Catching the Temporal Regions-of-Interest for Video Captioning

Anonymous Author(s)

ABSTRACT

As a crucial challenge for video understanding, exploiting the spatial-temporal structure of video has attracted much attention recently, especially on video captioning. Inspired by the insight that people always focus on certain interested regions of video content, we propose a novel approach which will automatically focus on regions-of-interest and catch their temporal structures. In our approach, we utilize a specific attention model to adaptively select regions-of-interest for each video frame. Then a Dual Memory Recurrent Model (DMRM) is introduced to incorporate temporal structure of global features and regions-of-interest features in parallel, which will obtain rough understanding of video content and particular information of regions-of-interest. Since the attention model could not always catch the right interests, we additionally adopt semantic supervision to attend to interested regions more reasonably. We evaluate our method for video captioning on two public benchmarks: Microsoft Video Description Corpus (MSVD) and Montreal Video Annotation (M-VAD). The experiments demonstrate that catching temporal regions-of-interest information really enhances the representation of input videos and our approach obtains the state-of-the-art results on popular evaluation metrics like BLEU-4, CIDEr, and METEOR.

KEYWORDS

Deep Learning, Video Captioning, Regions of Interest, LSTM

ACM Reference format:
Anonymous Author(s). 2017. Catching the Temporal Regions-of-Interest for Video Captioning. In *Proceedings of ACM multimedia, Mountain View, CA USA, October 23-27, 2017 (MM'17),* 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Visual captioning targets to automatically generate natural language descriptions for images or videos, based on visual content understanding [12, 16, 31], which is regarded as a crucial challenge in multimedia content analysis and computer vision. Generating meaningful natural language for visual content plays an important role in machine understanding and may benefit many practical applications, such as human-robot interaction, impaired people auxiliary, and video retrieval. Thus it has attracted much attention recently.

Visual content contains complicated information, such as motions, objects, and scenes, as well as relationships among them and their attributes. The earliest approaches [9, 13, 14, 17] for visual captioning firstly predict semantic concepts or words (e.g., objects, subjects, and verbs) by different classification methods. Then they employ a pre-defined sentence template to generate a natural language description, which may be grammatically correct but limited to specific structures or fixed length. With the development of deep learning, recent researches [8, 18, 19] turn to employing different neural networks for constructing visual representation as well as a recurrent neural network, especially Long Short-Term Memory (LSTM) [11], for variable-length captions generation.

As a consecutive visual content, videos deliver rich information, yet include a lot of clutter. The early work [30] for video captioning simply leveraged a mean pooling over frames to create video representation and used a recurrent neural network for sentence generation, which has the drawback of losing temporal information. Therefore, S2VT [29] proposed a stacked LSTM framework to encode consecutive frame inputs. But it remains another problem that LSTM shows a poor capacity of capturing long-range dependencies. The later work [20] designed a hierarchical recurrent neural encoder to incorporate video temporal information during a long range and fused consecutive inputs at different levels. Baraldi *et al.* [3] proposed a specific LSTM cell to detect discontinuous segments boundaries and discovered hierarchical video structures. These methods demonstrate that exploiting temporal structure of video inputs improves the discriminative ability of video representations. However, they treat consecutive video inputs in the same way and do not consider different characteristics they own for understanding the video content.

Different segments and regions of videos may have different impacts on video content understanding, as they usually receive different attentions from people. When we are asked to utter the content of a certain video, we have a large possibility to only mention the semantic of the regions of interest. For instance, a video about playing football in a stadium may contain events about the game as well as the reactions from audiences. We always talk about what happens about the game, rather than how the audience are cheering and making noises. Following similar insights, when dealing with video captioning, it would be better to automatically select significant or discriminative regions for describing the visual content. The previous work [33] used a temporal attention mechanism to adaptively select the most important segments for word prediction, which seems similar to focus on some interested snippets in human visual system. And some methods [6, 31] of image captioning also utilized regions-of-interest, which has shown better performance on recognizing details and object attributes. Though these methods attempt to construct flexible visual understanding system with attention mechanism, few work catches interested regions for each video frame and incorporates their temporal cues into the process of captions generation.

In this paper, we propose a novel video captioning method, which will adaptively select interested regions from each frame and utilize their temporal structure. As shown in Fig. 1, we firstly employ

MM'17, Mountain View, CA USA

DOI: 10.1145/nnnnnnn.nnnnnnn

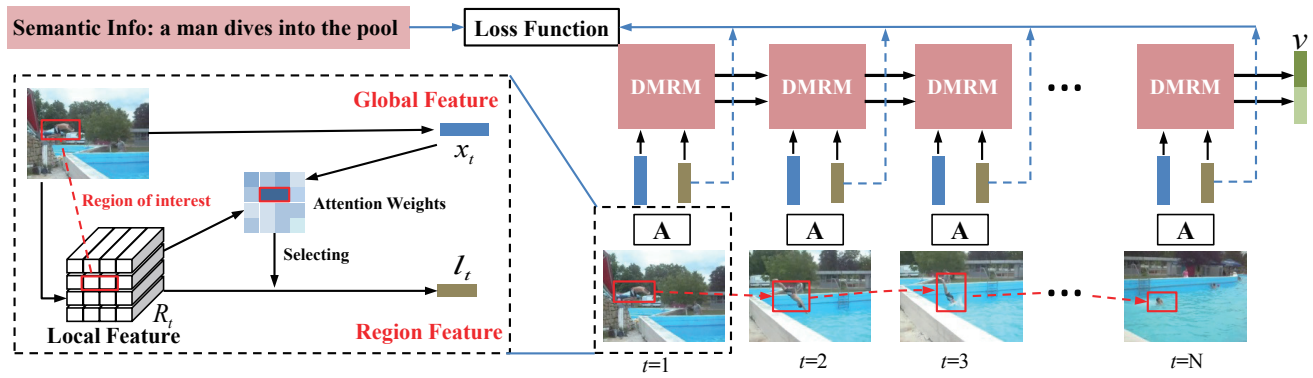


Figure 1: The overall framework of our method. "x_t", "R_t" and "l_t" represent global feature, local feature and interested region feature at time t respectively. The red boxes denote regions of interest. "A" denotes the module to select regions-of-interest. "N" is the length of input frames. "v" is final video representation.

a well-designed attention model to computer attention weights for current local feature, which represent the extent of interest of corresponding regions. And the regions of interest are selected according to the attention weights. Then the features of global content and regions-of-interest will be fed into a Dual Memory Recurrent Model (DMRM) to construct integral video representation recursively. Since the features extracted by a pre-trained CNN represent rough semantic concepts of corresponding video frames, we simply regard global CNN features as rough video frame understandings. Our attention model devotes to focus on semantically important regions, guided by the correlation with the current video frame understanding. For example, if the global feature depicts the concept "a people dives into a pool", our attention model attempts to select regions involving details of "people", "dives", or "pool". These regions seem to be more discriminative for identifying whether the rough concept is right or not. Since coarse global features extracted by the pre-trained CNN sometimes cause confused understanding, our attention model could not always catch the correct interests. Towards this problem, we additionally employ a semantic supervised method on the attention results to make the attention process more reasonably.

Our contributions of this work are as follows: (1) A well-designed attention model to adaptively select interested regions for each frame; (2) A Dual Memory Recurrent Model which simultaneously incorporates temporal structures for global features and interested region features; (3) A semantic supervision approach on attention model to modify interests with correct semantic concepts.

2 RELATED WORK

Describing visual content with natural language has been taken as a grand challenge for machine understanding. In recent years, several approaches have been proposed to deal with it. Early visual captioning approaches [9, 13, 23] attempt to detect semantic concepts (e.g., objects, subjects, and verbs) with multiple visual classifiers, and fill predicted concepts in pre-defined sentence templates for caption generation. These methods always create sentences with specific syntactical structure, which may lose flexibility of natural language and highly depend on pre-defined templates of sentence. Inspired by the crucial achievements of recurrent neural network in machine translation [1, 25], research on visual captioning attemps to employ recurrent neural network (RNN) as language model for generating sentences.

Recent RNN-based video captioning approaches have greatly enhanced the performance than early approaches [9, 23], owning to the merits that differentiable recurrent models can directly model variable-length inputs and outputs. Venugopalan *et al.* [30] simply utilized mean pooling over CNN features of frames and fed video representation into a two layer LSTM for generating captions. Donahue *et al.* [7] explored video content representation by using CRFs to get semantic predictions of activity, location, object, and tool. Then a LSTM layer was applied to generate sentences. Pan *et al.* [21] introduced visual-semantic embedding to video captioning framework and exploited relationship between sentence and visual content.

However, above attempts just regard video as a collection of images and neglect temporal dependencies between video frames. Although Yao *et al.* [33] proposed a temporal attention mechanism to automatically select the most relevant temporal segments for word prediction and creating video descriptions, it merely used a weight sum of frame-level CNN features. Therefore, plenty of later methods desire to incorporate better temporal structure into video representation by modeling video frame-level features as consecutive inputs with sequence learning method. S2VT [29] firstly employed a sequence-to-sequence model, which was already applied to machine translation [25], to address both video encoding and sentence decoding stages. In this method, a stacked LSTM is used to incorporate sequential video inputs and produce natural language description. HRNE [20] leveraged a hierarchical recurrent neural encoder to capture video temporal structure in a long range and model transitions between frames as well as segments. In order to explore temporal continuity of video inputs, Baraldi *et al.* [3] proposed a novel recurrent video encoding method to identify the boundaries of discontinued video snippets and modify corresponding input encoding layer. It incorporates a fine hierarchical structure of video and really performs well on movie captioning datasets. All these approaches mostly consider global information

for each frame but not exploit local details, which may be helpful for identifying accurate semantic concepts and their attributes.

In order to exploit local information, Yao *et al.* [33] utilized a 3D CNN to capture spatial structure of frames and Ballas *et al.* [2] stacked several GRUs to capture visual percepts from different convolutional levels, which contain different spatial information. Our method also exploits local context and spatial information. Differing from previous works, we focus on regions-of-interest for consecutive video inputs and incorporate temporal dependencies of these regions. By catching temporal information of interested regions, we can obtain more discriminative regional features, which will contribute to identifying objects and their attributes.

Besides exploring discriminative representation of visual content, combining visual attention with visual captioning system is also regarded as a potential direction. Xu *et al.* [31] introduced two variants of attention mechanism for image captioning to learn a latent alignment when predicting corresponding words. Yang *et al.* [32] enhanced current encoder-decoder frameworks using a review network, which incorporates visual content into a compact representation and uses attention mechanism both in encoder and sentence generating stages. Furthermore, You *et al.* [34] employed attention mechanism on semantic concepts or attributes obtained before and injected the most relevant semantic concepts into neural-based sentence generator. In this paper, we devise a well-designed attention model to adaptively select interested regions, which are most relevant to rough semantic concepts represented by global features for each frame.

3 THE PROPOSED APPROACH

Our approach targets to exploit temporal structure of both global features and regions-of-interest, which will depict rough semantic concepts and certain details of these concepts respectively. The main difference between our work and related methods [2, 10, 33] lies in adaptively and quickly selecting interested regions with the guidance of rough semantic concepts for every frame. This framework is a sequential process which will maintain temporal dependencies of both global features and region features. In this section, we firstly give an overall introduction of our approach. Then we explain how to select regions-of-interest (or salient regions) by a well-designed attention model. And we propose a Dual Memory Recurrent Model (DMRM), which incorporates temporal structure of both region and global features in parallel. Finally, we will devise a semantic supervision method to modify interested regions with semantic information of captions.

3.1 The Overall Framework

Given an input video with N frames and corresponding description sentence, our approach utilizes an encoder-decoder framework to model their mapping, which has been successfully applied to machine translation [25] and visual captioning [20, 29].

The encoder part of our framework is composed of two modules: regions-of-interest selecting and temporal structures encoding. The first module is implemented by an attention model, which regards global features as rough semantic concepts of current frame and employs the semantic information to guide the selection of regions-of-interest. The second module uses a Dual Memory Recurrent Model (DMRM) to incorporate temporal dependencies of global and region features. As shown in Fig. 1, we extract global feature x_t and local feature $R_t = (r_1^{(t)}, r_2^{(t)}, ..., r_k^{(t)})$ at time t by a pre-trained CNN, where k denotes the number of regions we extract from single video frame. With the guidance of rough semantic concepts depicted by global feature, we calculate the attention weights for k regions. The attention weights represent the correlation between semantic concepts and corresponding regions. The regions of interest will be selected according to attention weights. We denote the feature of regions-of-interest as l_t. Hence we obtain a sequence of global features $(x_1, x_2, ..., x_N)$ and consecutive regions-of-interest features $(l_1, l_2, ..., l_N)$. In view of the distinction between global and region features, we do not simply fuse them together at every time step. Instead, we employ the DMRM to incorporate their temporal structures separately. The details of DMRM will be elaborated in Section 3.3. At the end of encoder phrase, the outputs of DMRM will be concatenated as an integrated video representation v. We formulate the encoder process as the following equations:

$$l_t = f_{att}(R_t, x_t), \tag{1}$$

$$h_N^G, h_N^L = \text{DMRM}(x_1, x_2, ..., x_N, l_1, l_2, ..., l_N), \tag{2}$$

$$v = \text{Concat}(h_N^G, h_N^L), \tag{3}$$

where f_{att} represents regions-of-interest selector, composed by a specific attention model. DMRM denotes temporal structures encoder. N is the length of video inputs. h_N^G and h_N^L are the final outputs of DMRM at time N.

For decoding video representation to sentence description, we utilize a plain LSTM as language model according to related methods [29, 33]. A sentence with length T is represented as $(y_1, y_2, ..., y_T)$, encoded with one-hot vectors (1-of-D encoding, where D is the size of the vocabulary). We initialize hidden state and memory cell of the decoder LSTM by DMRM outputs at the last encoder step. Conditioned on the previous $t-1$ words of the caption and video representation v, the decoder is trained to predict next word y_t. We model the conditional probability of next word with a softmax layer:

$$p(y_t | y_{t-1}, y_{t-2}, ..., y_0, v) \propto (y_t^T W_y s_t). \tag{4}$$

W_y denotes parameters of a linear transformation layer which maps the outputs of LSTM to word space. s_t is the output of decoder LSTM at time t and v is the integrated video representation.

We define the objective function of generating words as negative logarithm of the likelihood:

$$Loss_1 = -\sum_{t=1}^{T} \log p(y_t | y_{t-1}, y_{t-2}, ..., y_0, v; \theta), \tag{5}$$

where θ are all the parameters of video captioning model and T is the length of sentence. We minimize the objective function over all the training set.

3.2 Selecting Regions-of-Interest

As mentioned before, video captioning usually utters important snippets and interested regions rather than all the frames and details. In order to model temporal structure of regions-of-interest, we firstly select interested regions for each frame. The features of interested regions are more discriminative for identifying particular

semantic concepts than global features. Now we introduce how we design the selector of regions-of-interest.

As shown in the left part of Fig. 1, we extract global feature x_t and local feature R_t from original video frame at time t. The global feature is regarded as a rough semantic representation of current video content but suffers from a lack of detail information. Local feature $R_t = r_1^{(t)}, r_2^{(t)}, ..., r_k^{(t)}$ contains features of k regions, where $r_i^{(t)}$ represents i-th region of the frame at time t. Since region feature only involves content of certain area, it contains less redundant information and is more discriminative on distinguishing different semantic concepts. In the paper, we design the regions-of-interest selector by a soft attention model, which is guided by global feature. In order to reduce computational cost, we use feature maps before fully-connected layer as our local features. Each regional feature in feature maps represents content of a certain frame location.

Since there will be many regions relevant to global feature, the representation of interested regions l_t at time t is formulated as dynamic weight sum of regional features in feature maps:

$$l_t = \sum_{i=1}^{k} \alpha_i^{(t)} r_i^{(t)}, \tag{6}$$

where $\sum_{i=1}^{k} \alpha_i^{(t)} = 1$. $\alpha_i^{(t)}$ is the attention weight at t-th time step for i-th region and is modeled as an attention neural network.

The relevance between global feature x_t and regional features will be measured by attention weights. They represent the extent of interest we have in corresponding regions. We use the following equation to calculate the relevance score:

$$e_i^{(t)} = w^\top \tanh(W_x x_t + W_l r_i^{(t)}), \tag{7}$$

where w, W_x, and W_l are all parameters. $r_i^{(t)}$ denotes the feature of the i-th region at time step t. The relevance scores will be normalized over k regions to calculate attention weights. The operation is formulated as:

$$\alpha_i^{(t)} = exp(e_i^{(t)}) / \sum_{j=1}^{k} exp(e_j^{(t)}). \tag{8}$$

3.3 Temporal Structures Encoding

With pre-trained CNN and the regions-of-interest selector, we can obtain global and interested region features for every frame. Because of the distinction between global and local features, simply blending them together at every time step may cause confusion. Thus, we devise a Dual Memory Recurrent Model (DMRM) to incorporate temporal structures of global and interested region features separately.

Our DMRM is based on the general LSTM structure which only contains one memory cell and one hidden state. The Fig. 2 illustrates the structure of our DMRM, composed of two memory cells, two hidden states and three control gates. The memory cells and hidden states are used to store temporal information of two different inputs and output current states with the control of three gates. Similar to LSTM, we also use one input gate i_t, one output gate o_t, and one forget gate f_t to steer input, output, and forget modules of DMRM respectively. At time instance t, given global feature x_t and

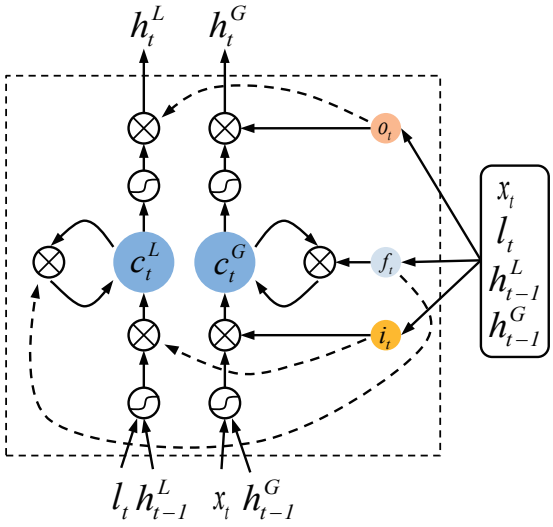


Figure 2: Dual Memory Recurrent Model (DMRM): two memory cells for incorporating temporal structures of global and local features in parallel.

interested region feature l_t, the control gates will be calculated in the following manner:

$$i_t = \sigma\left(W_{ih} h_{t-1}^G + U_{ih} h_{t-1}^L + W_{ix} x_t + U_{il} l_t\right), \tag{9}$$

$$f_t = \sigma\left(W_{fh} h_{t-1}^G + U_{fh} h_{t-1}^L + W_{fx} x_t + U_{fl} l_t\right), \tag{10}$$

$$o_t = \sigma\left(W_{oh} h_{t-1}^G + U_{oh} h_{t-1}^L + W_{ox} x_t + U_{ol} l_t\right), \tag{11}$$

where W_{*h},W_{*x} are learned weight matrices for hidden state h_{t-1}^G and global feature x_t. U_{*h} and U_{*l} are parameters for hidden state h_{t-1}^L and interested region feature l_t. σ is the sigmoid function. Under the control of three gates, the memory cells and hidden states of DMRM will be updated recursively:

$$\begin{aligned} c_t^G &= f_t \odot c_{t-1}^G + i_t \odot \phi\left(W_{ch} h_{t-1}^G + W_{cx} x_t\right), \\ h_t^G &= o_t \odot \phi\left(c_t^G\right), \end{aligned} \tag{12}$$

$$\begin{aligned} c_t^L &= f_t \odot c_{t-1}^L + i_t \odot \phi\left(U_{ch} h_{t-1}^L + U_{cx} x_t\right), \\ h_t^L &= o_t \odot \phi\left(c_t^L\right). \end{aligned} \tag{13}$$

For global feature, memory cell c_t^G and hidden state h_t^G are calculated by Eq. (12). W_{ch}, W_{cx} are the shared weight matrices to be learned. ϕ denotes hyperbolic tangent function tanh. And $\odot$ denotes an element-wise multiplication. We tackle memory cell c_t^L and hidden state h_t^L for regions-of-interest in the similar way with global feature. The process is formulated in Eq. (13), where U_{ch} and U_{cx} are weight matrices of DMRM to update memory cell and hidden state for interested regions.

3.4 Semantic Supervision

Although attention mechanism has been introduced in many previous methods [33, 34] and benefits most visual captioning systems, there remains improvement of catching the correspondence between attention alignment and human focus. The previous work [6]

has demonstrated that adding whether strong or weak supervision to attention mechanism improves the performance of implicitly-learned attention model. Since the global feature extracted by a pre-trained CNN may contain confusing and wrong information, our attention model sometimes will be leaded to a wrong direction and produce a chaotic weight alignment. Therefore, we add semantic supervision information to attention results for modifying them with correct semantic concepts.

In our method, we employ the ground-truth sentences of training videos as semantic supervision information. Similar to [32], we firstly apply a linear transformation on attention results for mapping region features to semantic space. Then we use a max-pooling over all results of linear mapping to get most significant signals. The process is formulated as:

$$B = max\_pool(W_b l_1, W_b l_2, ..., W_b l_N), \tag{14}$$

where W_b is a transfer matrix. $B = (b_1, b_2, ..., b_D)$ denote the most significant signals for each word and a multi-label margin loss function formulated by Eq. (15) will be optimized.

$$Loss_2 = \frac{1}{q} \sum_{j \in D} \sum_{i \neq j} max\left(0, 1 - (b_j - b_i)\right), \tag{15}$$

where D is the vocabulary size and q is the amount of all the valid i,j pairs. b_i denotes the signal for word i after max-pooling operation.

While training the model with semantic supervision, the whole loss function will be formulated as:

$$Loss = Loss_1 + \lambda Loss_2. \tag{16}$$

λ is the weight factor which trades off impacts of negative conditional log-likelihood $Loss_1$ and semantic supervision $Loss_2$. We set λ to 0.5 empirically.

4 EXPERIMENTS

Our proposed method is validated with extensive experiments on two public video captioning datasets: Microsoft Video Description Corpus (MSVD) [4] and Montreal Video Annotation Dataset (M-VAD) [27]. The first one is widely used by most of the state-of-the-art video captioning methods and another one is recently released large-scale movie description dataset, which contains more complicated activities and annotations.

4.1 The Datesets

The Microsoft Video Description Corpus (MSVD): The Microsoft Video Description Corpus is a popular benchmark for evaluating video captioning methods. It contains 1,970 video clips with about 8,000 English descriptions labeled by the Amzon Mechanical Turkers. The original dataset is composed of multi-lingual descriptions, but our experiments only consider English descriptions like many previous work [29, 33]. For consideration of fair comparisons with the state-of-the-art video captioning methods, we use the common split which divides the dataset into training, validation, and testing with 1,200 clips, 100 clips, and 670 clips respectively.

The Montreal Video Annotation Dataset (M-VAD): The Montreal Video Annotation Dataset (M-VAD) is a large-scale movie description dataset. It contains 46,523 movie clips, which are collected from 92 popular DVD movies. Each movie clip describes movie plot in a short duration with a single sentence. The ground truth sentences are generated in a semi-automatically transcribed way and have complicated grammar. Similar to previous works [20, 33], this paper uses the standard spilt provided by [27], which separates the M-VAD dataset into 36,921 clips for training, 4,651 clips for validation, and 4,951 clips for testing.

4.2 Evaluation Metrics

We employ three metrics to evaluate our captioning results: BLEU [22], CIDEr [28], and METEOR [15]. The BLEU measures the precision of n-grams between generated sentence and ground-truth descriptions. The METEOR measures the word correlation between candidate and reference sentences by generating an alignment based on exact token matching. The author of [28] empirically shows that METEOR is better than BLEU according to the consistency with human judgment. Therefore, we consider METEOR as our main metric for caption evaluation. The same as many previous captioning methods [20, 29, 33], we utilize the Microsoft COCO evaluation server [5] to compute all the scores in this paper for fair comparison. We also use beam search with size 5 during sentence generation as [36] does. Our experiments are implemented using the Torch Framework.

4.3 Experimental Settings

In order to construct discriminative video features, we employ the pre-trained GoogLeNet [26] and VGG-16 as our initial feature extractors. These two networks have been widely used on majority computer vision tasks and usually obtain good performance. To unify the input length of DMRM, we gather 10 frames from each video with different intervals. The global feature we use is the output of pool5/7x7_s1 layer with 1024-dimension vector from GoogLeNet. We use GoogLeNet inception 5b feature and VGG-16 pool5 feature as local features. Each regional part of local features depicts a certain region in corresponding video frame.

Before training the model, we firstly encode the ground truth captions to lower case and remove punctuation characters. Then the processed sentences are divided into word tokens which will be encoded as one-hot vectors later. The collection of word tokens composes the vocabulary of corresponding dataset. The vocabulary on the MSVD dataset contains 12,593 words and the M-VAD dataset has a vocabulary with about 16,000 words.

During the training phase, we add a begin-of-sentence <BOS> tag at the beginning of each sentence and end-of-sentence <EOS> tag at its end. The size of hidden units used in our DMRM is 512. The hidden size of decoder LSTM is 1024 and the word embedding size is set to 512. All the attention sizes in our method are set to 100 empirically. We rely on the RMSPROP algorithm to update parameters for better convergence, with learning rate 2×10^{-4} and other parameters using default values. The training batch size is set to 64. In order to alleviate overfitting problem when training our model, we apply dropout [24] with rate of 0.5 on the output of fully connected layers and the output of decoder LSTM. We clip the gradients to [-5,5] to prevent gradient explosion.

4.4 Experimental Results on MSVD

4.4.1 Comparison of Model Variants. On the MSVD dataset, we firstly discuss the performance of different variants of our method.

Table 1: Comparison of model variants on the MSVD dataset. The experiments are implemented with the same global feature GoogleNet pool5 and different local features. "SS" means semantic supervision. v, g denote VGG-16, GoogLeNet respectively. "DA" means decoder attention.

Methods	BLEU-4	METEOR	CIDEr
DMRM+pool5(v)	45.2	31.6	66.5
DMRM+in5b(g)	47.7	32.4	68.4
DA	46.5	32.0	70.1
DMRM+pool5(v)+DA	48.5	32.5	70.6
DMRM+in5b(g)+DA	50.0	33.2	73.2
DMRM+in5b(g)+DA with SS	**51.1**	**33.6**	**74.8**

Table 2: Experiment results on the MSVD dataset compared to the state-of-the-art methods.

Methods	BLEU-4	METEOR	CIDEr
S2VT [29]	-	29.8	-
Temporal Attention [33]	41.9	29.6	51.7
p-RNN [35]	49.9	32.6	65.8
HRNE with attention [20]	43.8	33.1	-
Boundary-aware encoder [3]	42.5	32.4	63.5
mGRU+pre-train [36]	49.5	33.4	**75.5**
DMRM+in5b(g)+DA	50.0	33.2	73.2
DMRM+in5b(g)+DA with SS	**51.1**	**33.6**	74.8

We use GoogLeNet pool5/7x7_s1 as our global feature and compare the experimental results of different local features. We denote pool5 output from VGG-16 with pool5(v) and inception 5b feature from GoogLeNet with in5b(g). And we use "DMRM" to represent Dual Memory Recurrent Model combined with regions-of-interest selecting module in all result tables.

Table 1 shows that in5b(g) outperforms pool5(v) 0.8% in the METEOR score. It demonstrates that a suitable local feature improves behavior of our DMRM. Then we introduce temporal attention used in [33] to decoder stage of our framework. So the framework can adaptively focus on interested regions of each frame in encoder stage and select important segments in decoder stage. It is more integrated and seems similar to how people understand video content. We evaluate the performance of DMRM with temporal attention with different settings. The decoder attention (DA) represents the framework which only uses temporal attention in decoder stage and utilizes a plain LSTM encoder. The methods with "DA" represent applying temporal attention in decoder part of corresponding methods.

From Table 1, we find that our "DMRM+DA" outperforms the basic decoder attention (DA) method by 0.5% in the METEOR score with VGG-16 pool5 and 1.2% with GoogLeNet inception 5b. The results demonstrate that exploiting temporal structure of regions-of-interest really enhances the video representation in video captioning. In addition, we validate the performance of semantic supervision on regions-of-interest selection process and the result shows that this improves BLEU-4, CIDEr, and METEOR scores compared with results without semantic supervision.

Table 3: Experiment results on the M-VAD dataset.

Methods	METEOR
SA-GoogLeNet+3D-CNN [33]	5.7
S2VT: RGB (VGG) [29]	6.7
HRNE with attention [20]	6.8
Boundary-aware (C3D+ResNet) [3]	**7.3**
DMRM+in5b(g)+DA with SS	6.9

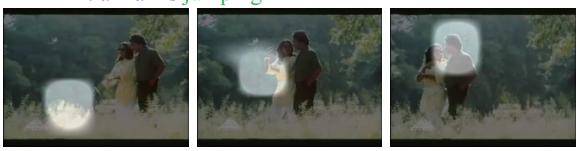
4.4.2 Comparison with the State-of-the-Art Methods. To further validate the effectiveness of our method, we compare our method with several the state-of-the-art methods on video captioning. S2VT [29] was a general encoder-decoder model which firstly introduced a stacked LSTM structure in both encoder and decoder stages. Temporal Attention [33] applied attention mechanism over features extracted from GoogLeNet and 3D-CNN. p-RNN [35] utilized a hierarchical decoder as sentence generator to create multiple descriptions for each video. Different with p-RNN, HRNE [20] introduced a hierarchical structure of encoder to incorporate long-dependencies for video temporal information. Boundary-aware encoder [3] detected boundaries of discontinuous video snippets and discovered the hierarchical structures of video inputs. mGRU [36] was a recent work which incorporated temporal information with different motion speeds.

The Table 2 lists the performance of our method and the state-of-the-art methods on the MSVD dataset. Since S2VT is regarded as foundation of most sequence-to-sequence based methods, our method outperforms it with improvement of 11.4% in METEOR score, which demonstrates that our model actually enhances video captioning system and catching interested region details assists in video understanding. Besides, our method outperforms "HRNE with attention" model in both BLEU-4 and METEOR. It demonstrates that not only temporal information of global feature but temporal regions-of-interest makes contribution to video captioning. The Table 2 also shows that our method outperforms other representative methods in both BLEU-4 and METEOR scores, which proves the effectiveness of our method again.

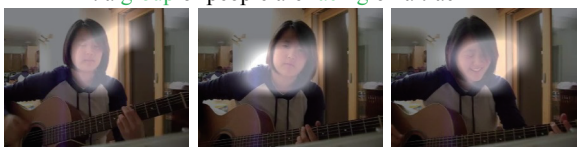
4.4.3 Video Captioning Examples. We show some representative examples in Fig. 3 and Fig. 4 to illustrate the performance of our caption results. Fig. 3 contains the sentences obtained from basic decoder attention (DA), our DMRM and ground truth. The bright circles are interested regions selected by an attention model. We find that the interests of different frames are located in different regions and the regions usually correspond to relevant semantic concepts of captions. The results demonstrate that catching temporal regions-of-interest makes contribution to distinguishing semantic concepts. For example, with more discriminative details from interested regions, our model improves "running" to a more explicit concept "racing on a track". Fig. 4 shows examples with sentence descriptions generated by our DMRM with semantic supervision (SS). It demonstrates that semantic supervision will lead our model to a more correct semantic concept. The representative examples give an intuitive sense for the effectiveness of our method and semantic supervision on attention results.

Ground truth: a man is jumping over things
DA: a man is riding a bicycle
DMRM: a man is jumping

Ground truth: a group of men are racing down a track
DA: a man is running on a track
DMRM: a group of people are racing on a track



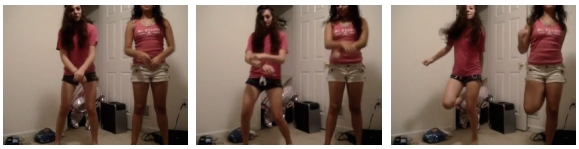
Ground truth: a man and a woman are walking in the woods
DA: a man and woman are riding a horse
DMRM: a man and a woman are walking



Ground truth: a girl is playing a guitar
DA: a man is playing a guitar
DMRM: a girl is playing a guitar

Figure 3: Captions generated by basic decoder attention (DA), our DMRM with decoder attention, and ground truth on MSVD test set. The bright circles indicate interested regions selected by our method for each frame.

Ground truth: elephants are playing with soil
DA: a man is walking
DMRM: a rhino is walking
DMRM SS: an elephant is walking



Ground truth: two girls are dancing
DA: a woman is doing exercise
DMRM: a woman is dancing
DMRM SS: two girls are dancing

Ground truth: a man is riding a horse in the desert
DA: a woman is riding a horse
DMRM: a woman is riding a horse
DMRM SS: a man is riding a horse

Ground truth: a man is lifting a car
DA: a man is riding a car
DMRM: a man is pushing a car
DMRM SS: a man is lifting a car

Figure 4: Examples on MSVD dataset. The sentences after "DMRM SS" are generated by our method with semantic supervision. The figure demonstrates that semantic supervision can lead to a correct concept.

4.5 Experimental Results on M-VAD

Table 3 lists the result of our method on the M-VAD dataset. Since video clips of M-VAD dataset are collected from realistic movies, they usually contain plentiful visual concepts and sentence descriptions with complicated syntactic structure. We compare our method with the state-of-the-art methods in METEOR score and omit other scores with the suggestion by [20]. The experiment result shows that our method outperforms the most state-of-the-art methods and gets comparable performance on complex movie captioning dataset. The performance of Boundary-aware [3] is better than ours on this dataset, since it is well designed to detect boundaries of discontinuous segments and more suitable for movie captioning datasets. Our method also outperforms it on the MSVD dataset.

5 CONCLUSIONS

In this paper, we propose a novel method for video captioning, which automatically selects regions-of-interest and incorporates their temporal dependencies. We evaluate our method on MSVD and M-VAD datasets. The experimental results demonstrate that our method achieves state-of-the-art performance on both evaluation datasets. Different from previous methods, we firstly encode regions-of-interest for each frame as sequential inputs and devise a specific attention model to select interested regions at each time step. In addition, we also find that employing semantic supervision on attention results improves the performance. We argue that our method is more human-like as important regions in the frames usually attract more attention for understanding video content.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[2] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. 2016. Delving deeper into convolutional networks for learning video representations. In *ICLR*.

[3] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2017. Hierarchical Boundary-Aware Neural Encoder for Video Captioning. In *CVPR*.

[4] 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.

[5] 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).

[6] Liu Chenxi, Mao Junhua, Sha Fei, and Yuille Alan. 2017. Attention Correctness in Neural Image Captioning. In *AAAI*.

[7] 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 *CVPR*. 2625–2634.

[8] 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 ACM International Conference on Multimedia*. 1082–1086.

[9] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*. 2712–2719.

[10] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. 2016. Attention-based LSTM with Semantic Consistency for Videos Captioning. In *Proceedings of the ACM International Conference on Multimedia*. 357–361.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Tushar Karayil, Philipp Blandfort, Damian Borth, and Andreas Dengel. 2016. Generating Affective Captions using Concept And Syntax Transition Networks. In *Proceedings of the ACM International Conference on Multimedia*. 1111–1115.

[13] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. In *AAAI*, Vol. 1. 2.

[14] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *CVPR*. Citeseer.

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

[16] Guang Li, Shubo Ma, and Yahong Han. 2015. Summarization-based video caption via deep neural networks. In *Proceedings of the ACM International Conference on Multimedia*. 1191–1194.

[17] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *CoNLL*. 220–228.

[18] Xiangyang Li, Xinhang Song, Luis Herranz, Yaohui Zhu, and Jiang Shuqiang. 2016. Image Captioning with both Object and Scene Information. In *Proceedings of the ACM International Conference on Multimedia*. 1107–1110.

[19] Yuan Liu and Zhongchao Shi. 2016. Boosting video description generation by explicitly translating from frame-level captions. In *Proceedings of the ACM International Conference on Multimedia*. 631–634.

[20] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*. 1029–1038.

[21] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *CVPR*. 4594–4602.

[22] 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.

[23] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *ICCV*. 433–440.

[24] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* 15, 1 (2014), 1929–1958.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.

[26] 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 *CVPR*. 1–9.

[27] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070* (2015).

[28] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*. 4566–4575.

[29] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *CVPR*. 4534–4542.

[30] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. *NAACL-HLT* (2015).

[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] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. 2016. Review networks for caption generation. In *NIPS*. 2361–2369.

[33] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*. 4507–4515.

[34] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*. 4651–4659.

[35] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*. 4584–4593.

[36] Linchao Zhu, Zhongwen Xu, and Yi Yang. 2017. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos. In *CVPR*.