Video Description with Spatial-Temporal Attention

Yunbin Tu
Institute of Information and Control, Hangzhou Dianzi University
Hangzhou, China
tuyunbin1995@foxmail.com

Xishan Zhang
Key Lab of Intelligent information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS
Beijing 100190, China
zxs@ict.ac.cn

Bingtao Liu
Institute of Information and Control, Hangzhou Dianzi University
Hangzhou, China
liubingtao@hdu.edu.cn

Chenggang Yan
Institute of Information and Control, Hangzhou Dianzi University
Hangzhou, China
cgyan@hdu.edu.cn

ABSTRACT

Temporal attention has been widely used in video description to adaptively focus on important frames. However, most existing methods based on temporal attention suffer from the problems of recognition error and detail missing, because only coarse frame-level global features are employed. Inspired by recent successful work in image description using spatial attention, we propose a spatial-temporal attention (STAT) method to address such problems. In particular, first, we take advantage of object-level local features to address the problem of detail missing. Second, the STAT method further selects relevant local features by spatial attention and then attend to important frames by temporal attention to recognize related semantics. The proposed two-stage attention mechanism can recognize the salient objects more precisely with high recall and automatically focus on the most relevant spatial-temporal segments given the sentence context. Extensive experiments on two well-known benchmarks suggest that STAT method outperforms the state-of-the-art methods on MSVD with BLEU4 score 0.511, and achieves superior BLEU4 score 0.374 on MSR-VTT-10K. Compared to the method without local features, the relative improvements derived from our STAT method are 10.1% and 0.8% respectively on two benchmarks. Compared to the method using only temporal attention, the relative improvements derived from our STAT method are 18.3% and 9.0% respectively on two benchmarks.

KEYWORDS

Video Descripton; Temporal Attention; Spatial Attention.

1 INTRODUCTION

Recently, automatic video description has received increasing attention in the fields of multimedia, computer vision and natural language processing, because it enables a variety of practical applications. For example, it helps users of video sites to retrieve video efficiently, and benefits visually impaired people to better understand the video contents.

Compared to image captioning, describing videos is more challenging because the videos are composed of consecutive frames, involving both static objects and dynamic human actions. A typically video clip lasts 5 to 10 seconds, containing 120 to 240 frames. Though videos contain such vast quantity of information, people do not describe everything in videos, and it is hard to determine the most relevant objects and describe the event appropriately. Therefore, a description generation model should be clever enough to attend to the most relevant part of the videos.

Regarding the difficulty of these problems, visual attention mechanism [8, 14, 37, 39] has been proposed recently to selectively focus on part of the information in the video. To the best of our knowledge, most of existing temporal attention-based methods

Ground truth: A man is cutting a tree.
TAT: A man is cutting a head.
STAT: A man is cutting a tree.

Ground truth: A man is calling.
TAT: A man is talking.
STAT: A man is talking on the phone.

Figure 1: Illustration of problems of the detail missing (bottom: missing 'phone') and misrecognition (top: misrecognizing 'tree' as 'head'), where TAT and STAT are short for temporal attention and spatial-temporal attention.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123354

only utilize coarse frame-level global features. As a video contains complex interactions of humans and objects, there are always multiple salient objects in single frame. Though the important frames are selectively focused using temporal attention, it is still hard to attend to multiple meaningful objects on each frame, which will lead to detail missing in video description. Taking the temporal attention result in the bottom of Figure 1 as an example, the local detail 'phone' is missing. In addition, frame-level global features are extracted at a coarse level, which are incapable of representing and localizing small objects. Therefore, using global features alone will result in recognition error of small objects during the process of description generation. Taking the temporal attention result in the top of Figure 1 as an example, the local detail 'tree' is wrongly recognized as 'head'.

There have been a few attempts to capture multiple objects information in video description. Shetty *et al.* [21] introduced object-level local features extracted by pre-trained SVM classifier and integrated these features into global features. These local features from each frame, are then collapsed via simple average or maximum pooling to result in a single vector representation of each frame. However, the indiscriminative average or maximum pooling of all the objects ignores the important differences among local features. It is also worth pointing out that Yu *et al.* [39] considered using spatial-temporal attention for video captioning. There are several important differences between our work and [39]. Firstly, in [39] local features have fixed resolution (220×220) and are extracted at pre-defined spatial locations. We argue that their detection method is likely to cause false judgment on the objects. Secondly, they do not differentiate the order between spatial and temporal attention. When human describe a video, they always first focus on specific objects on the frame, and then study the interaction between objects over time. Therefore, we argue that it is reasonable to first calculate spatial attention weights for local features and then compute the temporal attention weights for frames. Thirdly, [39] only uses local features instead of global features, which will overlook the context information.

To address the above issues, in this paper, first, we take advantage of local features extracted by Faster R-CNN [19] to address the problem of detail missing. Faster R-CNN can generate variable-size bounding boxes according to the actual size of objects, and detects multiple objects more accurately. Second, we introduce a spatial-temporal attention (STAT) method to selectively attend to not only specific subset of frames, but also salient objects in that subset.

In summary, we make the following contributions:

- We study of the importance of using local feature in video description, and improve the recognition and localization of multiple small objects on video frames. In addtion, we discover that the introducing of local feature will make the temporal attention insufficient, because temporal attention alone is hard to distinguish multiple salient objects on one frame, thus generating worse descriptions.
- We propose a spatial-temporal attention (STAT) method for video description. By assigning different weights to the spatial features on each frame and the temporal features on consecutive frames, the STAT method is able to capture the key details while keeping the global and motion information in the video, thus it can address the problems of recognition error and detail missing.
- Extensive experiments conducted on two well-known video description benchmarks, MSVD and MSR-VTT-10K demonstrate that our STAT method achieves noticeable gains by appropriately integrating spatial attention into temporal attention.

2 RELATED WORK

Video/Image captioning: In the image captioning work, [3, 9, 10, 13, 29] first try to use RNN for visual text translation. In the task of image captioning, the input is a single image without temporal structure, and the output is a natural language description. Thus, the overall structure of an image captioner (instance-to-sequence) is also usually simpler than that of a video captioner (sequence-to-sequence) [11]. Inspired by the successful application of RNN in image captioning, Venugopalan *et al.* [28] has applied neural approach to video description. However, a main shortcoming of this method is that this representation completely ignores the ordering of the video frames and fails to utilize any temporal structure [27]. To solve this problem, Yao *et al.* [37] proposed to exploit global temporal structure which lets the decoder selectively focus on only a small subset of frames at a time. Our STAT method is closely related to [37], because we also mainly use attention mechanism to selectively focus on video features. However, an obvious difference is that our attention model to selectively attend to not only specific subset of frames, but also specific objects in that subset.

Attention mechanism in image/video captioning: Attention mechanism has been widely used in captioning tasks [12, 25, 31, 36–38, 43]. On one hand, the tasks of image captioning mainly exploit spatial attention mechanism. Taking Xu *et al.* [31] for instance, they explored two attention-based image caption approaches, which are able to generate a target word according to the most relevant regions in an image. On the other hand, the tasks of video captioning mainly utilize temporal attention mechanism. For example, Yao *et al.* [37] first introduced a temporal attention mechanism to exploit global temporal structure, which is able to generate a target word based on the most relevant frames in a video. As we know, the tasks of video captioning not only have a temporal structure on the consecutive frames, but also a spatial structure on each frame. Therefore, Yu *et al.* [39] has explored both spatial attention and temporal attention to capture quite small and difficult to be localized objects. However, when they caculated attention weights of all the patch features, they ignored the order between spatial and temporal attention. In contrast, we believe that the order is important due to the visual attention mechanism of human beings. Therefore, we argue that a clever decoder first should focus on salient objects on each frame by spatial attention, and then selects relevant frames on consecutive frames by temporal attention. Compared to the above methods, our STAT method is built upon attention and extends it one step further, which is able to generate a target word based on both spatial attention and temporal attention while considering the order between two attentions.

The use of object-level local features in video captioning: As we know, the videos contain more types of features than images,

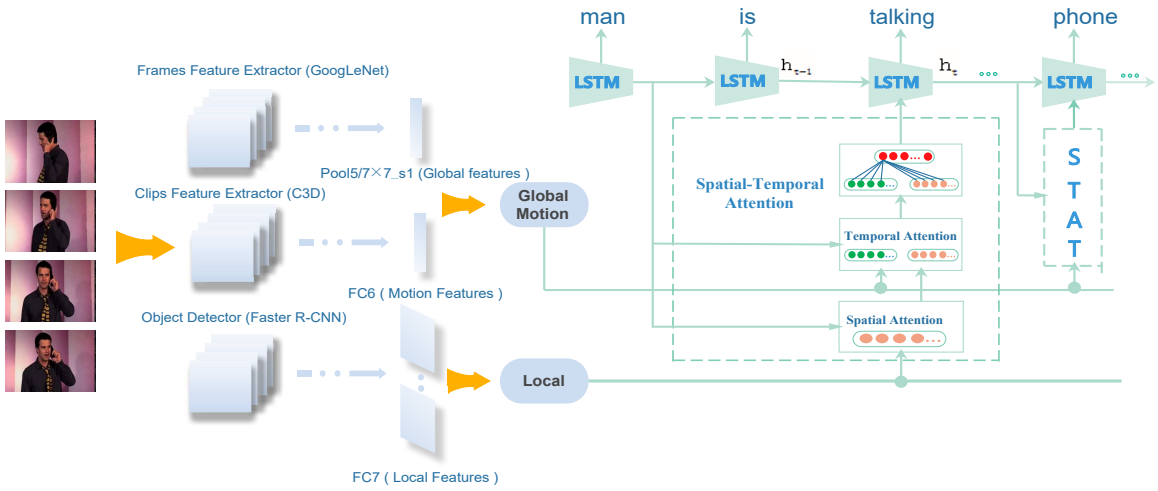

Figure 2: The video description based on spatial-temporal attention (STAT) is displayed. STAT is mainly composed of two parts: spatial attention (SA) and temporal attention (TA), the detailed STAT unit will be shown in Figure 4.

such as appearance features, motion features. However, most existing work in video captioning mainly use frame-level appearance features. Shetty *et al.* [21] utilized pre-trained SVM classifier and integrated these features into global features. However, they only adopted simple averaging strategy to deal with these local features. This approach risks ignoring the spatial structure underlying each frame. For instance, it is not possible to tell the importance between two objects from the collapsed features. Yu *et al.* [39] utilized optical flow to roughly detect and extract patch features on each frame and pooled all the patch features together. However, rough detection is likely to cause false judgment on the objects. In addition, they only used patch features which will overlook context information. Hence, we exploit pre-trained Faster R-CNN model [19] that detects objects more accurately, and generates variable-size bounding boxes according to the actual size of objects. At the same time, we propose a STAT method, which use two type appearance features such as frame-level global features and object-level local features simultaneously. Thus, our method can capture more significant details while keeping global context information.

3 EXPLOITING SPATIAL-TEMPORAL ATTENTION IN VIDEO DESCRIPTION

In this section, we delve into the main contributions of this paper and propose a method for exploiting spatial-temporal attention in video description.

3.1 Overall Framework

We build our video description framework based on the popular ConvNet + LSTM architecture [15, 17, 22, 29, 38], which consists of two neural networks: the encoder and decoder as shown in Figure 2. The encoder network is intended for learning a good visual representation, and the decoder network generates a corresponding description from the output of the encoder. In the encoder network, we extract global features vg_i and local features vl_i from the video frames while extracting the motion features vm_i from the video clips. Thus, a video inputted to encoder network can be converted into a feature set $V = \{v_1, ..., v_k\}$, where each $v_i = \{vg_i, vl_i, vm_i\}$. Moreover, we intend to fuse vg_i and vm_i into $v[gm]_i$ in that both of them reflect context information. In the decoder network, visual features can be converted into a word sequence $Y = \{y_i, y_2, ..., y_m\}$, which describes the video content.

In the image description, all the visual features are encoded into a single feature vector into the LSTM unit. But for a video, it is obviously unrealistic to cram the visual information of the whole video into a single vector. Therefore, we will follow the implementation of [37] to introduce visual features with the generation of each target word. It is necessary to add a new visual input part $\varphi_t(V)$ to the LSTM unit, which is formulated as follows:

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

$$f_t = \sigma(W_f E[y_{t-1}] + U_f h_{t-1} + A_f \varphi_t(V) + b_f); \quad (2)$$

$$o_t = \sigma(W_o E[y_{t-1}] + U_o h_{t-1} + A_o \varphi_t(V) + b_o); \quad (3)$$

$$g_t = \sigma(W_g E[y_{t-1}] + U_g h_{t-1} + A_g \varphi_t(V) + b_g); \quad (4)$$

$$c_t = c_{t-1} \odot f_t + i_t \odot g_t; \quad (5)$$

$$h_t = o_t \odot \phi c_t, \quad (6)$$

where σ is a *sigmoid* activation function, ϕ is a *tanh* function, y_{t-1} is the previous word, h_{t-1} is the previous hidden state, $\varphi_t(V)$ is the encoder representation. E is a word embedding matrix, and we denote by $E[y_{t-1}]$ an embedding vector of word y_{t-1}. Besides, W_i (W_o, W_f, W_g), U_i (U_o, U_f, U_g), A_i (A_o, A_f, A_g) and b_i (b_o, b_f, b_g) are, in order, the weight matrices for the input, the previous hidden state, the context from the encoder and the bias. Finally, the probability distribution of a series of target words at each time will be obtained through a single hidden layer:

$$\hat{y_t} = softmax(U_y \phi(W_y[h_t, \varphi_t(V), E[y_{t-1}]]) + b_y), \quad (7)$$

where $[h_t, \varphi_t(V), E[y_{t-1}]]$ denotes the concatenation of the three vectors.

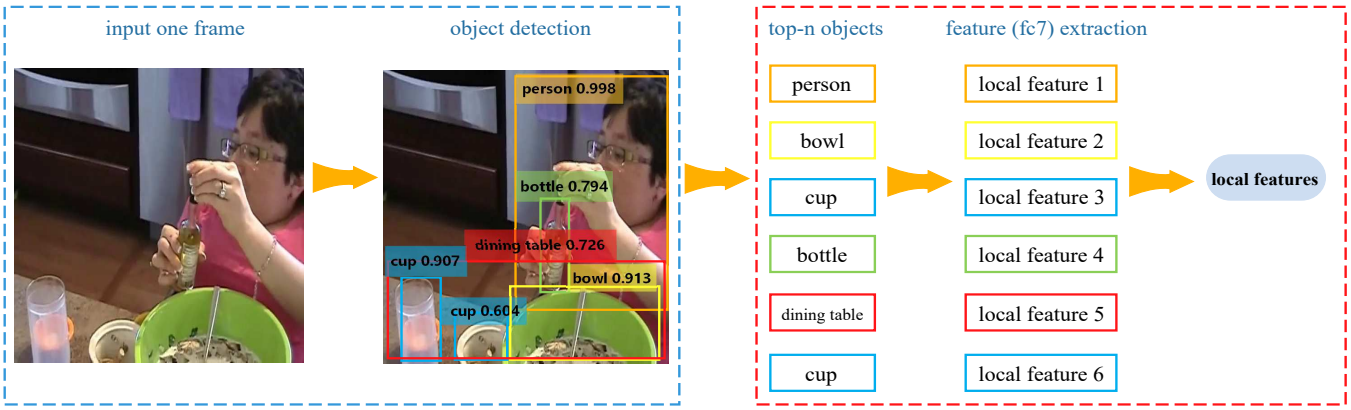

Figure 3: Example detection using Faster R-CNN on a frame of a video (Left: blue dotted wireframe). The process of local feature extraction (Right: red dotted wireframe).

3.2 Object Detection and local Feature Extraction

Nowadays, object detection has attracted more and more attention in the field of compute vision [4, 32–35, 41, 42, 44]. Meanwhile, extraction of multiple local features vl_i is a key component of our STAT method in both training and testing. In order to detect and locate multiple objects on video frames, Yu *et al.* [39] exploited optical flow to roughly detect and extract n image patches of size 220×220 along the lower part of the box border (*c.f.* Section 2 for details). In addition, Donahue *et al.* [5] and Rohrbach *et al.* [20] designed a specialized hand detector which could accurately detect and locate multiple objects. However, we find that both of their methods require a lot of engineering efforts. Inspired by the recent success of Region Proposal Network (RPN) [19] and Region-based Convolutional Neural Networks (R-CNNs) [7] in object detection, we will exploit Faster RCNN model [19] to directly detect multiple objects from the input video frames.

We use a Faster R-CNN model which takes an image (of any size) as input and outputs a set of rectangular object proposals, each with a class confidence score. The higher the score, the more likely there is an object for a certain class. In order to alleviate unnecessary computation complexity, we have made some efforts to address it. First, we reduce the number of proposals, which is the maximum number bounding boxes for a frame, from 300 (*e.g.* default setting) to 100, because it still leads to a competitive result when using the top-ranked 100 proposals at a test-time [19]. In effect, the average number of proposals is smaller after NMS. Besides, considering that there is little change between a subset of frames of a video, we select 28 equally-spaced frames to detect possible objects, thus further alleviating computation complexity. The Faster R-CNN model is pre-trained on MS COCO detection dataset and can detect 80 objects. Compared to the methods [5] and [39], the Faster R-CNN model not only detects multiple objects more accurately, but also decreases the detection time largely. Last but not least, it generates variable-size bounding boxes which is more flexible for object detection.

After detecting objects on each video frame, we select the top-n objects to represent important local objects according to their class confidence scores $\{s_1, s_2, ..., s_n\}$. Then, we represent each object as a 4096-dimensional local feature, which is extracted from the fc7 layer of the Faster R-CNN network. Finally, we obtain a set of local features $vl_i = \{vl_{i1}, ..., vl_{in}\}$ where $vl_{ij} \in \mathbb{R}^{4096}$ on each frame. Figure 3 shows results of object detection (blue dotted wireframe) and local feature extraction (red dotted wireframe) on a frame of one video in Figure 5.

3.3 Spatial-Temporal Attention

Visual attention is an important mechanism in the visual system of primates and human beings. It is a feedback process that selectively maps a representation from the early stages in the visual cortex into a more central non-topographic representation that contains the proprieties of only particular regions of objects in the scene [38]. Thus, we exploit visual attention mechanism of human beings to design a spatial-temporal attention (STAT) method. Our proposed method enables the decoder to first focus on specific objects on video frames, and then studies the interaction between objects over time during the process of video description.

As shown in Figure 4, the input of STAT are composed of global-motion features, local features and model status information. The outputs of STAT unit are dynamic visual representation, which feed each iteration of LSTM decoder. First, we let local features vl_i go through Layer 1 and exploit spatial attention to select semantically more relevant local features $\Psi_i^t(VL)$ to Layer 2; Second, in Layer 2, we utilize temporal attention to generate global-motion temporal representation $\varphi_t(VGM)$ and local temporal representation $\varphi_t[\Psi(VL)]$ from global-motion features $v[gm]_i$ and local features $\Psi_i(VL)$ respectively; Finally, global-motion temporal representation and local temporal representation go through Layer 3 and are concatenated into a new temporal representation $\varphi_t(V)$ to feed each iteration of the LSTM decoder.

Spatial Attention: We exploit spatial attention mechanism to encode the top-n local features $vl_i = \{vl_{i1}, ..., vl_{in}\}$ of each frame obtained from Section 3.2 into variable-length local features:$\Psi(VL) = \{\Psi_1(VL), ..., \Psi_k(VL)\}$. Each of the $\Psi_i(VL)$ is the dynamic weighted

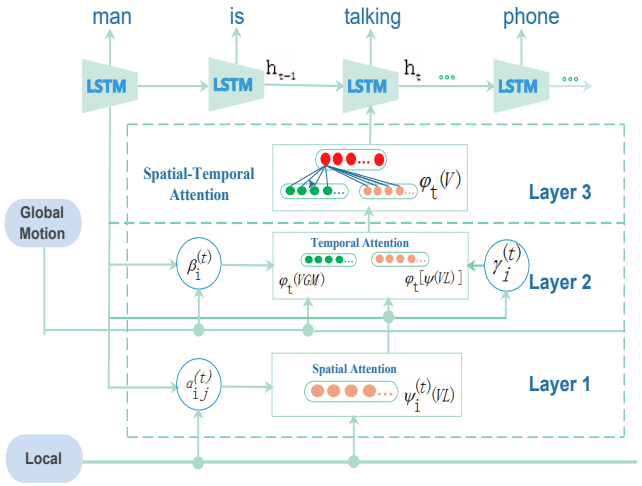

Figure 4: The STAT unit is shown. Local features, global-motion features, and model status information are inputted into STAT, and STAT generates dynamic visual features to each iteration of the LSTM decoder. Layer 1 indicates that spatial attention is applied to local features. Layer 2 expresses temporal attention on global-motion features and local features. Layer 3 represents the new temporal representation is fused by two temporal representation.

sum of all the n local features through a spatial attention mechanism:

$$\Psi_i^{(t)}(VL) = \sum_{j=1}^{n} \alpha_{ij}^{(t)} vl_{ij}, \tag{8}$$

where $\sum_{j=1}^{n} \alpha_{ij}^{(t)} = 1$ and $\alpha_{ij}^{(t)}$ are computed at each time step t inside the LSTM decoder. We refer to $\alpha_{ij}^{(t)}$ as the spatial attention weights at time t. The spatial attention weights $\alpha_{ij}^{(t)}$ reflect the relevance of the j-th local features in the input video given all the previous words, i.e. $y_1, ..., y_{t-1}$. Hence, we design a function that takes as input from the previous hidden state of the LSTM decoder, and the j-th local features and returns the unnormalized relevance scores $e_{ij}^{(t)}$:

$$e_{ij}^{(t)} = w_l^T tanh(W_e h_{t-1} + U_e vl_{ij} + z_e), \tag{9}$$

where w_l^T, W_e, U_e, z_e are the parameters to be learned by our model and shared by all the local features at all the time steps.

Once the relevance scores $e_{ij}^{(t)}$ for all the local features $j = 1, ..., n$ are computed, we normalize them through *softmax* function to obtain the $\alpha_{ij}^{(t)}$:

$$\alpha_{ij}^{(t)} = exp\{e_{ij}^{(t)}\} / \sum_{j'=1}^{n} exp\{e_{ij'}^{(t)}\}. \tag{10}$$

In conclusion, the spatial attention mechanism allows the decoder to selectively focus on more salient objects by increasing the attention weights of the corresponding local features.

Temporal Attention: We encode the variable-length global-motion features $V[GM] = \{v[gm]_1, ..., v[gm]_k\}$ and local features $\Psi(VL) = \{\Psi_1(VL), ..., \Psi_k(VL)\}$ into a sentence-length temporal representation $\varphi(V) = \{\varphi_1(V), ..., \varphi_m(V)\}$. Each $\varphi_t(V)$ is a concatenation of global-motion temporal representation and local temporal representation:

$$\varphi_t(V) = \{\varphi_t(VGM), \varphi_t[\Psi(VL)]\}, \tag{11}$$

where $\varphi_t(VGM)$ is the dynamic weighted sum of all the k global-motion features, and $\varphi_t[\Psi(VL)]$ is the dynamic weighted sum of all the k local features through an temporal attention mechanism:

$$\varphi_t(VGM) = \sum_{i=1}^{k} \beta_i^{(t)} v[gm]_i; \tag{12}$$

$$\varphi_t[\Psi(VL)] = \sum_{i=1}^{k} \gamma_i^{(t)} \Psi_i(VL), \tag{13}$$

where $\sum_{i=1}^{k} \beta_i^{(t)} = 1$ and $\sum_{i=1}^{k} \gamma_i^{(t)} = 1$. We compute $\beta_i^{(t)}$ and $\gamma_i^{(t)}$ respectively at each time step t inside the LSTM decoder. We refer to $\beta_i^{(t)}$ and $\gamma_i^{(t)}$ as the temporal attention weights at time t.

Similarly, we design two temporal attention functions to calculate unnormalized relevance scores $b_i^{(t)}$ and $c_i^{(t)}$, which take the previous hidden state, the i-th global-motion features and the i-th local features as inputs:

$$b_i^{(t)} = w_k^T tanh(W_b h_{t-1} + U_b v[gm]_i + z_b); \tag{14}$$

$$c_i^{(t)} = w_r^T tanh(W_c h_{t-1} + U_c \Psi_i(VL) + z_c), \tag{15}$$

where w_k^T, W_b, U_b, z_b and w_r^T, W_c, U_c, z_c are shared by all the global-motion features and local features respectively.

Then, we also normalize them through the *softmax* function:

$$\beta_i^{(t)} = exp\{b_i^{(t)}\} / \sum_{i'=1}^{k} exp\{b_{i'}^{(t)}\}; \tag{16}$$

$$\gamma_i^{(t)} = exp\{c_i^{(t)}\} / \sum_{i'=1}^{k} exp\{c_{i'}^{(t)}\}. \tag{17}$$

Therefore, the temporal attention mechanism allows the decoder to selectively focus on a subset of frames by increasing the attention weights of the corresponding global-motion features and local features. In conclusion, the two proposed attention mechanism are integrated orderly into an encoder-decoder neural video caption generator, which can pay more attention to how to predict the salient objects more precisely with high recall while attending to semantically more relevant video frames.

4 EXPERIMENT

4.1 Dataset and Evaluation Metrics

Dataset: We conduct the experiments on two video captioning benchmarks: MSVD [1] and MSR-VTT-10K [30]. The MSVD has 1970 video clips with a variety of human annotated language descriptions. The dataset contains a total of 80839 sentences which can be divided into 13010 separate words. According to the method of [37], we spilt a training set of 1200 video clips, a validation set of 100 clips, and a test set consisting of remaining clips. The MSR-VTT-10K [30] contains 10,000 video clips, which is the most challenging dataset for video captioning to date. We use the official spilt with 6513 videos for training, 497 for validation and 2990 for testing.

We report the experimental results on both the validation and test splits on MSR-VTT-10K.

Evaluation Metrics: Various methods for the evaluation of generated sentences have been employed, such as BLEU [16], METEOR [11] and CIDEr [26]. BLEU is the most popular metric for the evaluation of machine translation which is only based on the n-gram precision. METEOR is based on the harmonic mean of unigram precision and recall with which the recall is weighted higher than precision. It is designed to fix some of the problems of BLEU metric. Different with the BLEU metric, the METEOR seeks correlation at the corpus level. CIDEr is designed for evaluating image descriptions using human consensus. We utilize the Microsoft COCO evaluation server [2] to obtain all the results in this paper, which makes our results directly comparable with the previous work.

4.2 Implementation Details

Features Extraction: For frame-level global features, we adopt 1024-dimension $pool5/7 \times 7\_s1$ layer from GoogLeNet [23] and denote them as $VG = \{vg_1, ..., vg_k\}$. For object-level local features, we denote them as $VL = \{vl_1, ..., vl_k\}$. These local features are extracted by Faster R-CNN [19] (*c.f.* Section 3.2 for details). In this paper, we set n to 8 on the MSVD. In order to reduce the amount of computation and memory consumption, we reduce the number of n to 5 on the MSR-VTT-10K because the number of object contained in an image is usually below 10. For motion features, we use the 4096-dimensional fc6 layer from C3D [24] and pre-trained on the Sports-1M video dataset [9]. On the MSVD, we take continuous 16 frames as the input short clips for the C3D. On the MSR-VTT-10K, we increase the interval and take continuous 32 frames as the input short clips for the C3D. The C3D features are denoted as $VM = \{vm_1, ..., vm_{k'}\}$. At last, we select 28 equally-spaced frame global features, local features and clip motion features as visual inputs.

Model and Training: An overview of our video description framework is shown in Figure 2. We use one-layer LSTM unit and set hidden layer size as 1024. The word embedding size is set to 512 and learning rate is set to 2×10^{-4} empirically. In training, all video description generation models are trained end-to-end by minimizing the penalized negative log-likelihood. Training continued until the validation log-probability stopped increasing for 6,000 updates. Then, we use the Adadelta algorithm [40] with the gradient computed by the back propagation algorithm, which is widely used for optimizing attention model to update the parameters. Finally, we estimate the parameters by maximizing the log-likelihood:

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{t_m} = logp(y_i^m | y_{<i}^m, x^m, \theta), \tag{18}$$

where there are N training video-description pairs (x^m,y^m),and each description y^m is t_m words long.

4.3 Experimental results

Baseline Methods: First, we compare our spatial-temporal attention method (STAT) with the method using none local features (TAT-NL). Second, we compare STAT with the method using none attention (NAT), which adopts simple averaging strategy for all the features. Finally, we compare STAT with using only temporal attention method (TAT), which uses the averaging strategy for local features.

State-of-the-art Methods: On MSVD, we compare STAT with five methods: TA [37], LSTM-E [15], h-RNN [39], HRNE [14] and M-Fusion [8]. TA is the first work exploring temporal attention in video description. LSTM-E simultaneously explores learning of LSTM and visual semantic embedding. h-RNN presents an approach that exploits hierarchical RNNs to tackle the video captioning problem. HRNE models video temporal information using a hierarchical recurrent encoder. M-Fusion explores attention model which selectively attends not only specific times, but also specific modalities of input. Little work have been done on MSR-VTT-10K. We compare STAT with three methods: SA-LSTM [30], C3D+ResD [18] and v2t_nagvigator [6]. SA-LSTM is the basic method of publication on MSR-VTT-10K, but it is done on a different split from ours. C3D+Res studies the fusion of multiple features. v2t_nagvigator on this dataset has the best results.

Results on MSVD: We report the results on MSVD in Table 1. Our STAT method achieves the best BLEU and CIDEr scores among all the methods. BLEU4 has shown good performance for corpus-level comparisons over which a high number of n-gram matches exit [2]. For CIDEr, this is a consensus-based metric, which rewards a sentence for being similar to the majority of human written descriptions. Thus, the description based on our STAT method can be as accurate as possible on the basis of maintaining human language habits.

We also compute the relative improvements obtained by STAT method. Compared to the TAT-NL method, our STAT method obtains 10.1% relative improvements in terms of BLEU4, 2.8% relative improvements in terms of METEOR, and 8% relative improvements in terms of CIDEr. The results show that we integrate local features into global features and motion features indeed improve recognition and localization multiple small objects on video frames. Compared to the NAT method, our STAT method obtains 23.1% relative improvements in terms of BLEU4, 3.8% relative improvements in terms of METEOR, and 7.9% relative improvements in terms of CIDEr. In contrast, compared to the NAT method, although TAT method also has relative improvements in terms of BLEU, it has shown worse results in other two metrics. The comparison result from STAT method and TAT method shows that temporal attention is hard to distinguish the small objects on video frames. Hence, the spatial attention is an essential part of the video description method. In conclusion, we observe that the improvements brought by exploiting spatial and temporal information are complimentary, with the best performance achieved when both the spatial attention and the temporal attention are used together.

Results on MSR-VTT-10K: We report the results on MSR-VTT-10K in Table 2. We find that our STAT method has less improvements over TAT-NL in that this dataset has more objects than MSVD. Thus, we expect to have even better performance if our number of object detection is increased. In addition, our STAT method also has better improvements over NAT method. In contrast, we note that TAT method has shown worse results in all the evaluation metrics of Test split. Hence, we further argue that temporal attention hardly enables decoder to attend to small objects on each frame.

Table 1: Performance evaluation on MSVD

	B@1	B@2	B@3	B@4	METEOR	CIDEr
TAT-NL (G+C)	0.803	0.676	0.572	0.464	0.318	0.625
NAT (G+C+R-fc7)	0.764	0.627	0.521	0.415	0.315	0.629
TAT (G+C+ R-fc7)	0.773	0.642	0.540	0.432	0.307	0.597
STAT (G+C+ R-fc7)	**0.826**	**0.714**	**0.616**	**0.511**	0.327	**0.675**
TA[37](G+3D CNN)	0.800	0.647	0.526	0.419	0.296	0.517
LSTM-E[15](V+C)	0.788	0.660	0.554	0.453	0.310	-
h-RNN[39](V+C)	0.815	0.704	0.604	0.499	0.326	0.658
HRNE[14](G+C)	0.811	0.686	0.578	0.467	**0.339**	-
M-Fusion[8](V+C)	0.811	0.703	0.607	0.499	0.318	0.634

[1] (G=GoogLeNet, C=C3D, R-fc7=Faster R-CNN fc7, V=VGG)

Table 2: Performance evaluation on MSR-VTT-10K

	Test split			Valid split		
	B@4	METEOR	CIDEr	B@4	METEOR	CIDEr
TAT-NL (G+C)	0.371	0.264	0.398	0.379	0.269	0.405
NAT (G+C+R-fc7)	0.348	0.250	0.365	0.347	0.252	0.350
TAT (G+C+ R-fc7)	0.343	0.243	0.319	0.358	0.247	0.316
STAT(G+C+ R-fc7)	0.374	0.266	0.415	0.380	0.271	0.402
v2t_nagvigator[6]	**0.408**	0.282	0.448	**0.394**	**0.275**	**0.480**
C3D+Res[18]	-	-	-	0.385	0.267	0.411
SA-LSTM[30]	0.405	**0.299**	-	-	-	-

[1] (G=GoogLeNet, C=C3D, R-fc7=Faster R-CNN fc7)

In this more challenging dataset, our experimental results are not satisfactory due to STAT method only outperforms C3D+Res [18] with METEOR score. As to v2t_nagvigator [6] and SA-LSTM [30], both of them have higher scores than our method. There are two likely reasons accounting for the limited improvement on MSR-VTT-10K corpus. First, the challenging corpus included diverse and tiny objects may lead to imprecision when detecting objects, so weakening the strength of spatial attention. Besides, SA-LSTM is done on a different split from our available data, which makes it unsuitable for comparison. For v2t_nagvigator, it exploits sentence re-ranking method which promotes relevant captions by re-scoring a list of candidate sentences [6]. However, our goal is to improve the visual encoder, which is quite different from their research. In conclusion, the attention model in encoder part of video description is worth further studying.

4.4 Qualitative Analysis

Although the evaluation mechanisms introduced in [2] can reflect the degree of matching between the descriptions generated by our STAT method and the reference descriptions by human, the scores in Table 1 and Table 2 are not straightforward for understanding of our model. Thus, we visualize the spatial attention weights and temporal attention weights for some video clips from MSVD and MSR-VTT-10K, as shown in Figure 5. SA and TA represent the spatial attention results and temporal attention results in our STAT method, respectively. We also present the ground truth descriptions, the descriptions generated by temporal attention (TAT) and the descriptions generated by spatial-temporal attention (STAT) .

In Figure 5, we can clearly see that the description generated using the STAT method is able to capture more details (*e.g.* 'paper airplane', 'tv show', 'bowl', 'bear') and less misrecognition, because our model is able to attend to those key evidences. We observe that our model can not only focus on key frames relating to each word, but also can focus on some key objects on the frames. For example, in the video clip *No.1*, when generating the word 'boy', our model focuses on the first frame through temporal attention. Meanwhile, the discriminative face area of the boy is focused through the spatial attention. After that, when generating the word 'dog', our model switch attention to the third and fourth frames according to previous generated words. We also find that using temporal attention method alone, the 'dog' is misrecognized as the 'baby'. In the video clip *No.4*, it is difficult to judge the place where 'a man and woman talking' when using temporal structure alone. In contrast, our STAT method takes pride in identifying the 'dog' accurately, and successfully recognize the 'couch' and other details, because we incorporate the local features and automatically attend to related local objects. These examples further confirms that integrating spatial attention into temporal attention to select the object-level local features are critical in the video description.

5 CONCLUSION

In this paper, we have studied the existing problems of video description in depth in terms of detail missing and recognition error. We identify and underscore the importance of order between spatial structure on each frame and temporal structure on consecutive frames. To this end, we propose a novel video description method

Figure 5: Eight sample videos on MSVD and MSR-VTT-10K and their ground truth descriptions, the descriptions generated by temporal attention (TAT) and the descriptions generated by spatial-temporal attention (STAT). The white aperture area on each frame represents the change in the degree of importance of each local object 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 temporal attention stage.

which integrates the local objects information into global and motion information based on spatial-temporal attention mechanism. Compared to existing methods, our STAT method can pay attention to multiple prominent objects, thus generating detailed and accurate descriptions. Extensive experiments conducted on two well-known benchmarks show that the integrating of local features and spatial-temporal attention mechanism are critical in the video description. In the future, we will model the visual relation between multiple objects on consecutive frames.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported by National Nature Science Foundation of China (61671196, 61525206, 61327902), Zhejiang Province Nature Science Foundation of China LR17F030006.

REFERENCES

[1] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 190–200.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[3] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2422–2431.

[4] Gong Cheng, Peicheng Zhou, and Junwei Han. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 54, 12 (2016), 7405–7415.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[6] Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees GM Snoek. 2016. Early Embedding and Late Reranking for Video Captioning. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1082–1086.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[8] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R Hershey, and Tim K Marks. 2017. Attention-Based Multimodal Fusion for Video Description. *arXiv preprint arXiv:1701.03126* (2017).

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[10] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[11] Michael Denkowski Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *ACL 2014* (2014), 376.

[12] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image Caption with Global-Local Attention. In *AAAI*.

[13] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* (2014).

[14] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1029–1038.

[15] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4594–4602.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[17] Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. 2015. Learning temporal embeddings for complex video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*. 4471–4479.

[18] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal Video Description. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1092–1096.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[20] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition*. Springer, 184–195.

[21] Rakshith Shetty and Jorma Laaksonen. 2015. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. *arXiv preprint arXiv:1512.02949* (2015).

[22] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Yan Song, Zechao Li, and Liyan Zhang. 2017. Concurrence-Aware Long Short-Term Sub-Memories for Person-Person Action Recognition. *arXiv preprint arXiv:1706.00931* (2017).

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[24] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. C3D: generic features for video analysis. *CoRR, abs/1412.0767* 2 (2014), 7.

[25] Daksh Varshneya and G Srinivasaraghavan. 2017. Human Trajectory Prediction using Spatially aware Deep Attention Models. *arXiv preprint arXiv:1705.09436* (2017).

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[27] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. 4534–4542.

[28] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729* (2014).

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[30] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.. In *ICML*, Vol. 14. 77–81.

[32] C Yan et al. 2017. Effective Uyghur language text detection in complex background images for traffic prompt identification. *IEEE Transactions on Intelligent Transportation Systems* (2017).

[33] C Yan et al. 2017. Supervised hash coding with deep neural network for environment perception of intelligent vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2017).

[34] Chenggang Yan, Yongdong Zhang, Jizheng Xu, Feng Dai, Liang Li, Qionghai Dai, and Feng Wu. 2014. A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Processing Letters* 21, 5 (2014), 573–576.

[35] Chenggang Yan, Yongdong Zhang, Jizheng Xu, Feng Dai, Jun Zhang, Qionghai Dai, and Feng Wu. 2014. Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 12 (2014), 2077–2089.

[36] Yichao Yan, Bingbing Ni, and Xiaokang Yang. 2017. Predicting Human Interaction via Relative Attention Model. *arXiv preprint arXiv:1705.09467* (2017).

[37] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515.

[38] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.

[39] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4584–4593.

[40] Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR* abs/1212.5701 (2012). http://arxiv.org/abs/1212.5701

[41] Dingwen Zhang, Junwei Han, Lu Jiang, Senmao Ye, and Xiaojun Chang. 2017. Revealing event saliency in unconstrained video collection. *IEEE Transactions on Image Processing* 26, 4 (2017), 1746–1758.

[42] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. 2016. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision* 2, 120 (2016), 215–232.

[43] Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. 2017. Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[44] Xishan Zhang, Hanwang Zhang, Yongdong Zhang, Yang Yang, Meng Wang, Huanbo Luan, Jintao Li, and Tat-Seng Chua. 2016. Deep fusion of multiple semantic cues for complex event recognition. *IEEE Transactions on Image Processing* 25, 3 (2016), 1033–1046.