Visual Oriented Encoder: Integrating Multimodal and Multi-Scale Contexts for Video Captioning

Bang Yang* and Yuexian Zou†
*ADSPLAB, School of ECE, Peking University, Shenzhen, China
†Peng Cheng Laboratory, Shenzhen, China
Email: {yb.ece,zouyx}@pku.edu.cn

Abstract—Video captioning is a challenging task which aims at automatically generating a natural language description of a given video. Recent researches have shown that exploiting the intrinsic multi-modalities of videos significantly promotes captioning performance. However, how to integrate multi-modalities to generate effective semantic representations for video captioning is still an open issue. Some researchers proposed to learn multimodal features in parallel during the encoding stage. The downside of these methods lies in the neglect of the interaction among multi-modalities and their rich contextual information. In this study, inspired by the fact that visual contents are generally more important for comprehending videos, we propose a novel Visual Oriented Encoder (VOE) to integrate multimodal features in an interactive manner. Specifically, VOE is designed as a hierarchical structure, where bottom layers are utilized to extract multi-scale contexts from auxiliary modalities while the top layer is exploited to generate joint representations by considering both visual and contextual information. Following the encoder-decoder framework, we systematically develop a VOE-LSTM model and evaluate it on two mainstream benchmarks: MSVD and MSR-VTT. Experimental results show that the proposed VOE surpasses conventional encoders and our VOE-LSTM model achieves competitive results compared with state-of-the-art approaches.

I. INTRODUCTION

Video captioning, aiming to automatically describe video contents with complete and natural sentences, has drawn considerable attention from computer vision researchers and nature language processing communities. It has a wide range of practical applications like human-robot interaction, video retrieval and auxiliary aid for visually impaired people, etc. With the rapid development of deep learning, the encoder-decoder architecture is widely adopted for video captioning, where deep Convolutional Neural Networks (CNNs) are used to encode a video whereas Recurrent Neural Networks (RNNs) like LSTM [1] and GRU [2] are usually applied to generate sentences word by word.

Compared with image captioning, video captioning is much more challenging in two aspects. Firstly, videos are complex due to spatio-temporal dynamics. To address this complexity problem, temporal attention [3] was explored to exploit the temporal structure of videos, while spatial attention [4] was to capture the region-of-interest within static frames. Secondly, videos naturally contain multi-modalities like image (appearance of static frames), motion (spatio-temporal dynamics) and audio (acoustic clues), which inherently contain multi-scale contextual information. Some efforts have been made to integrate multi-modalities to boost captioning quality [5], [6].

For video captioning, videos can be grouped into two categories: visual dominant and visual subordinate. To illustrate the contribution of each modality, one example for each video category is shown in Fig. 1, where (a) is a visual dominant video and (b) is a visual subordinate video. In Fig. 1 (a), the contents can be understood with image modality easily since “jockeys”, “horses” and “race track” can be directly identified from these four frames. From the ground truth sentences, we can infer that motion and audio modalities are beneficial for generating vivid sentences. However, the visual subordinate video in Fig. 1 (b) cannot be understood with only image modality. With four images in (b), only “landscapes”, “trees” and “beautiful scenery” can be captured. From the ground truth sentences given besides the images, we understand that the motion and audio modalities in this video provide essential reference for captioning “woman” and “sing a song”. From these examples, we can conclude that the complementary nature of non-visual information should be exploited to assist visual perception for better video understanding.

Inspired by the discussions above, we differentiate the importance of multi-modalities within videos and define two
key components for comprehending video, that is, the visual contents of image modality and multi-scale contexts provided by auxiliary modalities, as illustrated in Fig. 2. Unlike previous works that learn multimodal features separately [5], [7], [8], [9], we propose a novel hierarchical encoder, Visual Oriented Encoder (VOE), to progressively integrate multimodal features into compact video representations. In our design, the bottom layers of VOE are utilized to extract multi-scale contexts from auxiliary modalities while the top layer isexploited to generate joint representations by considering both visual and contextual information. Finally, following the successful encoder-decoder framework [10], we develop a VOE-LSTM captioning model by simply employing a single layer LSTM as the decoder. To highlight, our main contributions are listed below:

- We propose a novel Visual Oriented Encoder (VOE), where multimodal and multi-scale contexts are progressively integrated in an interactive manner, to learn better joint video representations.
- Extensive results on two mainstream benchmark datasets, MSVD and MSR-VTT, verify the effectiveness of our proposed VOE-LSTM, which achieves competitive performance compared with state-of-the-art approaches.

II. RELATED WORKS

A. Exploiting Temporal Structure of Videos

The work proposed by Venugopalan et al. [10] is one of the first works that utilize the CNN plus LSTM architecture, but it simply applies mean pooling along the temporal axis, discarding the temporal structure of videos. To address this problem, Venugopalan et al. [11] propose a sequence to sequence model, a CNN followed with a stacked LSTMs, to process both input (sequence of frames) and output (sequence of words) of variable length. Besides, some efforts have been made to the decoder. Yao et al. [3] propose temporal attention to focus on only a subset of frames by increasing the attention weights of the corresponding temporal feature. Song et al. [12] further propose an adjusted temporal attention to decide when and where to use the visual information of videos to predict visual words. Except of the explorations in the decoder, researchers also investigate elaborate encoders. By imitating the convolution operation in CNN, Pan et al. [13] design a hierarchical LSTMs encoder to capture long-range dependencies of visual features at different time scales and granularities. Baraldi et al. further propose a boundary-aware encoder [14] to automatically discover and leverage the hierarchical structure of the video.

B. Integrating Multi-Modalities

Videos naturally contain multi-modalities like image, motion and audio. To utilize the complementary nature of these modalities, the prevalent approach usually generates joint representations by concatenating the multimodal features [9], [5]. In [15], authors term such concatenation method as static fusion and argue that it may result in inaccurate captioning since the different modalities contribute differently in generating descriptive words for different videos. To address their concerns, they provide a fusion strategy to dynamically determine which fusion method to use to describe different visual entities. With the same consideration, a two-stage attention mechanism named multimodal attention [7], [8] is introduced to dynamically decide the importance of different modalities when generating words. Most recently, the authors in [6] provide a new solution to determine the contribution of each modality where the models related to the different modalities are learnt individually and combined in a late fusion way. Except for modalities contained in video, external textual modality is also explored to improve the captioning performance [16], [17]. Specifically, the authors in [17] utilizes the frame-based image captions to acquire external knowledge. However, most of these works focus on differentiating the importance of multi-modalities in the decoding stage. But they neglect multimodal interaction in the encoding stage, which may results in semantics mismatch and misalignment [18].

C. Contextual Information

It is undeniably that rich spatial and temporal contextual information exists in images and videos, respectively. In recent years, a great number of approaches have been developed to explore the contextual information in images and videos for different learning tasks. For example, for segmentation and object detection tasks, [19] exploits both local context around each candidate object as well as the global scene-level context. For video captioning tasks, bidirectional LSTM is widely adopted to capture temporal contextual information including both past and future contexts [20], [21]. Literature studies show that the temporal contextual information is usually extracted for single modality, but how to effectively integrate intrinsic contextual information of multi-modalities in videos has not been well investigated.

III. APPROACH

The overall architecture of our VOE-LSTM model is shown in Fig. 3. Our VOE-LSTM follows the successful encoder-decoder framework for video captioning, which consists of the encoding and decoding stages. In the following, we will
first introduce an overview of our methodology, then present the details of our Visual Oriented Encoder (VOE), and finally describe the decoding stage.

A. Overview

As shown in Fig. 3, the encoding stage is composed of the CNN-based encoder and our proposed Visual Oriented Encoder (VOE). In this study, we consider three modalities including audio (rectangles in purple), motion (rectangles in blue) and image (rectangles in green). For a given video, $N$ key frames are uniformly sampled for image feature extraction. To ensure the semantic consistency, we treat each key frame as the center to generate corresponding motion and audio features so that each snippet of them contain both the past and future contexts for image features. With the pre-trained CNN models, multi-modalities are encoded into a sequential representations, namely $V = \{v_1, v_2, \ldots, v_n, v_N\}$, where $v_n = \{v_{n1}, v_{n2}, \ldots, v_{nK}\}$ and $K$ is the number of modalities. In particular, we denote the representations of $k$-th modality as $V^{(k)} = \{v_{1k}, v_{2k}, \ldots, v_{Nk}\}$, where $k \in [1, K]$. As illustrated in Fig. 3 (a), audio, motion and image features usually have different time-span. So in our design, we empirically process these three modalities in order as shown in Fig. 3 (b), and this is what we call “visual oriented processing order”. For clarity, $V^{(1)}$, $V^{(2)}$ and $V^{(3)}$ actually stand for audio modality, motion modality and image modality in this work (i.e., $K = 3$). Then $V$ is fed into our proposed VOE to learn joint representations $\nabla$, which can be briefly formalized as:

$$\nabla = \text{VOE}(V)$$

During decoding stage, with the given output of our VOE ($\nabla$), a sentence $S = \{s_1, s_2, \ldots, s_T\}$ with length of $T$ will be generated. We adopt a temporal attention mechanism $\varphi(\cdot)$ to dynamically determines the context feature $C_t$ at $t$-th time step, where $t \in [1, T]$. Therefore, $C_t$ can be computed as:

$$C_t = \varphi_t(\nabla)$$

At every time step $t$, the decoder is trained to predict the $t$-th word ($s_t$) conditioned on the previous $t - 1$ words ($s_{<t}$) and $C_t$. The output of the decoder is thus a conditional probability distribution over words, which can be expressed as:

$$P(s_t|s_{<t}, C_t; \theta)$$

where $\theta$ represents all trainable model parameters. Finally, our VOE-LSTM model can be jointly trained by minimizing the negative log likelihood as follows:

$$\mathcal{L}(\theta) = \min_{\theta} \sum_{t=1}^{T} -\log P(s_t|s_{<t}, C_t; \theta)$$

B. Visual Oriented Encoder

As shown in Fig. 4, our VOE is designed as a hierarchical GRUs, where each GRU layer is to process one modality and does not share parameters with other layers. The detailed computation of GRU is formalized as below:

$$r_n = \sigma(W_r x_n + U_r h_{n-1} + b_r)$$

$$z_n = \sigma(W_z x_n + U_z h_{n-1} + b_z)$$

$$\tilde{h}_n = \tanh(W_h x_n + r_n \odot U_h h_{n-1} + b_h)$$

$$h_n = h_{n-1} \odot z_n + \tilde{h}_n \odot (1 - z_n)$$

where $\sigma$ stands for the sigmoid function, $\odot$ denotes element-wise multiplication and $x_n$ represents the input of the GRU at $n$-th time step. $W$, $U$ and $b$ with different subscripts are all trainable parameters. For simplicity, Eq. (5) can be briefly formalized as below:

$$h_n = \text{GRU}(x_n, h_{n-1})$$
To clarify the mechanism within VOE, we divide the learning process of VOE into two stages, namely context learning and joint representation learning. Lines in red and blue represent regional contexts and global contexts, respectively.

To predict the word $s_t$, $h_t$ is mapped into a distribution over the vocabulary through a fully connected layer:

$$P(s_t|s_{<t}, C_t; \theta) = softmax(W_p h_t + b_p)$$

where $W_p$ and $b_p$ are trainable parameters.

### IV. Experiments

#### A. Datasets

The Microsoft Video Description Corpus [23] (MSVD) contains 1,970 video clips with an average duration of 9.6 seconds, approximately 80,000 description pairs. Following the split settings in prior works [3, 13, 12, 24, 4], we split dataset into training, validation and testing set with 1,200, 100 and 670 videos, respectively. The resulting training set has a vocabulary size of 9,760.

The MSR-VTT corpus [25], consisting of 10,000 web video clips with 20 human-annotated captions per clip, is a relative large and challenging video to language dataset. Following the official split, we utilize 6513, 497 and 2990 video clips for training, validation and testing, respectively. After keeping words appear more than twice, the resulting training set has a vocabulary size of 10544.

#### B. Implementation Details

About feature extraction. For each video, we divide it into 8 snippets and randomly/uniformly sample 8 frames from these snippets in training/evaluation process. Three modalities are considered: image, motion and audio. For image modality, we extract 1536-D image features from Inception-Resnet-V2 [26] pre-trained on the ImageNet. For motion modality, we extract 1536-D motion features from the motion modality. The vocabulary size of 10544.

When it comes to the joint representation learning stage shown in Fig. 4, i.e., $K$-th GRU layer, the global context of previous layer $h^{(K-1)}_N$ contains multimodal clues by learning overall semantic information of top $K-1$ modalities. By fusing $V^{(K)}$ (features from image modality) and multi-scale contexts (learnt from the context learning stage), semantic joint representations can be generated, that is $V = \{v_1, v_2, ..., v_N\}$.

#### C. The Decoding Stage

With the capabilities of modeling long-term temporal dependencies, LSTM is able to achieve remarkable performance in many language modeling tasks [22]. For video captioning, LSTM has been well employed. Specifically, following the most recent state-of-the-art work [6], we adopt a single LSTM layer as our decoder (Fig. 3 (c)), which can be briefly formalized as follows:

$$h_t, c_t = LSTM([s_{t-1}, C_t], (h_{t-1}, c_{t-1}))$$

where $h_t, c_t$ are the hidden state and cell state of the LSTM at time step $t$. The context feature $C_t$ here is determined by leveraging $V$. We follow [3] and adopt temporal attention mechanism to compute $C_t$. So Eq. (2) can be specified as follows:

$$C_t = \varphi_t(V) = \sum_{n=1}^{N} \alpha^n_t \pi_n$$

where $\alpha^n_1, \alpha^n_2, ..., \alpha^n_N$ are attention weights at $t$-th time step. Specifically, $\alpha^n_1$ reveals the relevance between $\pi_n$ and $s_t$ to be generated. With the output of our VOE ($V$) and previous hidden state $(h_{t-1})$, $\alpha^n_1$ can be calculated by:

$$\alpha^n_1 = \exp \left\{ e^n_t \right\} / \sum_{j=1}^{N} \exp \left\{ e^j_t \right\}$$

$$e^n_t = w^T \tanh(W_a h_{t-1} + U_n \pi_n + b_a)$$

where $w$, $W_a$, $U_n$ and $b_a$ are trainable parameters that are shared across all time steps. After computing $C_t$, as shown in Eq. (9), the hidden state at $t$-th time step $h_t$ can be obtained. To predict $t$-th word $s_t$, $h_t$ is mapped into a distribution over the vocabulary through a fully connected layer:

$$P(s_t|s_{<t}, C_t; \theta) = softmax(W_p h_t + b_p)$$

where $W_p$ and $b_p$ are trainable parameters.
network [27] pre-trained on Sports-1M dataset, resulting in 4096-D motion features. For better computation efficiency, we reduce motion features to 512-D features by a fully connected layer. For MSR-VTT, category information and audio modality are considered. We extract audio features by considering 256-D BoAW [28], 260-D Fisher Vector [29], and 128-D VGGish [30] pre-trained on AudioSet. We pad zeros for video without the soundtrack.

About training and evaluation settings. The encoder GRUs and decoder LSTM all have 512 hidden units, and word embedding size is set to 512. To mitigate overfitting, we utilize dropout regularization with a rate of 0.5 in all layers. The whole model is trained by ADAM [31] optimizer with weight decay of $5 \times 10^{-4}$ and an initial learning rate of $10^{-3}$. The batch size for MSVD and MSR-VTT is 64 and 128 respectively. We stop training our model until 500 epochs are reached and the best model is selected according to the performance on the validation set. During the evaluation process, beam search with size 5 is used to generate sentences. We adopt three common metrics including BLEU [32], METEOR [33] and CIDEr [34]. All the metrics are computed using the API released by Microsoft COCO Evaluation Server.

C. Compared Methods

To empirically verify the merit of our proposed VOE-LSTM, we compare our method with the following state-of-the-art methods, which can be mainly grouped into two categories. 1) Temporal and Spatial Structure Exploiting, HRNE [13] and hLSTM [12] are designed for exploiting temporal information of videos, while MAM-RNN [24] and DMRM w/o SS [4] focus on spatial attention. 2) Multimodal-Based Method, including the winner of the MM16 VTT challenge v2t-navigator [5], utilizing textual features from image captions HMVC [17], task-driven dynamic fusion TDDF [15], multimodal attention MA-LSTM [8] and Attentional Fusion [7], multimodal memory modeling M$^2$ [9], and multimodal ensemble with latent topic guidance Enhanced TGM [6]. To verify the effectiveness of our VOE, we implement two other encoders, namely LP (linear projection [5]) and MSLSTM (modality-specific LSTM, [8]). Specifically, LP uses a fully connected layer to learn linear projection of concatenated multimodal features, which disregards non-linear relationships between modalities. MSLSTM utilizes different two-layer LSTMs to learn representations of each modality, which neglects the interaction among modalities. Therefore, we take LP and MSLSTM as baselines to compare with our proposed VOE.

D. Ablation Studies on Model Variants

As MSVD is relative small and excludes audio tracks, we thus only conduct experiments on MSR-VTT in this subsection to evaluate the performance of model variants.

Firstly, we explore the impact of different processing order within VOE. Based on the experimental results shown in Table I, one observations can be obtained that processing the the visual contents of image modality on the top layer of our VOE is able to achieve the best performance. This is consistent with the intuition that auxiliary modalities with longer time-span can provide rich contextual information to assist visual perception. Therefore, the design of visual oriented processing order is reasonable.

Secondly, we analyze the influence of global and regional contexts. Specifically, global contexts are excluded from our VOE when $h_0^{(\lambda)}$ in Eq. (8) is set to zero vector, while regional contexts are excluded when the $\lambda$-th GRU does not take $h_0^{(\lambda-1)}$ as input in Eq. (8). The experimental results are shown in Table II. Obviously, utilizing both global and regional contexts are conducive to generating captions of higher quality. Moreover, it is noted that global contexts matter more when including audio features. This indicates that audio modality can provide comprehensive video-level information.

Finally, we verify the effectiveness of our VOE by comparing it with two other encoders introduced in Sec. IV-C, namely LP and MSLSTM. For fair comparison, the same decoder structure, i.e., a single layer LSTM as proposed in this work, is adopted for LP, MSLSTM and our VOE. As shown in Table IV, our VOE outperforms LP and MSLSTM in all cases. These results indicate that VOE acquires better joint representations, which demonstrate the validity of our VOE and rationality of integrating multimodal and multi-scale contexts in an interactive manner.

\begin{table}[h]
\centering
\caption{The evaluation of different processing order of multi-modalities within our VOE on MSR-VTT. The numbers 1, 2 and 3 explicitly represent the order (small number first).}
\begin{tabular}{|c|c|c|c|c|}
\hline
Processing Order & Image & Motion & Audio & BLEU@4 & METEOR & CIDEr \\
\hline
1 & 2 & - & - & 41.82 & 28.71 & 50.06 \\
2 & 1 & - & - & 42.99 & 28.83 & 50.50 \\
3 & 2 & 3 & - & 44.56 & 29.47 & 51.64 \\
3 & 2 & 1 & - & 45.18 & 29.88 & 52.74 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{The performance of our VOE-LSTM on MSR-VTT with different contexts, regional contexts, Inception-ResNet-V2, C3D and audio features, respectively.}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Features & Context & BLEU@4 & METEOR & CIDEr \\
\hline
I+C & ✓ & 43.21 & 28.80 & 50.09 \\
I+C & ✓ & 42.61 & 28.57 & 49.53 \\
I+C+A & ✓ & 42.99 & 28.83 & 50.50 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{The performance of our VOE-LSTM on MSR-VTT with different contexts, regional contexts, Inception-ResNet-V2, C3D and audio features, respectively.}
\begin{tabular}{|c|c|c|c|c|}
\hline
Features & Context & BLEU@4 & METEOR & CIDEr \\
\hline
I+C & ✓ & 43.21 & 28.80 & 50.09 \\
I+C & ✓ & 42.61 & 28.57 & 49.53 \\
I+C+A & ✓ & 42.99 & 28.83 & 50.50 \\
\hline
\end{tabular}
\end{table}

1https://github.com/tylin/coco-caption
TABLE III
CAPTIONING PERFORMANCE COMPARISON ON MSVD AND MSR-VTT, WHERE B@4, M AND C ARE SHORT FOR BLEU@4, METEOR AND CIDEr. THE SHORT NAME IN FEATURES: G, V, C, R-N, I, A AND T DENOTE GOOGLENET, VGGNET, C3D, N-LAYER RESNET, INCEPTION-RESNET-V2, AUDIO FEATURES AND TEXTUAL FEATURES, RESPECTIVELY. “-” MEANS THAT THE CORRESPONDING RESULTS ARE NOT REPORTED.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Features</th>
<th>MSVD</th>
<th>MSR-VTT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B@4</td>
<td>M</td>
<td>C</td>
</tr>
<tr>
<td>HRNE w/ Attention [13]</td>
<td>G+C</td>
<td>46.7</td>
<td>33.9</td>
<td>-</td>
</tr>
<tr>
<td>hLSTM [12]</td>
<td>G</td>
<td>48.5</td>
<td>31.9</td>
<td>-</td>
</tr>
<tr>
<td>MAM-RNN [24]</td>
<td>G</td>
<td>41.3</td>
<td>32.2</td>
<td>53.9</td>
</tr>
<tr>
<td>DMRM w/o SS [4]</td>
<td>G</td>
<td>50.0</td>
<td>33.2</td>
<td>73.2</td>
</tr>
<tr>
<td>v2 Navigator [5]</td>
<td>V+T</td>
<td>44.3</td>
<td>32.1</td>
<td>68.4</td>
</tr>
<tr>
<td>HMVC [17]</td>
<td>V+C</td>
<td>45.8</td>
<td>33.3</td>
<td>73.0</td>
</tr>
<tr>
<td>TDDF [15]</td>
<td>V+C</td>
<td>52.4</td>
<td>32.0</td>
<td>68.8</td>
</tr>
<tr>
<td>Attentional Fusion [7]</td>
<td>G+C</td>
<td>52.3</td>
<td>33.6</td>
<td>70.4</td>
</tr>
<tr>
<td>MA-LSTM [8]</td>
<td>I+C</td>
<td>52.82</td>
<td>33.31</td>
<td>-</td>
</tr>
<tr>
<td>M3 [9]</td>
<td>I+C</td>
<td>49.26</td>
<td>33.91</td>
<td>83.02</td>
</tr>
<tr>
<td>Enhanced TGM [6]</td>
<td>I+C</td>
<td>51.44</td>
<td>34.40</td>
<td>84.04</td>
</tr>
<tr>
<td>VOE-LSTM (ours)</td>
<td>I+C</td>
<td>45.81</td>
<td>29.88</td>
<td>52.74</td>
</tr>
</tbody>
</table>

TABLE IV
THE PERFORMANCE OF USING DIFFERENT ENCODERS ON MSR-VTT, WHERE I, C, A DENOTE INCEPTION-RESNET-V2, C3D AND AUDIO FEATURES, RESPECTIVELY. FOR FAIR COMPARISON, ALL EXPERIMENTS ADOPT A SINGLE LAYER LSTM AS DECODER.

<table>
<thead>
<tr>
<th>Features</th>
<th>Encoder</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>I+C</td>
<td>LP</td>
<td>41.75</td>
<td>28.14</td>
<td>48.12</td>
</tr>
<tr>
<td></td>
<td>MSLSTM</td>
<td>41.65</td>
<td>28.34</td>
<td>48.97</td>
</tr>
<tr>
<td></td>
<td>VOE</td>
<td>42.99</td>
<td>28.83</td>
<td>50.50</td>
</tr>
<tr>
<td>I+C+A</td>
<td>LP</td>
<td>42.56</td>
<td>28.53</td>
<td>48.82</td>
</tr>
<tr>
<td></td>
<td>MSLSTM</td>
<td>44.33</td>
<td>29.31</td>
<td>50.04</td>
</tr>
<tr>
<td></td>
<td>VOE</td>
<td>45.81</td>
<td>29.88</td>
<td>52.74</td>
</tr>
</tbody>
</table>

E. Experimental Results on MSVD

Table III shows the performances of different models on MSVD. Overall, our VOE-LSTM achieves the best METEOR and CIDEr scores among all these methods, indicating that our method is more likely to generate discriminative words and diverse sentences. Specifically, though both Attentional Fusion and MA-LSTM dynamically assign different weights to different modalities when generating words, they neglect the interaction among multi-modalities. From another prospective, it shows that integrating multimodal and multi-scale contexts is beneficial for promoting video understanding. Moreover, compared with Enhanced TGM, one of the most recent works, our VOE-LSTM surpasses it with relative improvements of 4.4%, 1.4% and 1.2% for BLEU@4, METEOR and CIDEr, respectively. This superior performance shows the necessity of considering multimodal interaction in the encoding stage to solve semantic misalignment. However, in terms of BLEU@4 metric, the performance of our model is slightly worse than Attentional Fusion, MA-LSTM and M3, and decreases by 1.8%, 1.6% and 2.6%, respectively. The reason for the inferior performance may be due to the small size of MSVD, which leads to over-fitting.

F. Experimental Results on MSR-VTT

The performance comparisons on MSR-VTT are summarized in Table III. Our method outperforms other methods. Excluding audio features, our VOE-LSTM (I+C) still surpasses most of the methods (hLSTM, HMVC, TDDF, Attentional Fusion, MA-LSTM and M3). When audio features are included, VOE-LSTM (I+C+A) outperforms all other methods. Specifically, though Attentional Fusion has a better BLEU@4 score than our method on MSVD, however, our VOE-LSTM (I+C+A) outperforms it on MSR-VTT, with relative improvements of 15.4%, 17.2% and 31.9% for BLEU@4, METEOR and CIDEr, respectively. It is also noted that the competitor Enhanced TGM ensembles several models via late fusion to achieve state-of-the-art results, while our VOE-LSTM is one model solution and gains relative improvements of 2.0%, 0.9% and 1.8% for BLEU@4, METEOR and CIDEr, respectively. These remarkable results listed in Table III validate the effectiveness of our VOE in terms of exploiting the inherent contextual information of multimodal features in a visual oriented processing order.

G. Qualitative Analysis

Fig. 5 visualizes six examples on MSR-VTT testing set. The sentences generated by our proposed VOE-LSTM, and two baselines LP and MSLSTM are provided. It is clear to see that our model can capture salient details and generate scene-related keywords, e.g., “gun” in example (a), “golf” in (b), “waterfall” in (c) and “blue” in (f). Moreover, descriptions from our VOE-LSTM are more detailed and diverse. For example, in Fig. 5 (e), though both LP and MSLSTM recognize that this video is related to “cooking”, the sentence generated by our VOE-LSTM is in more details, where the action of the woman (“mixing”) is identified and specific objects like “ingredients” and “bowl” are captured. Similarly, the example (f) also shows the vividness of our generated caption. These examples further confirm that our VOE-LSTM
can learn better joint semantic representations by integrating multimodal features in an interactive manner.

V. CONCLUSION

In this work, we focus on exploring how to effectively integrate multi-modalities in the encoding stage to generate better joint representations for video captioning. Motivated by the way human beings understand the videos with multi-modalities, we propose a novel hierarchical encoder, Visual Oriented Encoder (VOE), to progressively integrate multimodal features in an interactive manner. With such well-designed multimodal fusion, VOE can avoid semantics misalignment and thus learn compact and semantic video representations. Extensive experiments on both MSVD and MSR-VTT video captioning datasets demonstrate the validity and rationality of our proposed method. In our further study, we will focus on improving the capability and generalization of our VOE such as dealing with the over-fitting problem. Additionally, we will analyze a more explainable mechanism for multimodal fusion.

REFERENCES

of recurrent network architectures,” in International Conference on Machine Learning, 2015, pp. 2342–2350.


