Sports Video Captioning by Attentive Motion Representation based Hierarchical Recurrent Neural Networks

Mengshi Qi
Beijing Advanced Innovation Center for Big Data and Brain Computing
School of Computer Science and Engineering
Beihang University
Beijing, China

Yunhong Wang
Beijing Advanced Innovation Center for Big Data and Brain Computing
School of Computer Science and Engineering
Beihang University
Beijing, China

Annan Li*
Beijing Advanced Innovation Center for Big Data and Brain Computing
School of Computer Science and Engineering
Beihang University
Beijing, China

Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY, USA

ABSTRACT

Sports video captioning is a task of automatically generating a textual description for sports events (*e.g.* football, basketball or volleyball games). Although previous works have shown promising performance in producing the coarse and general description of a video, it is still quite challenging to caption a sports video with multiple fine-grained player's actions and complex group relationship among players. In this paper, we present a novel hierarchical recurrent neural network (RNN) based framework with an attention mechanism for sports video captioning. A motion representation module is proposed to extract individual pose attribute and group-level trajectory cluster information. Moreover, we introduce a new dataset called *Sports Video Captioning Dataset-Volleyball* for evaluation. We evaluate our proposed model over two public datasets and our new dataset, and the experimental results demonstrate that our method outperforms the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**; **Natural language generation**;

KEYWORDS

Sports Video Captioning; Video Analysis; Motion Representation; Pose Attribute; Volleyball

ACM Reference Format:
Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2018. Sports Video Captioning by Attentive Motion Representation based Hierarchical Recurrent Neural Networks. In *1st International Workshop on Multimedia Content Analysis in Sports (MMSports'18), October 26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3265845.3265851

*Corresponding author: liannan@buaa.edu.cn.

MMSports'18, October 26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5981-8/18/10…$15.00
https://doi.org/10.1145/3265845.3265851

1 INTRODUCTION

Sports video captioning, which describes events and actions happened in the match with language, has captured more attention in multimedia and natural language processing communities [8, 45]. In sports videos, a host of different categories of players' actions and interactions among players occur at the same time, *e.g.* in a volleyball game (see Figure 1). The complex variations of dynamic event and temporal structures make sports video captioning an arduous problem.

Recently, more researchers strive to this emerging topic. Conventional algorithms for video captioning can be divided into two categories: one is template-based language model [13, 34, 48], which generates captions based on predefined grammar rules, templates of sentences, and correlates each part of sentence with detected object; and the other is sequence learning method [7, 10, 15, 27, 28, 30, 41, 49, 51] that is inspired by Recurrent Neural Nework (RNN), such as Long Short-Term Memory (LSTM) [14] and Gated Recurrent Unit (GRU) [5]. Sequence learning methods achieve state-of-the-art performance at present for visual captioning. They are based on the encoder-decoder architecture: the encoder is used to translate input original video frame to the compact visual feature, while the decoder is used to generate words and sentences by sequence. However, all these methods can only generate coarse description, and present a video sequence simply by a collection of the frame-level feature, which ignore

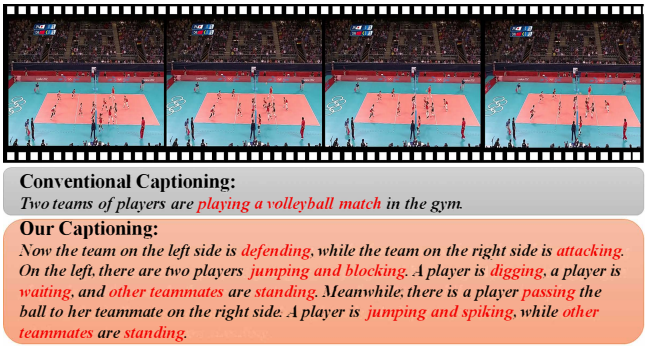

Figure 1: Illustration of sports video captioning task. Conventional captioning generates coarse-level text description for a video. In contrast, sports video captioning task needs to capture more fine-grained individual action details and group relationships.

the motion details of player's action and group activity, and are not inappropriate for sports video captioning.

Sports video captioning should take not only global visual appearance, but also the fine-grained individual motion information into consideration. Player's individual action is the main fine-grained motion information in sports event, which involves player's articulated movements/pose estimation [3, 4, 25] and motion trajectory [42, 52]. Capturing and representing these motion accurately and effectively from the untrimmed video would provide more informative cues for captioning. Meanwhile, attention mechanism [23, 47, 50] is often introduced to identify the salient visual regions with high objectness score and meaningful visual pattern of an image. For video captioning, the performance can be also improved by attending the spatial salient object/player and the temporal motion information. The key player's action or movement, such as dunking in basketball and shooting in soccer, invariably play a significant role in a sports event, thus precisely attending to these highlights and retrieving the crucial movement are overwhelmingly critical for sports video captioning.

In this paper, to address the above-mentioned issues, we propose a novel hierarchical LSTM-based deep framework for sports video captioning with attentive motion representation. In particular, individual pose attribute features and group-level trajectory cluster information will be fed into an encoder-decoder network. Then, we fuse the motion representation and global frame-level features by the attention mechanism and decode them into natural language utilizing a sequence to sequence architecture. Precisely, the main contribution of this work are summarized as:

- We introduce a novel framework for sports video captioning with attentive motion representation based hierarchical recurrent neural networks.
- A motion representation module is presented to extract player's pose and trajectory information, where we capture semantic attribute from player's skeletons, and cluster trajectory from team-level movement.
- We collect a new dataset called *Sports Video Captioning Dataset-Volleyball* that mainly contains volleyball games for evaluation. Meanwhile, extensive experiments on two public benchmarks and our dataset demonstrate the effectiveness and general applicability of our framework. To the best of our knowledge, we are the first to propose such a volleyball video captioning dataset.

2 RELATED WORK

Video Captioning Early efforts adopt template-based language methods [13, 34, 48] that align sentence elements with detected words from visual content. Rohrbach *et al.* [34] learned a Conditional Random Fields (CRF) [21] to model the relationships between different components of video content, and generated sentence descriptions for video. Xu *et al.* [48] proposed a unified framework to jointly model video and language by utilizing a compositional language model and a deep neural network.

Recently more sequence learning approaches [2, 7, 10, 27, 28, 30, 35] are used to learn probability distribution in space of video and textual sentence for video captioning. Venugopalan *et al.* [41] proposed an end-to-end sequence-to-sequence model to generate captions for videos, and their model can directly encode the temporal information by LSTM. Yao *et al.* [49] proposed a temporal attention mechanism to automatically select temporal segments for generating video caption. Yu *et al.* [51] proposed a hierarchical recurrent neural network, which consists of a sentence generator and a paragraph generator for video captioning. Furthermore, Hori *et al.* [15] incorporated audio features with image and motion feature for jointly video captioning via multi-model attention mechanism. Krishna *et al.* [19] introduced a dense video captioning model that combines the proposal and the captioning module to caption each event by single sentence. However, the captions generated by these works are coarse-level and missing lots of fine-grained level or detailed movement occurring in sports videos.

Sports Video Analysis Works on sports video analysis [16, 31, 32, 45] on group activity or team activity recognition are also relevant. Xu *et al.* [45] presented an approach for event detection from live sports game with text and video on the Internet. Zhu *et al.* [52] detected the goal event through extracting tactic information from broadcast soccer video. Duan *et al.* [8] proposed a mid-level representation between audio-visual processing and semantic analysis for sports video analysis. Ibrahim *et al.* [16] presented a hierarchical LSTM model for group activity recognition in the volleyball game.

However, these previous methods are inappropriate for sports video captioning task. In our work, we propose a novel framework to describe details of sports games with attentive motion representation.

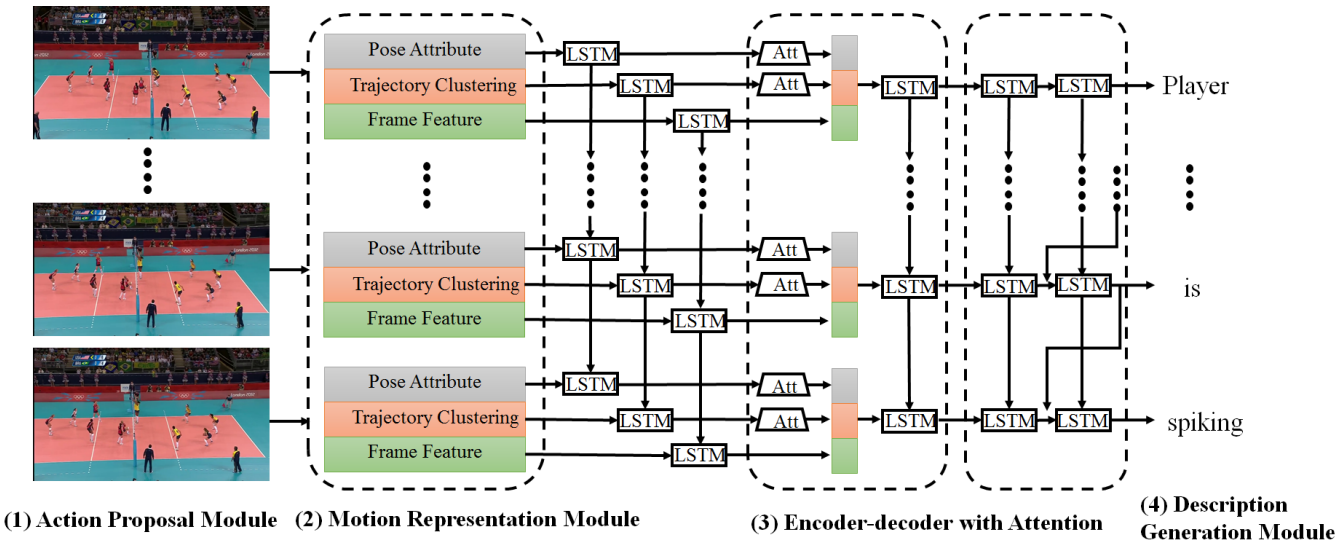

Figure 2: The overall framework of our sports video captioning model. (1) Action proposal module segments the whole video into activities of players. (2) Motion representation module employs detected pose attribute, trajectory clustering and frame-level feature to encode the individual action and group activity information. (3) Finally, all features are fused and decoded through an LSTM-based sequence-to-sequence structure with an attention mechanism. (4) Description generation module is used to generate textual caption.

3 THE PROPOSED APPROACH

The framework of our proposed approach for sports video captioning is illustrated in Figure 2. In particular, our framework includes: (1) action proposal module; (2) motion representation module; (3) encoder-decoder with attention mechanism and (4) description generation module. We adopt a sequence-to-sequence based architecture [41], where the input is the sequence of video frames, and the output is the sequence of words. And the lengths of the input and output are variable. Because the success of Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) in the visual captioning task, we employ this paradigm to our framework.

3.1 Action Proposal Module

Given a video, retrieving and localizing temporal segments that likely contain paramount spatio-temporal group events (*i.e.* attack, defend) or individual actions (*i.e.* spiking, passing) is the first task. In our work, we adopt Deep Action Proposals (DAP) [9] method for generating temporal action proposals. We infer the temporal location and duration of the action proposals from a T-frame video. And each proposal is associated with a confidence score. In practice, the input feature of video frames is extracted from the top layer of a 3D convolutional network (C3D) [38], and then an LSTM network is utilized to encode the sequential information.

3.2 Motion Representation Module

For describing sports video in natural language, more fine-grained details concerning player's actions would be of great assistance. Therefore, we design a motion representation module to model player's action, which consists of a pose attribute detection part and a trajectory clustering part.

3.2.1 Pose Attribute Detection. Pose estimation is often used to recognize the individual action in a fine-grained manner. Given a sequence of frames (both RGB and optical flow), we wish to determine the precise location of critical keypoints of the human body, for understanding individual posture and limb articulation as shown in Figure 3. In our work, we firstly utilize Faster RCNN [11] for localizing all players and extracting the bounding boxes. Then we have a set of candidate objects with the bounding box that represents their location and appearance feature. Based on the probability map, we select 12 bounding boxes with high confidence score per frame. After that, we adopt hourglass model [24] to extract the keypoints of each player. Center point of the detected skeleton is utilized to measure the relative offset of each body part, and optical flow values represent the motion of every joint, which manifest the characteristics of the player's movement (*e.g.* velocity and direction).

Then we extract pose-based CNN (Convolutional Neural Network) feature [4] from each body parts of a player in each frame. Based on the position of body joints (*i.e.* we select five pose parts per player: *right hand, left hand, upper body, bottom body* and *full body*), we crop corresponding RGB and optical flow patches and normalize them to 224×224. We adopt VGG-16 network pretrained on the ImageNet dataset [20] for RGB patches, and motion network in [12] pretrained on the UCF-101 dataset [37]. The pose-based

Figure 3: Pose attribute detection part in our model. The inputs are both RGB and optical flow images, and the outputs are binary representation of semantic attributes.

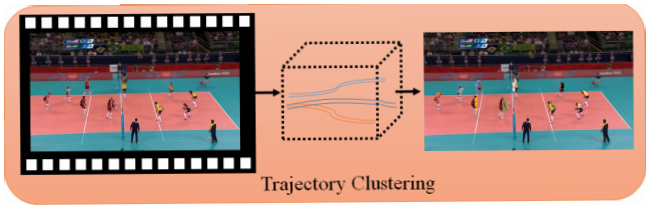

Figure 4: Trajectory clustering part in our model.

feature for each player in a frame can be denoted as F_{pose}, which concatenate all the features of each pose part.

However, directly and disorderly pooling all the players' pose-based feature is coarse, we desire to further capture more semantic attributes from raw features. We build an attribute vocabulary from the annotated sentences in dataset (*e.g.* UCF-101 [37], Volleyball [16]), and we use the top k high-frequency words of them. The pose attribute can be object (*e.g.* hand, leg, head, feet) or motion (*e.g.* spike, dig, pass). Taking n player's pose-based features $\{F^1_{pose}, ..., F^n_{pose}\}$ as input, we adopt the last fully connected layer of VGG-16 net to be a k-way classifier. We define $y^i = [y^i_1, ..., y^i_k]$ as the pose attribute vector of the i-th player, where $y^i_k = 1$ if the player is annotated with attribute k, and $y^i_k = 0$ otherwise. We define the attribute predict probability vector as $p^i = [p^i_1, ..., p^i_k]$, the loss function is as following:

$$L_{att} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[y^i_j \log(p^i_j) + (1 - y^i_j)\log(1 - p^i_j)], \quad (1)$$

After training, we formulate frame-level pose attribute vector by fusing y^i for all the players. Furthermore, we adopt the attention mechanism in the encoder network (will be described in 3.3) to get attentive pose attribute representation F_{pose_att} of n player's poss attribute vector:

$$F_{pose\_att} = \sum_{i=1}^{n}\{Attention\ Weight\} \cdot y^i, \quad (2)$$

3.2.2 Trajectory Clustering. The trajectory is good at representing motion in videos, and clustering them into groups can capture information of group or team activity. We adopt the method in [33, 44] to extract the dense point trajectories $Tra = \{Tra_1, Tra_2, ...Tra_M\}$ for a sequence of frames, where M is the number of trajectories. And we set the maximum length of the trajectory to 15 frames. Furthermore, we follow the distance metric in [33] to measure the similarity between trajectory pairs, considering temporal interval and spatial position. Then we partition all the detected trajectories into groups by computing the affinity matrix between each trajectory pair and utilizing a graph clustering method [33] as depicted in Figure 4. Given a video, we can obtain m clusters.

We assume the i-th trajectory cluster which contains L trajectories as $Tra(i) = \{Tra_{i1}, ..., Tra_{il}\}$. And defining each trajectory as a position point sequence $Tra_{il} = \{(x^1_{il}, y^1_{il}, z^1_{il}), ..., (x^T_{il}, y^T_{il}, z^T_{il})\}$, where $(x^t_{il}, y^t_{il}, z^t_{il})$ is the 3D coordinates of the t-th point in trajectory Tra_{il}, and T is the time step of trajectory.

Following [43], we employ convolutional neural networks to obtain the so-called trajectory-pooled deep-convolutional representation. We input each frame to the CNN (VGG-16 in our work), and obtain a feature map of size $H \times W \times N$, where H, W and N are the number of height, width and channel, respectively. Finally, we achieve an overall feature map $C \in R^{H \times W \times T \times N}$ through concatenating all the feature maps of the video, where T is the length of the video. Then, a trajectory point with coordinates (x^t, y^t, z^t) (center at $(r \times x^t, r \times y^t, r \times z^t)$) in the feature map, where r denotes the map size ration with respective to the input size. Thus the averaged feature of Tra_{it} is formulated as the following:

$$F_{tra_{il}} = \frac{1}{T}\sum_{t=1}^{T} C(r \times x^t_{il}, r \times y^t_{il}, r \times z^t_{il}), \quad (3)$$

and the representation of the trajectory cluster is computed via mean pooling of all trajectory features in the same cluster:

$$F_{tra_i} = \frac{1}{L}\sum_{l=1}^{L} F_{tra_{il}}, \quad (4)$$

For a given video, we extract m trajectory clusters and the visual feature of them is defined as $F_{tra_1}, F_{tra_2}, ..., F_{tra_m}$. Then we adopt attention mechanism in the encoder network (will be described in 3.3) to formulate the overall trajectory feature vector F_{tra}:

$$F_{tra} = \sum_{i=1}^{m}\{Attention\ Weight\} \cdot F_{tra_i}, \quad (5)$$

3.3 Encoder-Decoder with Attention Mechanism

We follow the popular encoder-decoder framework for video captioning. The encoder is a one layers bi-directional LSTM which encodes the input video features into a sequence of feature vectors. In our work, the LSTM based encoder takes both attentive motion representation (*i.e.* pose attribute feature F_{pose_att} and trajectory clustering feature F_{tra}) and frame feature F_{frame} as input, which would be concatenated to formulate total representation F_t in the t-th frame. The

updating procedure in LSTM is formulated as

$$h_t = LSTM(h_{t-1}, F_t), \\ F_t = [F_{pose\_att}, F_{tra}, F_{frame}] \tag{6}$$

where h denotes the hidden state of LSTM, and $[\cdot]$ denotes concatenate operation.

The decoder takes the encoder representation as input, then sequentially produce the output vector, where denotes the predicted word at each time step. At each time step t, the LSTM updates its hidden state h_t and output y_t based on its previous hidden state h_{t-1} and output y_{t-1} and the encoder embeddding V, as the following:

$$[^{y_t}_{h_t}] = Decoder(h_{t-1}, y_{t-1}, V), \tag{7}$$

Next, we will introduce the attention mechanism in the encoder network.

Attention Mechanism: Conventional methods (*e.g.* mean pooling operation) always ignore the importance of motion information for video captioning, where the key player often plays the most remarkable role in group event. We adopt a soft attention model to obtain dynamic weighted sum of the pose attribute feature and trajectory cluster representation (described in Sec 3.2.1 and 3.2.2). Given the motion representation F, we denote $F_i \in \{F_1, ..., F_n\}$ for the feature of the i-th players or the i-th trajectory clusters, where n is number of players or number of trajectory clusters. We feed them to a single linear transform layer followed by a softmax function to calculate the attention distribution over motion representation F_i (*i.e.* pose attribute feature F_{pose_att} and trajectory cluster feature F_{tra}), and define $s_i^t = (s_1^t, ..., s_n^t)^T$ as the importance score in i-th player or i-th trajectory cluster on T frames:

$$s_i^t = U_s \tanh(W_{fs} F_i + W_{hs} h_{t-1}^s + b_s), \tag{8}$$

where U_s, W_{fs}, W_{hs} are the training parameters, and b_s is the bias vector. h_{t-1}^s is the hidden variable from an LSTM unit. Then the attention weight is computed as a normalization of the scores:

$$\alpha_i^t = \frac{\exp(s_i^t)}{\sum_{i=1}^n \exp(s_i^t)}. \tag{9}$$

After that, the visual feature input to the encoder at time t is computed by the weighted sum of the frame features, *i.e.* $F_t^{'}$,

$$F_t^{'} = \sum_{i=1}^{n} \alpha_i^t F_i, \tag{10}$$

where n denotes the number of players or trajectory clusters, α_i^t is the attention weight of i-th player or i-th trajectory cluster at time t. With the attention mechanism, the encoder is able to attend on the salient trajectory movement and key player's pose information.

3.4 Description Generation Module

The aim of video captioning in our work is to generate a paragraph includes several word sequences. For generating the sentence, the likelihood of generating a word in the n-th sentence is formulated as the following:

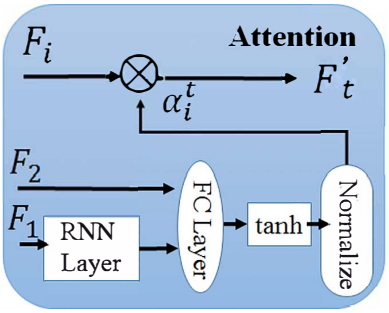

Figure 5: Attention Mechanism in our model.

$$P(w_t^n | s_{1:n-1}, w_{t-1}^n, F_t, W), \tag{11}$$

where $s_{1:n-1}$ represents all the preceding sentences in the paragraph, w_{t-1}^n means all the previous words in the n-th sentence, F_t are the feature that concatenate attentive motion representation and global features in the corresponding frames of the video, and W represents the model parameters. Furthermore, we define the overall loss function of generating the whole paragraph $s_{1:N}$ as:

$$L_{cap} = -\sum_{n=1}^{N}\sum_{t=1}^{T_n} \log P(w_t^n | s_{1:n-1}, w_{1:t-1}^n, F_t, W)) / \sum_{n=1}^{N} T_n \tag{12}$$

where N is the number of sentences in the paragraph, T_n is the number of words in the n-th sentence.

4 SPORTS VIDEO CAPTIONING DATASET

Sports Video Captioning Dataset-Volleyball (SVCDV) is a new dataset introduced by us that is focus on sports captioning. SVCDV has totally 55 videos with 4,830 short clips collected from *Youtube*, which are mainly high-resolution broadcast Olympic volleyball games. Specifically, the short clips are segmented into different types of group activities, and each short clips has more than 50 frames. It is annotated based on the Volleyball Dataset [16] that is introduced to address group activity recognition issue especially. We annotated natural language description of player action and group activity happened in each video, and each sentence with respect to one action or movement. Furthermore, SVCDV has totally 44,436 sentences, of which each video clip has 9.2 sentences on average. Meanwhile, the average sentences per second is 0.366, and verbs per sentence is 1.72, and verb ratio is 16.2% in the SVCDV dataset, which are all more than that in current general video captioning datasets (*e.g.* MSVD [13, 27], MSR-VTT [46], ActivityNet [19]). It demonstrates that SVCDV dataset is extremely suitable for sports captioning task. Besides, each player is labeled with a bounding box and one of the nine action labels: *waiting, setting, digging, falling, spiking, blocking, jumping, moving* and *standing*. The whole frame is annotated with one of the

eight group activity labels: *right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike* and *left set*. These labels can be utilized for individual pose attribute learning and group relationship modeling. In experiments, we split it into training, validation and testing sets of 65%, 5%, 30%, corresponding to 3,140, 241 and 1,449 video clips respectively.

5 EXPERIMENTS

In this section, we conduct extensive experiments to demonstrate the effectiveness of our approach on two public benchmark datasets: **MSVD Dataset**, **MSR-VTT Dataset**, and **Sports Video Captioning Dataset-Volleyball**. In the following, we first briefly introduce the datasets, evaluation metrics and implementation details. Then we describe the comparison methods, and present the experimental results and analysis.

5.1 Datasets

Microsoft Video Description Dataset (MSVD) [13, 27] contains 1,970 short videos collected from *YouTube*, where each video describes a single activity in a wide range of topics (*e.g.* animals, music, actions and sports). In total, the dataset consists of 80,839 sentences with regarding 40 English descriptions per video clip, and each sentence has about 8 words. Following the same setting in [13], we select 1,200 videos as the training set, 100 for validation and 670 as the testing set.

MSR Video-to-Text Dataset (MSR-VTT) [46] is the largest general video captioning dataset in the size of sentences and vocabulary. It contains 10,000 video clips with 41.2 hours and 200,000 clip-sentence pairs in 20 categories (e.g., news, sports), and 20 natural sentences annotated manually for each video clip. Following the same setting in [46], we split it into training, validation and testing sets of 65%, 5%, 30%, corresponding to 6,513, 497 and 2,990 clips respectively.

Metrics We choose four popular metrics for the evaluation: CIDEr (C) [40], BLEU (B) [29], METEOR (M) [6], and ROUGE-L(R) [22]. We adopt Microsoft COCO evaluation tools[1] to test the performance of video captioning, which has implemented the metrics and evaluation functions.

5.2 Implementation Details

For video preprocessing, we sample equally-spaced 25 frames for each video, and resize them to 224×224 resolution. A VGG-16 network [36] pretrained on the ImageNet dataset [20] is used for extracting visual appearance features. We select a sequence of 4096-dimensional feature vectors produced by the fully connected layer fc7. Besides, we employ pretrained C3D [39] on the Sports-1M dataset [17] to model motion and short-term spatio-temporal activity of videos. We extract activation vector from fully-connected layer fc6-1 of C3D network from frames of input video.

For text preprocessing, we convert all words to lowercases and split sentences into words and remove punctuation using *wordpunct-tokenizer* method from NLTK toolbox[2]. Consequently, we achieve the vocabulary with 12,593 words from MSVD, 13,065 words from MSR-VTT, and 7,296 words from SVCDV, where the word with the frequency less than 3 is removed. Furthermore, we utilize the one-hot vector to represent each word in our work.

For training our model, we add tag *BOS* and *EOS* to denote the begin and end of each sentence, respectively, which is aimed at making the length of sentences arbitrary. Then we input the *BOS* into video decoder to start generating video descriptions. The learning rates for training stage and the training batch size are set to 1×10^{-4} and 64 for MSVD, MSR-VTT and SVCDV, respectively. Meanwhile, we adopt Dropout for regularization with probability 0.5 on the input and output of encoder LSTMs and decoder LSTMs. For LSTMs in our model, the size of hidden states are set to 1,024, and size of embedding representation of video feature and words are set to 512. We select Adam optimizer [18] to update all the parameters in our model. We stop training our model until 200 epochs until the evaluation metric does not improve on the validation set. In the testing, we adopt the beam search strategy with the beam size 5. Our model is implemented using the TensorFlow [1] library with a single NVIDIA GTX 1080Ti GPU.

5.3 Compared Methods

In order to demonstrate effectiveness of our approach, we compare our model with following state-of-the-art methods: S2VT [41], LSTM-E [48], TA [49], HRNN [51], HRNE [26], DenseCap[19]. Following the experiments set in [19], we compare existing video captioning models using ground truth proposals.[3]. Since not all the papers report all the information, we only report results on the test set in all the dataset in our experiments.

5.4 Result and Analysis

Results On General Datasets: To evaluate the generality of our model, we conduct experiments on the MSVD and MSR-VTT datasets that are general video captioning datasets covering wider topics. The results and comparisons can be found in Table 1. As can be seen, the proposed method is able to achieve competitive results. On the MSVD dataset, the performance of our method is no more 2% worse than the state-of-the-art methods across all metrics. Meanwhile, the performance of our model can get the second place on MSR-VTT across most of metrics. The results imply that better fine-grained motion representation can also effectively enhance the performance of general video captioning. Particularly, it is worth noting that our model can be easily integrated with the compared methods for general video captioning. From the Table 1, we find that S2YT performs much worse than other models in the MSVD dataset since it encodes long sequences of video by mean pooling. H-RNN

[1] https://github.com/tylin/coco-caption

[2] http://www.nltk.org

[3] In the experiments, the parameter settings of above-mentioned methods are adopted from corresponding papers.

Table 1: Performance comparisons of our method and the state-of-the-art approaches with different video features on the MSVD/MSR-VTT Dataset. (V) denotes VGGnet, (O) denotes optical flow, (G) denotes GoogleNet, (C) denotes C3D and (R) denotes ResNet-152. All results are cited from corresponding papers. The best performance is highlighted in bold.

Methods	MSVD						MSR-VTT		
	B@1	B@2	B@3	B@4	M	C	B@4	M	C
basic LSTM(R)	80.6	69.3	59.7	49.6	32.7	69.9	-	-	-
S2VT [41](V)	-	-	-	-	29.2	-	-	-	-
S2VT [41](V+O)	-	-	-	-	29.8	-	31.4	25.7	**35.2**
S2VT [41](C)	73.5	59.3	48.2	36.9	29.8	48.6	31.4	25.7	**35.2**
LSTM-E [48](V)	74.9	60.9	50.6	40.2	29.5	-	-	-	-
LSTM-E [48](C)	75.7	62.3	52.0	41.7	29.9	-	-	-	-
LSTM-E [48](V+C)	78.8	66.0	55.4	45.3	31.0	-	-	-	-
TA [49](R)	**81.6**	70.3	**61.6**	**51.3**	**33.3**	**72.0**	-	-	-
TA [49](V)	-	-	-	-	-	-	35.6	25.4	-
TA [49](C)	74.1	58.9	48.2	36.6	29.4	48.1	36.1	25.7	-
TA [49](G+C)	80.0	64.7	52.6	42.2	29.6	51.7	-	-	-
TA [49](V+C)	-	-	-	-	-	-	**36.6**	**25.9**	-
HRNN [51](V)	77.3	64.5	54.6	44.3	31.1	62.1	-	-	-
HRNN [51](C)	79.7	67.9	57.9	47.4	30.3	53.6	-	-	-
HRNN [51](V+C)	81.5	**70.4**	60.4	49.9	32.6	65.8	-	-	20.2
HRNE [26](G)	78.4	66.1	55.1	43.6	32.1	-	-	-	-
HRNE+TA [26](G)	79.2	66.3	55.1	43.8	33.1	-	-	-	-
DenseCap[19](C)	-	-	-	-	-	-	-	-	-
Ours model (V+C+O)	80.7	69.5	60.7	50.5	32.7	69.6	36.2	25.6	33.5

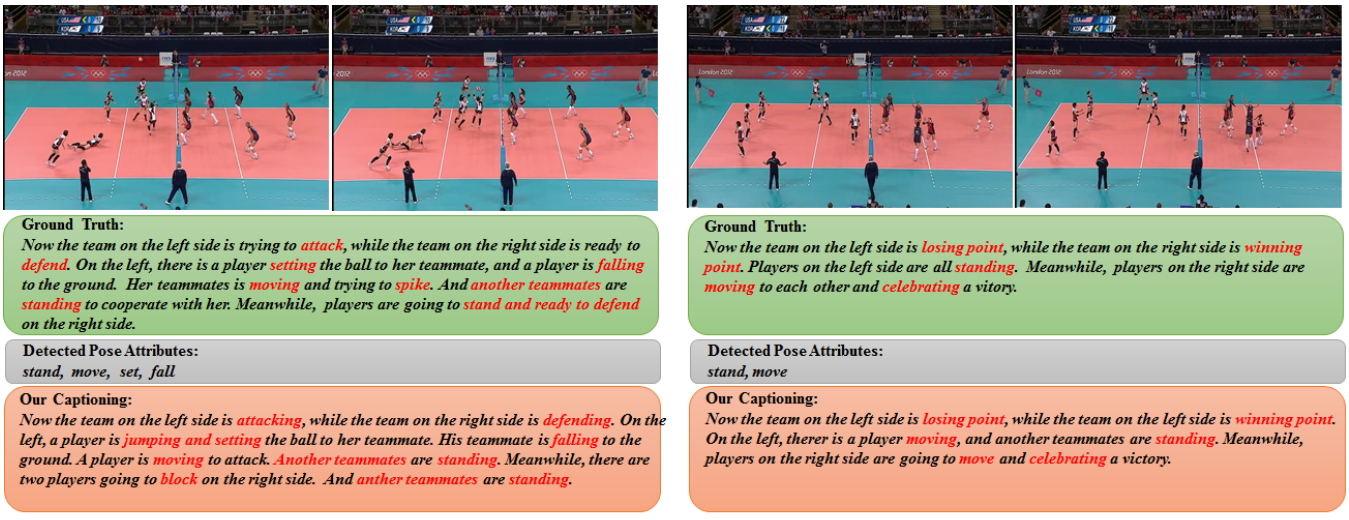

Figure 6: Qualitative video captioning results on the SVCDV dataset. The highlights in red denote important actions and activities in a sports event.

performs slightly better due to its attentive object-level features. In addition, we have seen how utilizing a more powerful feature representation can improve the performance, where methods with ResNet features perform significantly better than C3D features (*e.g.* TA with ResNet feature obtains the best performance than that with other features in the MSVD dataset). Specifically, it should be pointed out that the proposed method focuses on sports video captioning and also has the ability for general video captioning.

Results On SVCDV: We evaluate our method on the new SVCDV dataset that mainly contains sports videos. Table 2 reports the results and comparisons with the state-of-the-art and baseline methods. As can be noticed in Table 2, our approach improves plain techniques and achieves the

Table 2: Performance comparisons of our method and the state-of-the-art approaches on the SVCDV Dataset and the components analysis of our framework.

Methods	B@4	R	M	C
S2VT [41]	25.62	45.26	21.55	1.96
HRNN [51]	24.53	44.97	20.96	2.05
DenseCap[19]	26.77	46.78	23.33	2.29
Ours w/o motion	24.55	44.22	20.36	1.68
Ours w/o pose	25.12	44.57	21.21	1.99
Ours w/o trajectory	26.65	45.05	21.69	2.05
Ours w/o attention	27.18	46.38	23.19	2.06
Ours full model	**28.39**	**47.75**	**24.23**	**2.52**

state-of-the-art performance on SVCDV. We choose *S2VT* model as a baseline with only global video feature without attentive motion representation. The baseline achieves the worst performance that deteriorates our proposed framework by about 3% across all metrics. It obviously manifests that the introducing attentive motion representation is rewarding for improving the performance of sports video captioning. Although HRNN and DenseCap have the ability to extract context information from the video, more accurate articulate action information is neglected. It suggests that our framework is capable to generate sentences containing more fine-grained motion representation and group relationship. Figure 6 illustrates quite a few qualitative captioning results on the test data of SVCDV datasets. As can be seen, our framework can abstract more fine-grained action and activity details in the generated text description, and in more accordance with the ground truth. However, our generated caption fails to describe accurate players' actions or activities in several cases (*e.g.* 'blocking' is mistaken as 'standing'), due to that some actions in the video share high similarities and occlusions in the video. More training data and accurate action detection model can be beneficial to better distinguish these actions.

Components Analysis: To give evidence of the effectiveness of each component in our model (*i.e.* motion representation module and attention mechanism), we have conducted further experiments for comparison. As can be seen in Table 2, we analyze and identify the effect of each module in our framework. Firstly, we evaluate how much *motion representation module* can help sports video captioning. In our work, we utilize both pose attribute feature and trajectory clustering feature as our motion representation. As a comparison, we only extract the whole video features through C3D model and LSTM scheme (*i.e.* Ours w/o motion in Table 2), the performance is the worst. It indicates that the motion representation module is the most paramount component in our model, and the performance would drop drastically if missing this module (*i.e.* performances degrade more than 4% across all the metrics compared with the full model). Because the raw video feature cannot extract more individual motion details from each frame. Comparing pose attribute feature with trajectory clustering, our method without pose attribute feature achieves worse performance than that without trajectory clustering. Obviously, it proves the pose attribute feature is more crucial than trajectory cluster, because more individual player motion description exists in dataset and attributes can capture more semantic information. Especially, we can see improvements by introducing *attention mechanism* into our encoder, suggesting that attention is exceedingly useful for sports video captioning. Since the key players invariably play an considerable role for the sports event. Meanwhile, it also reveals that fusing features of video by mean-pooling is not the best choice.

6 CONCLUSION

In this paper, we proposed a novel deep framework for sports video captioning based on extracting attentive motion representation. Through capturing human pose attribute feature and group trajectory clustering, our model was capable of describing more fine-grained information regarding players and team-level movement in a sports game. We evaluated our model on two public datasets and a new introduced *Sports Video Captioning Dataset-Volleyball*. The experimental results demonstrated the effectiveness of our framework that achieves competitive or superior performance compared with the current state-of-the-art models.

7 ACKNOWLEDGEMENTS

This work was partly supported by the National Natural Science Foundation of China (No. 61573045). Jiebo Luo would like to thank the support of New York State through the Goergen Institute for Data Science and NSF Award (No. 1722847). Mengshi Qi acknowledges the financial support from the China Scholarship Council.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning.. In *OSDI*, Vol. 16. 265–283.
[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2017. Hierarchical boundary-aware neural encoder for video captioning. In *CVPR*. IEEE.
[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*. IEEE.
[4] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. 2015. P-cnn: Pose-based cnn features for action recognition. In *ICCV*. IEEE.
[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint* (2014).
[6] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*. 376–380.
[7] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. 2017. Improving interpretability of deep neural networks with semantic information. *arXiv preprint* (2017).
[8] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian, and Chang-Sheng Xu. 2003. A mid-level representation framework for semantic sports video analysis. In *MM*. ACM.
[9] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *ECCV*. Springer.

[10] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *CVPR*. IEEE.

[11] Ross Girshick. 2015. Fast R-CNN. In *ICCV*. IEEE.

[12] Georgia Gkioxari and Jitendra Malik. 2015. Finding action tubes. In *CVPR*. IEEE.

[13] 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*. IEEE.

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

[15] 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* (2017).

[16] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A hierarchical deep temporal model for group activity recognition. In *CVPR*. IEEE.

[17] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*. IEEE.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint* (2014).

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*. IEEE.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

[21] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[22] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 605.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*. IEEE.

[24] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*. Springer.

[26] 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*. IEEE.

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

[28] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *CVPR*. IEEE.

[29] 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.

[30] Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. *arXiv preprint* (2017).

[31] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. 2018. stagNet: An Attentive Semantic RNN for Group Activity Recognition. In *ECCV*. Springer.

[32] Mengshi Qi, Yunhong Wang, and Annan Li. 2017. Online Cross-Modal Scene Retrieval by Binary Representation and Semantic Graph. In *MM*. ACM.

[33] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto. 2012. Discovering discriminative action parts from mid-level video representations. In *CVPR*. IEEE.

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

[35] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly supervised dense video captioning. In *CVPR*. IEEE.

[36] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014).

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint* (2012).

[38] Kevin Tang, Bangpeng Yao, Li Fei-Fei, and Daphne Koller. 2013. Combining the right features for complex event recognition. In *ICCV*. IEEE.

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*. IEEE.

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

[41] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *ICCV*. IEEE.

[42] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *CVPR*. IEEE.

[43] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*. IEEE.

[44] Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. 2018. Interpretable Video Captioning via Trajectory Structured Localization. In *CVPR*. IEEE.

[45] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan. 2006. Live sports event detection based on broadcast video and web-casting text. In *MM*. ACM.

[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*. IEEE.

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[48] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework.. In *AAAI*.

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

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

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

[52] Guangyu Zhu, Qingming Huang, Changsheng Xu, Yong Rui, Shuqiang Jiang, Wen Gao, and Hongxun Yao. 2007. Trajectory based event tactics analysis in broadcast sports video. In *MM*. ACM.