluation mechanism described in [44], the scores do not directly reflect the performance of the generated sentences by STAT.

Therefore, we visualized the spatial and temporal attention results of some representative videos from MSVD and MSRVTT-10K, as shown in Figure 7. In the figure, SA and TA respectively represent the spatial and the temporal attention result. Figure 7 includes the sentences obtained from TAT, our STAT and ground truth.

In Figure 7, we can clearly see that compared to TAT using temporal attention, since we integrate the spatial attention with

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

temporal attention during the description generation, the sentences generated by STAT can contain more details (eg‘paper airplane’, ‘tv show’, ‘bowl’) and less error. Besides, STAT could not only select key frames which relate to each word, but could catch significant regions in these frames. For instance, in the No.1 video, while generating the word ‘boy’, first frame is selected via temporal attention mechanism. Then, the discriminative face area of the boy is caught via the spatial attention mechanism. And then, when the word of ‘dog’ is generated, STAT turns to select third and fourth frames according to previous generated words. However, in the sentence generated by TAT, the ‘dog’ is misrecognized as the ‘baby’. In another example, TAT is hard to determine the place where ‘a man and woman talking’ in No.4 video. By comparison, STAT recognizes the ‘couch’ accurately, and successfully identifies the ‘dog’ and other details, because we augment the local features, and STAT can selectively catch significant regions. Above representative examples further demonstrate that it is indispensable to integrate spatial attention mechanism with temporal attention mechanism in the task of video captioning.

V. CONCLUSION

Existing methods based on temporal attention mechanism cannot catch significant regions in frames, so they might suffer from the problems of recognition error and detail missing. In this paper, we propose a spatial-temporal attentional mechanism. Different from previous methods, we take into account both the spatial and temporal structures in a video, so it learns the ability to not only focus on a subset of frames, but further catch significant regions in that subset. Extensive experiments conducted on two well-known benchmarks show that our STAT achieves state-of-the-art performance on both evaluation benchmarks. We prove this by quantitatively analyzing the spatial and temporal attention weights in the process of sentence generation. Consequently, our method can generate detailed and accurate descriptions. In the future, we will attempt how to model the relationship among local features to improve the effectiveness of them.

Since both spatial and temporal information in a video are critical to understand contents of it, any encoder-decoder based on video understanding task can added our spatial-temporal attention method to promote the overall performance.

VI. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the constructive comments to improve the paper. The work is partially supported by National Nature Science Foundation of China (61671196, 61525206, 61701149), Shenzhen Fundamental Research fund (JCYJ20180306174120445, JCYJ20160331185006518), Zhejiang Province Nature Science Foundation of China LR17F030006, National Key Research and Development Program of

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

China (2017YFC0820600, 2017YFC0820605, 2017YFC0820604), 111 Project, No. D17019, Startup Research fund of Shenzhen University (2019041).

REFERENCES

[1] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," IEEE Trans. Multimedia, vol. 19, no. 9, pp. 2045–2055, 2017.

[2] X. Liu and W. Wang, "Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis," IEEE Trans. Multimedia, vol. 14, no. 2, pp. 482–489, 2012.

[3] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," IEEE Trans. Multimedia, vol. 10, no. 3, pp. 421–436, 2008.

[4] Y. Liao and J. D. Gibson, "Routing-aware multiple description video coding over mobile ad-hoc networks," IEEE Trans. Multimedia, vol. 13, no. 1, pp. 132–142, 2011.

[5] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: globallocal attention for image description," IEEE Trans. Multimedia, vol. 20, no. 3, pp. 726–737, 2018. [Online]. Available: https://doi.org/10.1109/TMM.2017.2751140

[6] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2045–2055, 2017.

[7] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Selfsupervised video hashing with hierarchical binary auto-encoder," IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3210–3221, 2018.

[8] L. Pang, S. Zhu, and C. Ngo, "Deep multimodal learning for affective analysis and retrieval," IEEE Trans. Multimedia, vol. 17, no. 11, pp. 2008–2020, 2015. [Online]. Available: https://doi.org/10.1109/TMM.2015.2482228

[9] N. Zhao, H. Zhang, R. Hong, M. Wang, and T.-S. Chua, "Videowhisper: Toward discriminative unsupervised video feature learning with attention-based recurrent neural networks," IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2080–2092, 2017.

[10] C. Hori, T. Hori, T.-Y. Lee, K. Sumi, J. R. Hershey, and T. K. Marks, "Attention-based multimodal fusion for video description," arXiv preprint arXiv:1701.03126, 2017.

[11] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1029–1038.

[12] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville,

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2022.28.04.189

"Describing videos by exploiting temporal structure," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4507–4515.