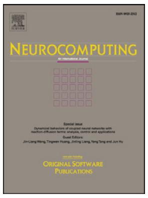
Neurocomputing 315 (2018) 362–370

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Hierarchical attention-based multimodal fusion for video captioning

Chunlei Wu[a], Yiwei Wei[a], Xiaoliang Chu[a], Sun Weichen[b,c], Fei Su[b,d], Leiquan Wang[a,*]

[a] *College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao, China*
[b] *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China*
[c] *First Research Institute of the Ministry of Public Security of PRC, China*
[d] *Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China*

ARTICLE INFO

Article history:
Received 28 January 2018
Revised 15 June 2018
Accepted 17 July 2018
Available online 8 August 2018

Communicated by Paolo Remagnino.

Keywords:
Video captioning
Multi-modal
Attention mechanism

ABSTRACT

Attention based encoder-decoder models have shown a great success on video captioning. Recent multimodal video captioning mainly focused on applying the attention mechanism to all modalities and fusing them in the same level. However, the connections among specific modalities have not been investigated in the fusion process. In this paper, the expressivity of uni-modal is firstly investigated. Due to the characteristic of attention mechanism, an instance-level of visual content is exploited to refine the temporal features. Then, a semantic detection architecture based on CNN+RNN is also employed on the spatiotemporal content to exploit the correlations between semantic labels for better video semantic representation. Finally, a hierarchical attention-based multimodal fusion model for video captioning is proposed by jointly considering the intrinsic properties of multimodal features. Experimental results on the MSVD and MSR-VTT datasets show that the proposed method has achieved competitive performance compared with the related video captioning methods.

1. Introduction

The task of video captioning is to summarize an input video with a natural language description automatically. It has a variety of practical applications, such as describing videos to the blind, improving index and search quality for online videos, etc. However, video captioning is a challenging domain in computer vision and machine learning society, which involves understanding of objects, persons, scenes, incidents, temporal relations and many others.

Recently, great progress have been made in image captioning, especially attention-based encoder-decoder framework [1,2]. The attention-based image captioning method transforms the visual feature (CNN encoder) to the target caption (RNN decoder) [3]. Exploiting better visual feature in the encoder procedure has been proved to be effective in attention-based approaches [4]. The same principle is also applied equally to video captioning.

Different from image captioning, video possesses a more complex situation than single images. Audio features, spatiotemporal features and semantic label features are all valuable information for video processing beyond visual features [5]. From a multimodal perspective, employing multi-modal features can also be a better solution for video captioning due to the complementariness among multi-modal features [6–8]. When abundant modalities are presented, choosing an optimal fusion architecture is more challenging.

Traditional early fusion method has an unsatisfying performance [9]. Meanwhile, late fusion and graph-based fusion strategies are not suitable for encoder procedure of sequence-to-sequence framework. Therefore, it is challenging to discover a multi-modal fusion mechanism during encoder procedure for video captioning. Attention mechanism is also a hotspot in video captioning [6,10]. The directed attention-based multi-modal fusion strategy (see Fig. 1 (a)) averages the different modal context vectors in the encoder phase at each step [11]. This strategy focuses on the effective part of each modality to generate descriptions. However, inherent structures of different modalities are destroyed and the importance of different modalities can not be discriminated. Attention-based multimodal fusion [7] is put forward to solve this problem by using inter-modality attention. The authors of [7] fuse all multimodal features in an attention layer to generate video descriptions (see Fig. 1 (b)). The importance of different modalities is discriminated by the learned attention weights. However, the features varied from low-level feature to high-level feature, from local feature to global feature. The combinations of various modalities with an inter-modality layer seem to be insufficient to highlight the dominant features. A hierarchical attention-based fusion

[*] Corresponding author.
E-mail addresses: wuchulei@upc.edu.cn (C. Wu), 360976808@qq.com (Y. Wei), 772063390@qq.com (X. Chu), weichen.sun@hotmail.com (S. Weichen), sufei@bupt.edu.cn (F. Su), richiewlq@gmail.com (L. Wang).

https://doi.org/10.1016/j.neucom.2018.07.029
0925-2312/© 2018 Elsevier B.V. All rights reserved.

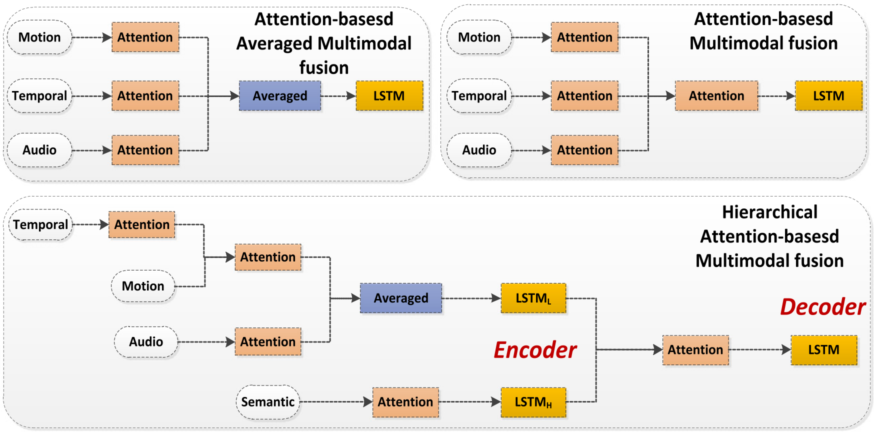


Fig. 1. Attention-based Averaged Multimodal Fusion and Attention-Based Multimodal Fusion represent two classic fusion strategies for video captioning. Hierarchical Attention-based Multimodal Fusion is the proposed framework in which different modalities are fused by different hierarchical attention layers.

strategy can take full advantage of informative modal feature parts, meanwhile, weaken the redundant modality.

Based on the considerations above, we propose a hierarchical attention strategy (HATT) to fuse information across different modalities (see Fig. 1 (c)). The effectiveness of uni-modal feature is firstly considered in this paper. Instances-level temporal features is exploited to replace the pretrained CNN features as a basis under attention mechanism. In addition, the correlations between semantic labels are also utilized as a key modality for video captioning. With the effective feature representation, the highly correlated modalities are firstly fused, then the less correlated ones. Low-level attention layer, high-level attention layer and sequential attention layer are used to fuse the multi-modalities with a progressive manner.

To summarize, the main contributions of this paper are as follows:

- Instance-level of visual feature is exploited to make up the defects of traditional attention-based methods, in which feature maps are split uniformly and the integrity of object is broken.
- A semantic detection method based on CNN+RNN is also proposed by exploiting the correlations between semantic labels for better semantic representation.
- A hierarchical attention-based multi-modal fusion model (HATT) is proposed for video captioning, which contains multiple attention fusion layers. The different modalities are fused with a progressive attention manner to utilize the complementarities of multiple modalities.
- Comprehensive experiments are conducted to empirically analyze the proposed method. The results on the challenging MSVD and MSR-VTT datasets show that the proposed method has achieved competitive performance compared with the related video captioning methods.

2. Related work

Lots of effective models have been developed for video captioning. In general, the video captioning methods can be divided into two categories: temporal-based video captioning [12–15] and multimodal-based video captioning [6,7,16].

Temporal-based approaches generate caption merely based on the temporal features extracted by the traditional CNN architectures [17]. In order to better capture temporal information, an end-to-end LSTM-based model [12] is presented for video captioning, in which frame-level local 2D CNN features are assembled with an average pooling operation to act as the inputs of LSTM decoder. However, time dependency information is lost since average pooling ignores the order of input sequences. To tackle the problem, an end-to-end sequence-to-sequence model (s2vt) [14] is proposed to generate captions for videos. It exploited a stacked LSTM to read the sequence of CNN features for each frame. However, it is deficient for traditional LSTM units to deal with long video-range dependencies [13]. For this reason, Pan et al. [13] proposed a hierarchical recurrent neural encoder (HRNE) for video captioning, which exploited multiple time-scale abstraction of the temporal information with a two-layer LSTM network. After that, Pan et al. [15] proposed another framework which explored the learning of LSTM, aiming to maximize the probability of the next word locally. Temporal-based approaches have achieved effective performances on video captioning. However, other features, such as audio, spatiotemporal and semantic label, are also valuable for video processing, have been neglected.

Multimodal-based approaches utilize temporal, motion, audio and other features to generate video caption. Recently, inspired by the attention mechanism [6,7,16], several multimodal-based approaches have been proposed. To obtain a representative and high-quality description for a target video, Yu et al. [16] proposed an RNN-based approach(h-RNN), in which both temporal and spatial features were exploited to selectively focus on specific visual elements during generation. However, the correlations between different modalities were not considered. In order to tackle the problem, Hori et al. [7] presented a modality-dependent attention mechanism which expanded the attention model to selectively attend not only to specific times or spatial regions, but also to specific modalities, such as image, motion and audio features. Nevertheless, high-level semantic information was not involved in. To introduce semantic information, an extensible approach was proposed by Long et al. [6] to jointly leverage several sorts of visual features and semantic attributes by utilizing modality-dependent attention layers. However, it seems to be insufficient to highlight the dominant features by fusing all features in the same level. In this paper, a hierarchical attention-based multimodal fusion method is exploited by jointly considering the intrinsic properties of multimodal features with a progressive attention manner for video captioning.

3. Traditional attention-based encoder-decoder model

Recurrent Neural Networks (RNNs) has shown a great success in the process of machine translation, speech recognition, image and video captioning [14,18,19]. On the basis of RNN, Long Short-Term Memory (LSTM)[20] is put forward to avoid the vanishing gradient problem without losing too much information.

3.1. LSTM unit

A basic LSTM unit is built up with three gates (input i_t, forget f_t and output o_t) and a memory cell m_t. Specifically, it brings all heterogeneous gates into effective teamwork. f_t determines the legacy of the information by the cell, and prevents the gradient from vanishing. Morever, o_t is used to control how much of the current cell state is filtered out in the transmitting process. In general, the memory cell and gates in a LSTM block are defined as follows:

$$i_t = \sigma(W_i y_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f y_t + U_f h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_o y_t + U_o h_{t-1} + b_o) \tag{3}$$

$$g_t = \phi(W_g y_t + U_g h_{t-1} + b_g) \tag{4}$$

$$m_t = f_t \odot m_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \phi(m_t) \tag{6}$$

where the weight matrices W, U and b are parameters to be learned. y_t represents the input vector for the LSTM unit at each time t. σ represents the logistic sigmoid nonlinear activation function and ϕ denotes the hyperbolic tangent function tanh.

3.2. Encoder-decoder framework

Given a video input x, different encoder networks encode it into various continuous representation spaces:

$$V = \zeta_{1\ldots n}(x) = \{v^1, v^2, \ldots v^n\} \tag{7}$$

where $\zeta_{1\ldots n}$ usually denote different feature-extraction networks (Resnet, C3D) and $\{v^1, v^2, \ldots v^n\}$ represent different multimodal features (temporal, motion). In most cases, LSTM is chosen as a decoder network to model V to generate a description $D = \{w_1, \ldots w_L\}$ for x, where $\{w_t, t = 1 \ldots L\}$ is the tth word in the description. Usually, the multimodal features $V = \{v^1, v^2, \ldots v^n\}$ and the last generated word w_{t-1} are combined as the current input vector y_t. The LSTM unit updates its hidden state h_t, memory cell m_t and the tth word w_t based on its previous hidden state h_{t-1}, previous memory cell m_{t-1} and the current input y_t:

$$(h_t, m_t, w_t) = LSTM(h_{t-1}, m_{t-1}, y_t) \tag{8}$$

The LSTM updates its hidden state recursively to generate each word until the end-of-sentence tag.

3.3. Attention mechanism

Attention mechanism produces a spatial map highlighting image regions relevant to each generated word. Assuming that we have a processed feature matrix F, which contains the specific feature or combinations of multimodal features. The attention mechanism calculates attention weights α_t for each element of F, conditioning on the last hidden states h_{t-1} of the decoder LSTM.

Each basic attention score [7] e_t^i is computed as:

$$e_t^i = w^T \phi(W_q F^i + U_q h_{t-1} + b_q) \tag{9}$$

where w, W_q, U_q and b_q are the learned parameters, F^i is the ith element of F, ϕ is the $tanh$ activation function. Then each basic attention score is fed to a sequential softmax layer to obtain the corresponding attention weight α_t^i:

$$\alpha_t^i = exp\{e_t^i\} / \sum_{j=1}^{M} exp\{e_t^j\} \tag{10}$$

where M represents the element number of F. Finally, the weighted context of F is defined as c_t:

$$c_t = \sum_{i=1}^{M} \alpha_t^i F^i \tag{11}$$

For convenience, we denote the attention mechanism as: $c_t = Attention(h_{t-1}, F)$.

4. Multi-modal features for video captioning

In this paper, temporal features, motion features, audio features and semantic label features are used for video representation. Particularly, temporal features and semantic label features are illustrated in detail.

4.1. Temporal features

Different from the previous methods [7,14] in which the extracted CNN feature maps are split uniformly with the attention mechanism, temporal features are extracted at the level of instances by adapting fully convolutional instance-aware semantic segmentation method (FCIS) [21]. The processed region proposed features are finally used as the target temporal features.

The temporal feature of the first frame is extracted by the traditional CNN to preserve the global information of the video. Except the first frame, a region proposal network (RPN) is performed as the ROIs generator on the other frames based on the pre-trained FCIS. The generated ROIs are then ranked by the region proposal scores for each frame. Consequently, the top-k scoring regions with the exact ranges for each frame are obtained. A set of features on the k regions are extracted from the $conv5$ feature maps. Mean pooling operations are then performed on the regional features. A frame is represented by a feature matrix of K*1024 dimensions. Finally, the mean pooling operations is performed on the feature matrix again to build a feature vector of 1024 dimensions. Suppose there are L frames for a video, a L*1024 target features will be generated.

The temporal features consist of a set of salient image regions, with each region represented by a mean pooling convolutional feature vector. It eliminates the background feature redundancy among different frames. In the context of attention mechanism, the traditional feature maps with equally-sized grid destroy the integration of objects. In contrast, the proposed instance level of temporal feature representation tries to keep the original structures of each object.

4.2. Semantic label features

Semantic concepts have been proved to be a key ingredient for video (image) captioning [22,23]. Most existed works extracted semantic labels independently. However, the correlations between semantic labels within the same video have been neglected. In this part, the semantic concepts detection are treated as a multi-label classification task with correlated semantic labels. A CNN+RNN network is exploited on the spatiotemporal visual content of videos to get semantic labels that rely on each other.

Fig. 2 shows the architecture of the proposed semantic label detection method. A label k is represented as a one-hot vector $e^k = [0, \ldots 0, 1, 0 \ldots, 0]$, in which 1 is at the kth location. Then, an

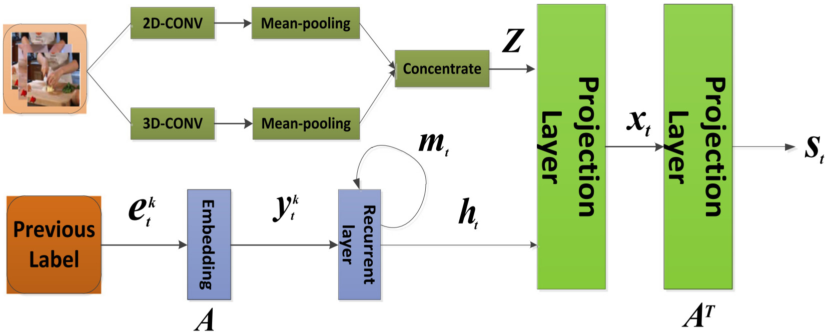


Fig. 2. The CNN+RNN based semantic label detection architecture. The predicting label depends on the video features and the previous label. The correlations between semantic labels are used for semantic concepts detection.

embedding matrix A is used to transfer the one-hot label vector e^k into an embedded label vector y_t^k:

$$y_t^k = A.e_t^k \tag{12}$$

Here, e_t^k is obtained by the probability of predicted labels at previous time step (S_{t-1}). After that, the label is transmitted into the recurrent layer to model the co-occurrence dependencies in its hidden recurrent states by learning nonlinear functions:

$$m_t = f_m(m_{t-1}, y_t^k) \tag{13}$$

$$h_t = f_h(h_{t-1}, y_t^k) \tag{14}$$

where h_t and m_t are the hidden state and the memory of the recurrent layer at time step t respectively, $f_m(.)$ and $f_h(.)$ are the nonlinear LSTM functions of Eqs. (5) and (6) in section 3.1.

The output of the recurrent layer o and the concentrated video features Z are projected into the same space.

$$x_t = \sigma(U_m^x m_t + U_Z^x Z) \tag{15}$$

where U_m^x and U_V^x are the projection matrices for recurrent layer output and video features respectively and σ is the sigmoid nonlinear activation function. The video features Z is produced by concatenating the mean pooling of 2D CNN features [17] and 3D CNN features [24].

Finally, the probability of predicted label can be computed by multiplying the transpose of embedding matrix A and x_t with a softmax.

$$s_t = softmax(A^T x_t) \tag{16}$$

Given the target ground truth labels, the objective function is to minimize the cross entropy loss (XE):

$$L(\theta) = -\sum_{t=1}^{T} log(p_\theta(s_t|s_1, ..., s_{t-1}, Z)) \tag{17}$$

where Z is the concatenation of 2D CNN and 3D CNN features and θ represents the model parameters.

5. Attention-based hierarchical multi-modal fusion model

In this section, our approach is described in detail. Fig. 3 shows the overview of our proposed approach for video captioning.

Different from traditional attention-based multimodal fusion methods [7,11], this new attention-based hierarchical multi-modal fusion model (HATT) is proposed for video captioning by exploiting the complementariness of multimodal features with a progressive manner. HATT consists of three attention layers: low-level attention layer, high-level attention layer and sequential attention layer. Low-level attention layer deals with temporal, motion and audio features with intra- and inter-modality attention. High-level attention layer selectively focuses on the semantic labels only. Sequential attention layer incorporates the hidden sequential information $(\widehat{h}^L, \widehat{h}^H)$ generated by the encoded low-level attention and high-level attention. HATT is set up with the sequence-to-sequence learning framework [14], which is an encoder-decoder framework. The first LSTM encodes the outputs of the high-level (H_t) and low-level (L_t) attention parts, and the second LSTM decodes the outputs of sequential attention ($\widehat{h}$) part into a sequence of words.

5.1. Intra- and inter- modality attention

Intra-modality attention. Attention mechanism is performed on a specific feature (such as, temporal or audio feature). Given a specific feature $v \in V$, the attention weights α_t and context vector c_t for the feature v is calculated based on the attention mechanism. The process is the same as the traditional attention introduced in Section 3.3. For simplicity, the intra-modality attention is defined as $c_t^{intra} = Intrattention(h_{t-1}, v) = Attention(h_{t-1}, v)$.

Inter-modality attention. Attention mechanism is performed on multimodal features. Given n multimodal features $V = \{v^1, v^2, ...v^n\}$, a feed-forward layer is added to reduce the dimensionality:

$$R^i = \phi(W_p^i v^i + b_p^i), i = 1, 2, \ldots, n \tag{18}$$

where W_p^i and b_p^i are the learned parameters for ith modality feature, ϕ is the *tanh* function and $R = [R^1, R^2, \ldots R^n]$ represents the translated multimodal features with the same dimension. Then, the attention weights α_t for the translated multimodal features R is computed based on the traditional attention mechanism. We denote the inter-modality attention at time step t as $c_t^{inter} = Interattention(h_{t-1}, \{v^1, v^2, \ldots v^n\}) = Attention(h_{t-1}, R)$.

5.2. Hierarchical attention-based multimodal fusion model

Then, the hierarchical attention-based multimodal fusion (HATT) will be discussed in detail based on the temporal feature (v^{tem}), motion feature (v^{mot}), audio feature (v^{aud}) and semantic labels (v^{sem}).

For low-level attention layer, we define audio and visual features as low-level features. Instead of a naive averaging of the feature vectors, different features are fused by different properties. Firstly, the **intra-modality attention** layer is applied to temporal features and audio features to exploit the time-frequency localization characteristics:

$$c_t^{tem} = Intrattention(h_{t-1}, v_t^{tem}) \tag{19}$$

$$c_t^{aud} = Intrattention(h_{t-1}, v_t^{aud}) \tag{20}$$

where c_t^{tem} and c_t^{aud} are respectively the temporal context and audio context at time step t and h_{t-1} represents the last hidden state

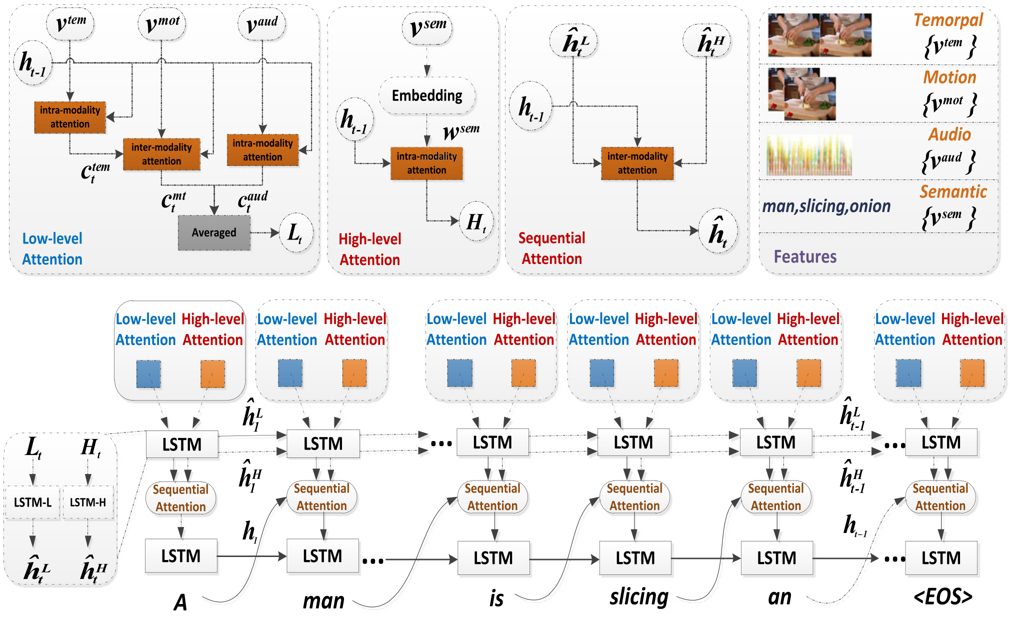


Fig. 3. Overview of the proposed hierarchical attention-based multimodal fusion (HATT) video captioning model. The s2s framework acts as the core and three attention layers are respectively deployed in different places, namely, low-level attention, high-level attention and sequential attention. In the language model, the decoder LSTM contains low-level branch and high-level branch. They use the outputs of the corresponding attentions as inputs. Note that the sequential attention works on the basis of sequential representation of the modals. HATT fuses various representations at different depths with a progressive manner.

of the decoder. Then, the motion features v^{mot} and audio context c^{tem} are fed into the **inter-modality attention** layer:

$$c_t^{mt} = Interattention(h_{t-1}, \{v_t^{mot}, c_t^{tem}\}) \tag{21}$$

where c_t^{mt} denotes the visual context at time step t.

Finally, the sequence of visual context is fed to the visual branch of the encoder LSTM by adding a multimodal layer:

$$c_t^{mat} = \phi(w^T[W^{mt} \odot c_t^{ma}, W^{aud} \odot c^{aud}]) \tag{22}$$

$$\widehat{h}_t^L = LSTM_L(\widehat{h}_{t-1}^L, \widehat{m}_{t-1}^L, c_t^{mat}) \tag{23}$$

where w^T, W^{ma} and W^{aud} are the learned parameters, ϕ is an activation function, $LSTM_L$ denotes the visual branch of the encoder LSTM and $\widehat{h}_t^L$ and $\widehat{m}_t^L$ represent the hidden state and memory cell of $LSTM_L$, respectively.

For high-level attention layer, we define semantic features as high-level features. Given the detected attributes vector v^{sem}, we feed them into an embedding layer and get a new embedded attribute matrix w^{sem}. Subsequently, an **intra-modality attention** layer is applied to the embedded attribute matrix:

$$c_t^{sem} = Intrattention(h_{t-1}, w^{sem}) \tag{24}$$

where h_{t-1} is the the last hidden state of the decoder LSTM and c_t^{sem} represents the semantic context at time step t.

Then, the sequence of semantic context is fed to the semantic branch of the encoder LSTM:

$$\widehat{h}_t^H = LSTM_H(\widehat{h}_{t-1}^H, \widehat{m}_{t-1}^H, c_t^{sem}) \tag{25}$$

Here, $LSTM_H$ is the semantic branch of the encoder LSTM, $\widehat{h}_t^H$ and $\widehat{m}_t^H$ represent the hidden state and memory cell of $LSTM_H$, respectively.

For sequential attention layer, we apply the **inter-modality attention** layer to the outputs, $\{\widehat{h}_t^L, \widehat{h}_t^H\}$ which are generated from different branches of the encoder LSTM. The sequential attention layer can be defined as:

$$\widehat{h}_t = Interattention(h_{t-1}, \{\widehat{h}_t^L, \widehat{h}_t^H\}) \tag{26}$$

where $\widehat{h}_t$ is the weighted hidden state of the encoder network.

In the following, the $\widehat{h}_t$ is transmitted into the decoder:

$$h_t = LSTM(h_{t-1}, m_{t-1}, \widehat{h}_t) \tag{27}$$

where h_{t-1} and m_{t-1} are the last hidden state and memory cell of the decoder, respectively. Finally, a softmax function is applied to get the probability distribution over the words in the vocabulary:

$$\varpi_t = softmax(W_l^T h_t) \tag{28}$$

here, W_l^T is the learned matrices and softmax(.) denotes a softmax function.

Given the ground-truth sentences, the loss is defined as:

$$L(\varphi) = -\sum_{t=1}^{m} logP_{\varphi}(\varpi_t|\varpi_1, ..., \varpi_{t-1}, V) \tag{29}$$

where V is the feature vector output by the video encoder and φ represents the learned model parameters.

By using different attention layers, we effectively consider the relationship between the different modals and optimize the quality of the generated captions.

6. Experimental results

6.1. Datasets

MSVD: The proposed model is evaluated on the Microsoft Research Video Description Corpus (MSVD) citeYouTube2Text. MSVD consists of 1,970 video clips with multiple natural language descriptions. Each video clip is annotated with multiple human generated descriptions in several languages. Only English descriptions are used, about 41 descriptions per video. In total, the dataset consists of 80,839 video-description pairs. Each description contains about 8 words on average. Following the previous work [14], 1,200 videos are used for training, 100 videos for validation and 670 videos for testing.

MSR-VTT: MSR Video-to-Text (MSR-VTT) dataset [25] is also used for evaluation. MSR-VTT provides 10,000 web video clips. Each video is annotated with about 20 natural sentences. There are 200,000 video-caption pairs in total. The proposed video captioning models are trained with the official training and validation set, which consists of 6,513 and 497 video clips respectively. The rest 2,990 video clips are used for testing.

6.2. Preprocessing

Low-level features: Motion features, temporal features, and audio features are extracted as low-level features in the experiments. **(1)** A pretrained C3D network [24] is used to extract motion features. The C3D net acquires a video and outputs a fixed-length feature vector every 16 frames. The activation vectors are extracted from fully-connected layer fc6-1, which has 4096-dimensional features. **(2)** Instance-level regional features are extracted as temporal features based on the methods described in Section 4.1. For a target video, each frame obtains a set of 1024-dimensional frame-wise feature vectors which are extracted from the positive region proposals. The number of regions for each frame is set to 45. **(3)** For audio preprocessing, we apply the extraction method which is proposed by Chiori [7]. It results in a sequence of 260-dimensional vectors.

High-level features: The 150 most common words for MSVD are selected in the training captions to determine the vocabulary of semantic concepts, which includes the most frequently nouns and verbs. The same operation is also performed on MSR-VTT, where the number of words is set to 240. For both datasets, 4 nouns and 3 verbs are used for semantic labels. According to the conclusion [26] that it is difficult to train a reliable semantic detector using the MSVD dataset alone, due to its limited amount of data. We also utilize additional training data from COCO dataset.

To generate correlated semantic labels, CNN+RNN based semantic labels detection method presented in 4.2 is applied. As illustrated in 4.2, 2D CNN [17] and 3D CNN [24] features are concatenated together with a mean pooling operation as the spatiotemporal visual content of videos. The previous label is embedded into the RNN layer. The output of RNN and the spatiotemporal visual features are combined to generate correlated high-level semantic label features.

6.3. Evaluation metrics

BLEU [27], METEOR [28], and CIDEr [29] are adopted as evaluation metrics. The Microsoft COCO evaluation server [30] is utilized to compute the metric scores. By all the three metrics, higher scores indicate that the generated captions are considered to be closer to the annotated captions created by humans.

6.4. Experimental settings

CNN+RNN model for semantic label detection. In the experiments, the proposed CNN+RNN model for semantic label detection is built based on Resnet-152 (2D-CONV) and C3D (3D-CONV). The dimensions of the label embedding and LSTM layer are 512 and 1024, respectively. The Adam optimizer is used with the weight decay rate 0.0001. In addition, the dropout rate is set to 0.5.

HATT model for video captioning. The number of hidden units in the encoder and decoder LSTMs are both 512. And the dropout rates of the encoder and decoder LSTMs are set to 0.5. In addition, Adam optimizer [31] is adopted to optimize the loss function, with the learning rate 10^{-4}. The training batchsize is 128 for MSVD and MSR-VTT. The number of epochs for MSVD and MSR-VTT are set to 20 and 50 respectively. All experiments in the proposed method are trained on a pair of Nvidia TitanX GPUs with 12GB memory.

Table 1
Comparisons with the state-of-the-art methods on MSVD dataset. "*" stands for the multi-modal fusion method, and "#" means that semantic labels are used in the fusion method.

Method	MSVD					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
LSTM-YT [15]	–	–	–	0.333	0.291	–
S2VT [14]	–	–	–	–	0.298	–
h-RNN [16]	0.815	0.704	0.604	0.499	0.326	0.658
HRNE-A [13]	0.792	0.663	0.551	0.438	0.331	–
TA* [32]	0.800	0.647	0.526	0.419	0.296	0.517
MA* [7]	0.801	0.688	0.593	0.496	0.310	0.655
TM-P-HRNE*# [6]	0.829	0.720	0.627	0.528	0.334	0.705
SCN*# [26]	0.810	0.697	0.606	0.511	0.335	**0.777**
HATT(45)*#	0.829	**0.725**	**0.631**	**0.529**	**0.338**	0.738

Table 2
Comparisons with the related methods on MSR-VTT dataset.

Method	MSR-VTT			
	B-4	METEOR	CIDEr	ROUGE-L
v2t navigator	0.408	0.282	0.448	**0.609**
Aalto	0.398	0.269	**0.457**	0.598
VideoLAB	0.391	0.277	0.441	0.606
HATT(45)	**0.412**	**0.285**	0.447	0.607

6.5. Experimental results and discussion

To validate the effectiveness of the proposed methods, experiments are conducted on both MSVD and MSR-VTT datasets. The experimental procedure are elaborated on MSVD. Meanwhile, the experimental results are also displayed on MSR-VTT with the best experimental conditions.

Comparisons with the related methods.

In this subsection, the comparison results of the proposed HATT with the related methods on the MSVD dataset are shown in Table. 1. HATT achieves the best performance among all methods listed in Table. 1.

Overall, the multimodal video captioning methods outperform those uni-modal methods (e.g. LSTM-YT [15], S2VT [14], h-RNN [16], HRNE-A [13]). The complementariness among multimodalities is beneficial for video captioning. In addition, the semantic labels are significant factors in multi-modal video captioning. The methods with semantic labels ("#") show superiority in the multi-modal video captioning.

Comparisons are also made on the attention-based multi-modal fusion methods, which are shown in Fig. 1. The proposed HATT exceeds TA [32] and MA [7] on all of the evaluation metrics. To further demonstrate the superiority of HATT, only low-level features are used in HATT to make a fair comparison with TA [32] and MA [7]. Specifically, Resnet101 are used to extract temporal features for all the three methods. The results are shown in Table. 3. HATT(Resnet101) also achieves the best performance than the other attention-based multimodal fusion methods. Fusing multiple modalities at different levels with attention mechanism leads to a significant improvement for captioning.

Particularly, experiments are also conducted on MSVD by changing the fusion order of low-level features. Temporal-motion-audio is 1% higher than temporal-audio-motion in CIDEr. The phenomenon is in accordance with [33] that used for gesture

Table 3
Comparison between attention-based multi-model fusion methods.

Method	MSVD					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
TA* [32]	0.800	0.647	0.526	0.419	0.296	0.517
MA* [7]	0.801	0.688	0.593	0.496	0.310	0.655
HATT* (Res101)	0.809	0.693	0.601	0.496	0.317	0.659
HATT*# (Res101-semantic)	**0.821**	**0.716**	**0.618**	**0.515**	**0.335**	**0.723**

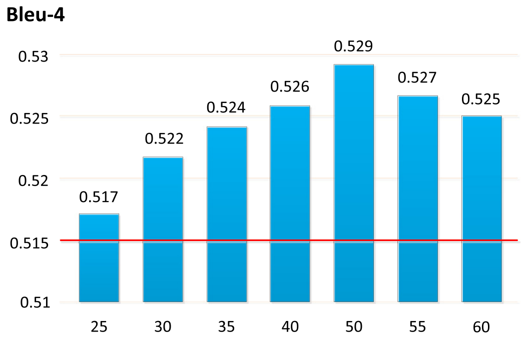


Fig. 4. Variations of performance (BLEU-4) with the increased number of proposed regions for each frame on MSVD. The red line (around 0.515) represents the model which takes resnet101 to extract temporal features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

recognition, in which highly correlated modalities are fused of a priority, then the less correlated ones.

To validate the effectiveness of proposed HATT, experiments are also conducted on the 2016 MSR-VTT dataset. The top 3 methods in **MSR Video to Language Challenge** 2016 are compared. The results are shown in Table 2. HATT also achieves the good performance in some evaluation metrics.

Influences of the number of proposed regions for temporal features representation. The number of proposed regions K of each frame for temporal feature representation is an important factor of HATT that should be investigated. Experiments with varied K proposed regions are conducted on the MSVD dataset to analyze the role of instance-level temporal features in HATT. Experimental results in Fig. 4 show the variations of BLEU-4 with the increase of K on the MSVD dataset. HATT is an attention-based video caption method. An advantage of attention-based methods is that the decoder makes a mapping between generated words and visual regions. The proposed FCIS-based instance-level temporal features try to retain the integrity of object, which obtains a more complete representation than the equally-splitted methods [7] under the attention mechanism. HATT achieves the best performance when $K = 45$ on the MSVD dataset. A small number of proposed regions will lost object information during the encoder phase, while a large number of proposed regions will mix excess noise. Furthermore, the red line in Fig. 4 represents the Bleu-4 scores of HATT(resnet101) in which the temporal features are extracted by the resnet101. It is clear that even the worst model HATT(25) performs better than HATT(resnet101) in BLEU-4 score. This phenomenon shows the effectiveness of the proposed temporal features extraction method.

Effectiveness of semantic representation. The semantic concept detection of HATT was trained with a CNN+RNN architecture in which all the layers of the entire network were unfrozen and finetuned with Adam optimizer. In order to prove the effectiveness of the proposed semantic representation, the semantic features which are extracted from [26] with the same data processing procedure are compared. The results are shown in Table 4. HATT(CNN+RNN) achieves better performance than HATT(CNN) on all metrics in HATT. CNN+RNN based semantic representation makes more important contributions than that with CNN only for video captioning. The phenomenon indicates that it is worthwhile to exploit correlations between semantic concepts for semantic representation.

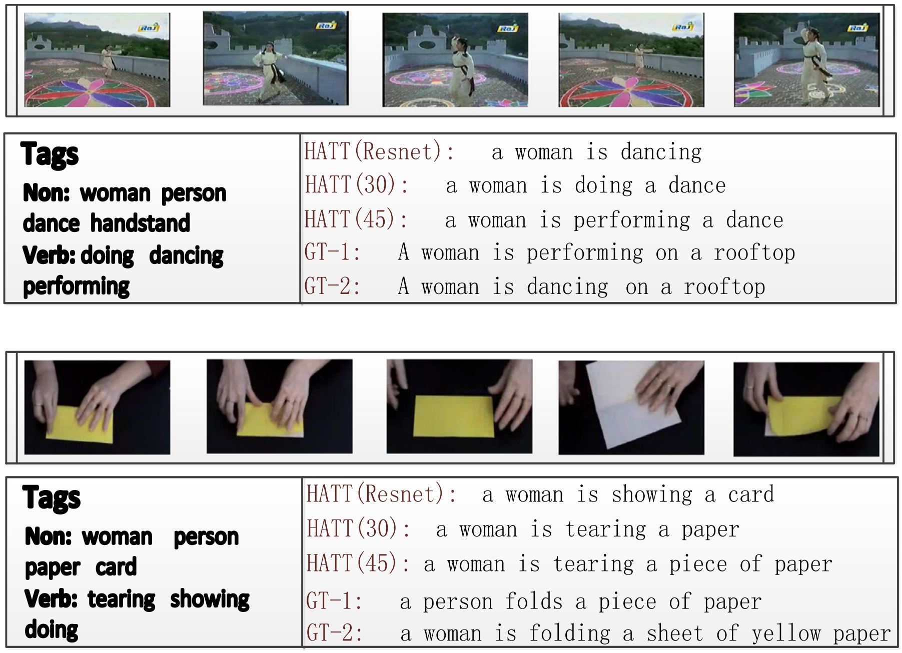
Table 4
Comparison between the hierarchical models based on different semantic detection methods.

Method	MSVD					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
HATT(CNN)	0.825	0.722	0.624	0.521	0.331	0.727
HATT(CNN+RNN)	**0.829**	**0.725**	**0.631**	**0.529**	**0.338**	**0.738**

Qualitative analysis.

To reflect the quality of the actual generated descriptions intuitively, two video clips from MSVD and the generated descriptions are presented in Fig. 5. Particularly, the corresponding labels are also presented with different semantic concepts detection methods. The generated semantic labels and descriptions correspond well with the video clips, especially HATT(45). The extracted representations (temporal, motion, audio and semantic features) of video streams are fused gradually with a hierarchical attention-based manner. The descriptions are generated based on the different characteristics of multi-modalities. For example, the description “a woman is performing a dance” tends to rely more on motion feature than audio feature. While, the description “a man is listening to music” depends more on audio feature than motion feature. The motion and audio features have relatively complemented relationship. For the description “a woman is tearing a piece of paper”, the motion feature “tearing” performs as a clue to guide the temporal feature “paper”. In most instances, the motion and temporal features have coexistence relationship. In summary, a simple fusion strategies can not fully exploit the relationship among multi-modalities. The proposed hierarchical attention-based multi-modal fusion strategy for video captioning is worth investigating.

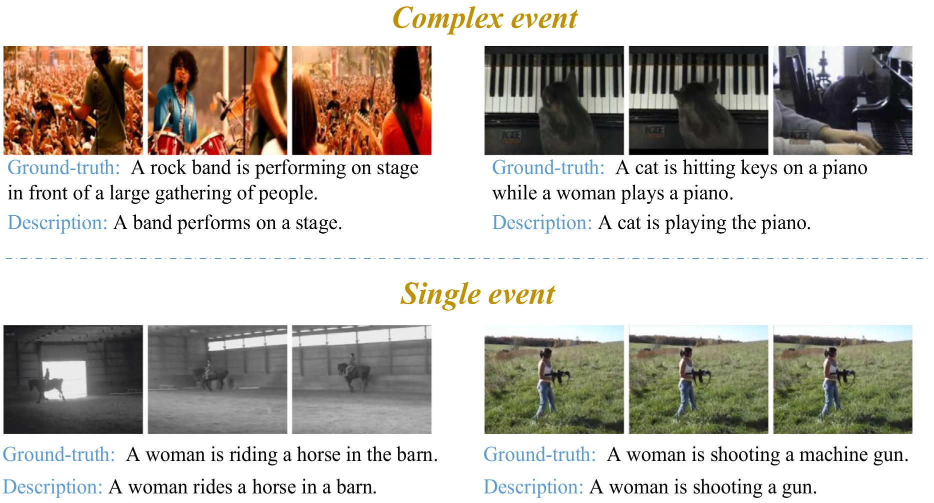
Although the proposed HATT performs well in the reported evaluation metrics, it still has some limitations. As can be seen in Fig. 6, HATT performs poorly when the target video clip contains a complex event. The phenomenon also existed in other attention-based multimodal fusion video caption methods. HATT is exploited by jointly utilizing the intrinsic properties of multimodal features with a progressive attention manner. The complementariness of multimodal features are used to generate description for the same scene. However, in complex scenes, the multimodal features may interfere with each other, leading to a worse performance. In addition, time cost is another limitation of the proposed HATT. Particularly, FCIS operations spend too much time on temporal feature extraction. About 0.2 s per frame to extract temporal features on a Nvidia Titanx(pascal) GPU. The time cost is longer than traditional feature extraction methods, such as Resnet101 [17].



Tags	
Non: woman person dance handstand Verb: doing dancing performing	HATT(Resnet): a woman is dancing HATT(30): a woman is doing a dance HATT(45): a woman is performing a dance GT-1: A woman is performing on a rooftop GT-2: A woman is dancing on a rooftop

Tags	
Non: woman person paper card Verb: tearing showing doing	HATT(Resnet): a woman is showing a card HATT(30): a woman is tearing a paper HATT(45): a woman is tearing a piece of paper GT-1: a person folds a piece of paper GT-2: a woman is folding a sheet of yellow paper

Fig. 5. Examples of generated captions on MSVD. The semantic labels is detected based on CNN+RNN by exploiting the correlations between labels.



Complex event

Ground-truth: A rock band is performing on stage in front of a large gathering of people.
Description: A band performs on a stage.

Ground-truth: A cat is hitting keys on a piano while a woman plays a piano.
Description: A cat is playing the piano.

Single event

Ground-truth: A woman is riding a horse in the barn.
Description: A woman rides a horse in a barn.

Ground-truth: A woman is shooting a machine gun.
Description: A woman is shooting a gun.

Fig. 6. Example results on the test set of MSVD.

7. Conclusion

A hierarchical attention-based multi-modal fusion video captioning model is proposed, which manages different modalities with progressive attention layers. In addition, effective approaches for temporal and semantic features extraction are also investigated in the work. The proposed method achieves competitive performance compared with the related video captioning methods.

Acknowledgment

This work is supported by the grants from the National Natural Science Foundation of China (61532018, 61673396, 61671482), the Fundamental Research Funds for the Central Universities (17CX02041A, 18CX02136A, 18CX06045A).

References

[1] C. Xiong, J. Lu, D. Parikh, R. Socher., Knowing when to look: adaptive attention via a visual sentinel for image captioning, arXiv preprint 2016 arXiv:1612.01887.

[2] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo., Image captioning with semantic attention., in: Proceedings of the CVPR, 2016.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio., Show, attend and tell: Neural image caption generation with visual attention., Proceedings of the ICML, 2015.

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang., Bottom-up and top-down attention for image captioning and visual question answering., arXiv preprint arXiv:1707.07998v2.

[5] Q. Jin, J. Chen, S. Chen, Y. Xiong, A. Hauptmann., Describing videos using multi-modal fusion., ACM Multim. Conf. (2016) 1087–1091.

[6] X. Long, C. Gan, G. de Melo., Video captioning with multi-faceted attention., arXiv preprint arXiv:1701.03126.

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

[8] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, M. Paluri., Learning spatiotemporal features with 3d convolutional networks., Proceedings of the ICCV, 2015.

[9] Q. Jin, J. Liang, X. Lin., Generating natural video descriptions via multimodal processing., Proceedings of the Interspeech, 2016.

[10] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, H.T. Shen., Hierarchical lstm with adjusted temporal attention for video captioning., arXiv preprint arXiv:1706.01231.

[11] V. Ramanishka, A. Das, H.P. Dong, S. Venugopalan, L.A. Hendricks, M. Rohrbach,

K. Saenko, Multimodal video description, in: ACM on Multimedia Conference, 2016, pp. 1092–1096.

[12] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R.J. Mooney, K. Saenko., Translating videos to natural language using deep recurrent neural networks., Proceedings of the NAACLHLT, 2015.

[13] P. Pan, Z. Xu, Y. Yang, F. Wu, Y. Zhuang., Hierarchical recurrent neural encoder for video representation with application to captioning., arXiv preprint arXiv:1511.03476.

[14] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko., Sequence to sequence-video to text., Proceedings of the ICCV, pages 4534C4542, 2015.

[15] Y. Pan, T. Mei, T. Yao, H. Li, Y. Rui., Jointly modeling embedding and translation to bridge video and language., CoRR, abs/1505.01861 (2015).

[16] H. Yu, J. Wang, et al., Video paragraph captioning using hierarchical recurrent neural networks., Proceedings of the CVPR, 2016.

[17] K. He, X. Zhang, S. Ren, J. Sun., Deep residual learning for image recognition., Proceedings of the CVPR, 2016.

[18] H. Fang, S. Gupta, et al., From captions to visual concepts and back., Proceedings of the CVPR, 2015.

[19] X. Chen, C.L. Zitnick., Learning a recurrent visual representation for image caption generation., arXiv preprint arXiv:1411.5654.

[20] Y. Bengio, P. Simard, P. Frasconi., Learning long-term dependencies with gradient descent is difficult., IEEE Trans. Neural Netw. 5 (2) (1994).

[21] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei., Fully convolutional instance-aware semantic segmentation., Proceedings of VPR, 2017.

[22] O. Maron, T. Lozano-Perez., A framework for multiple-instance learning, Proceedings of the NIPS, 1998.

[23] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo., Deep multimodal learning: a survey on recent advances and trends., Proceedings of the CVPR 2016.

[24] W. Xu, M. Yang, K. Yu., 3d convolutional neural networks for human action recognition., IEEE Trans. Pattern Anal. Mach. Intel. 35 (1) (2013) 221–231.

[25] J. Xu, T. Mei, T. Yao, Y. Rui., Msrvtt: A large video description dataset for bridging video and language., Proceedings of the CVPR, 2016.

[26] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, Semantic compositional networks for visual captioning, Proceedings of the CVPR, 2017.

[27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu., Bleu: a method for automatic evaluation of machine translation., Proceedings of the ACL, 2002.

[28] M. Denkowski, A. Lavie., Meteor universal: Language specific translation evaluation for any target language., Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014.

[29] R. Vedantam, C.L. Zitnick, D. Parikh., Cider: Consensus-based image description evaluation., Proceedings of the CVPR, 2015.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick., Microsoft coco:common objects in context., Proceedings of the ECCV, 2014.

[31] M.D. Zeiler., Adadelta: an adaptive learning rate method, CoRR (2012). Abs/1212.5701.

[32] L. Yao, A. Torabi, K. Cho, N. Ballas, C.J. Pal, H. Larochelle, A.C. Courville., Describing videos by exploiting temporal structure., Proceedings of the ICCV, 2015.

[33] N. Neverova, C. Wolf, G. Taylor, F. Nebout, Moddrop: Adaptive multi-modal gesture recognition, IEEE Trans. Pattern Anal. Mach. Intel. 38 (8) (2016) 1692–1706.

Chunlei Wu is a male associate professor in the college of computer and communication, China University of Petroleum (East China). He received the Ph.D. degree majoring in computer application technology from Ocean University of China in 2014. His current interests include image and video processing, and machine learning. He has authored and coauthored more than 30 journal and conference papers and textbooks.

YiWei Wei is a postgraduate student in college of computer and communication engineering, Chi na University of Petroleum. His current research interests include cross modal retrieval and neura l machine translation.

Xiaoliang Chu is a postgraduate in college of computer and communication engineering, China University of Petroleum. His current research interests include image caption, visual question answering and social media detection.

Weichen Sun is a postdoctoral researcher in First Researcher Institute of the Ministry of Public Security of PRC. He received the PH. D. degree majoring in Communication and Information System from Beijing University of Posts and Telecommunications in 2017. His current research interests include deep learning, image classification and object detection.

Fei Su is a female professor in the multimedia communication and pattern recognition lab, school of information and telecommunication, Beijing university of posts and telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.

Leiquan Wang received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT. Now he is a lecturer in college of computer and communication engineering, China University of Petroleum. His current research interests include multimodal fusion, cross modal retrieval and image/video caption.