Pattern Recognition 136 (2023) 109204

Contents lists available at ScienceDirect

Pattern Recognition

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

Relation-aware attention for video captioning via graph learning

Yunbin Tu[a,b], Chang Zhou[c], Junjun Guo[a,b], Huafeng Li[a,b], Shengxiang Gao[a,b], Zhengtao Yu[a,b,*]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Yunnan, Kunming, 650500, P.R. China
[b] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Yunnan, Kunming, 650500, P.R. China
[c] Department of information science and technology, Tsinghua Shenzhen International Graduate School, Guangdong, Shenzhen, 518000, P.R. China

ARTICLE INFO

Article history:
Received 16 September 2020
Revised 14 November 2022
Accepted 24 November 2022
Available online 25 November 2022

Keywords:
Video captioning
Relation-aware attention
Graph learning

ABSTRACT

Video captioning often uses an attentive encoder-decoder as the baseline model. However, the conventional attention mechanism still remains two problems. First, the attended visual feature is often irrelevant to the target word state, because the attention process only uses the unidirectional flow from vision to linguistics, while lacking the reverse flow. Second, each attention result is independent, because it is computed only based on the previous word states while not considering the attention information from the past and future. This does not suit the attention habits of human beings. In this paper, we improve the conventional attention mechanism to a relation-aware attention mechanism. To this end, we propose two kinds of graph learning strategies, namely the linguistics-to-vision heterogeneous graph (HTG) and the vision-to-vision homogeneous graph (HMG). The HTG aims to enhance the inter-relation of attention by reversely modeling the relation of each word with respect to every attended visual feature, supporting proper semantic alignment in between. The HMG aims to enhance the intra-relation of attention by capturing the relations among all of the attended visual features, which can leverage the attention information from the past and future to guide the current attention process. Extensive experiments on two public datasets show that our proposed method not only significantly improves the baseline model, but also outperforms state-of-the-art methods.

1. Introduction

Automatically describing a video with a natural language sentence has been flourishing, because it connects compute vision and natural language processing, which are two important applications in pattern recognition. Video captioning has many practical applications such as video title automatic generation and content-based video retrieval.

Recently, the attentive encoder-decoder framework has been widely used in image or video captioning methods [1–4]. Specifically, in the task of video captioning, a pre-trained CNN model is used to encode a sequence of frames into a sequence of visual features, which are then decoded into a sequence of word states by an RNN model or its variant LSTM [5]. At each decoding step, the attention mechanism is utilized to focus on a subset of key visual features to generate the target word state.

Despite the progress, there are two limitations for the conventional attention mechanism. First, it only exploits unidirectional flow from the attended visual feature to the target word sate, while the reverse flow is ignored. In this case, the attended visual feature is often not what the decoder really expects and thus is irrelevant to the target word state. For instance, in Fig. 1(a), although the attention mechanism wrongly recognizes the “cosmetic brush” as the “microphone”, this inaccurate attention result is still fed into the decoder that subsequently generates a totally irrelevant caption. Second, all of the attended visual features are only computed by the previously generated word states, while ignoring the previously attended results. As we see, in Fig. 1(b), the tiny objects “ping-pong” and “ping-pong bat” are both moving continually and quickly, so the previous attention results are central to the current attention process. Generally, when watching a video, our attentions are always attracted by moving and surprising objects along the temporal sequence [6,7]. But sometimes we might find that some important details in the previous snippets are ignored when the latter events appear, so we will be back to the previous frames to change our attention parts. This indicates that the current focus is potentially affected by the future attention results. Therefore, each attention result is not independent and should be closely relevant to the others.

* Corresponding author.
E-mail address: ztyu@hotmail.com (Z. Yu).

https://doi.org/10.1016/j.patcog.2022.109204
0031-3203/

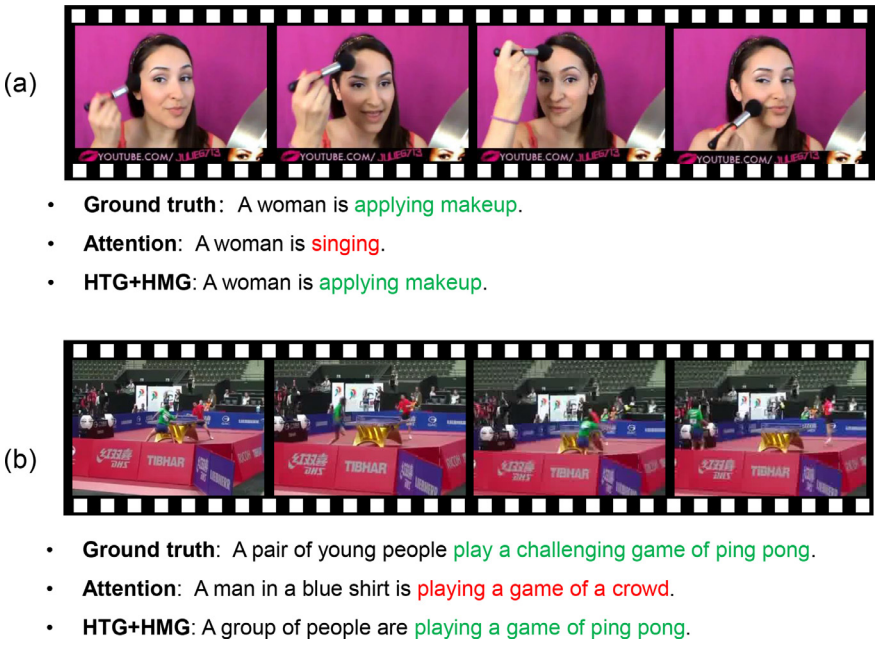

Fig. 1. Two examples of video captioning. (a) Conventional attention-based method (Attention) generates an irrelevant description. Our HTG+HMG correctly identifies the action of "applying makeup". (b) Attention generates an ambiguous description. Our HTG+HMG generates the informative words of "ping pong".

According to the above observations, in this paper, we propose a relation-aware attention model by enhancing the inter-relation and intra-relation of the conventional attention model for video captioning. To achieve this goal, we present a novel video captioning architecture which augments the conventional attentive encoder-decoder with a linguistics-to-vision heterogeneous graph (HTG) and a vision-to-vision homogeneous graph (HMG). Specifically, given a video, a pre-trained CNN model is first used to encode a sequence of frames into a sequence of visual features that are then fed into a bi-LSTM to model the temporal dependencies. Next, the attention-based LSTM decoder selectively attends to the key visual feature to generate the word state at each time step. Finally, the proposed two kinds of graphs are learned to enhance 1) the inter-relation between each attended visual feature and its corresponding word state, and 2) the intra-relation for each attended visual feature with respect to the others. On one hand, based on the bi-linear attention mechanism, the HTG is learned to reversely model the relations of each target word state with respect to every attended visual feature, thus enhancing the semantic alignment in between. On the other hand, we build the HMG which considers each attended visual feature as a homogeneous node. With the help of graph convolution network, the node representations receive the message from the neighbor nodes during the process of relation reasoning. Thus, each attended visual feature is computed based on not only previously generated words, but also the attention information from the past and future. Through two kinds of graph learning, the conventional attention mechanism can be improved to a relation-aware attention mechanism which models intra- and inter-relation during attention process. By jointly learning both the relations at the same time, the attention model is expected to not only attend to proper visual features the decoder really expects, but also suit the attention habits of human beings.

To summarize, the contributions of this work lie in three aspects:

- We improve the conventional attention mechanism to a relation-aware attention mechanism which aims to 1) support proper semantic alignment between target word states and attended visual features and 2) leverage the attention information from the past and future to guide the current attention process.
- A linguistics-to-vision heterogeneous graph (HTG) is learned to enhance the inter-relations between target word states and attended visual features. Moreover, a vision-to-vision homogeneous graph (HMG) is modeled to capture the intra-relations among all of the attended visual features.
- We incorporate the proposed two kinds of graphs (*i.e.*, HTG and HMG) and the gobal graph (HTG+HMG) into the attentive encoder-decoder framework to conduct extensive experiments on two well-known datasets, *i.e.*, MSVD [8] and MSR-VTT [9]. The experimental results indicate that both the intra- and inter-relation of the attention mechanism are well enhanced, and significant improvements in video captioning are achieved.

2. Related work

In this section, we will first briefly review the previous works for video captioning. Then, we will introduce the captioning works focusing on enhancing the relation for attention. Finally, we introduce the use of graph learning in captioning.

2.1. Video captioning

Recently, video captioning has draw extensive attentions in the community of multi-modal learning. Previous works in this task could be categorized into three dimensions, that is, (1) template-based methods, (2) CNN and LSTM-based methods and (3) CNN and Transformer-based methods.

Template-based Methods. In the template-based methods [10,11], a general pipeline is to first predict a number of visual concepts (*e.g.*, objects, relationships, and attributes) via different classification approaches. Then, these concepts constitute a caption according to a pre-defined sentence template and the basic grammar (*e.g.*, subjects-verbs-objects). Although this kind of approach is intuitive, it cannot generate flexible and meaningful captions due to the processing of complex data and the limitation of pre-defined templates.

CNN and LSTM-based Methods. Recently, with the rapid development of deep learning, the encoder-decoder framework using CNN and LSTM has been widely used in this task [12,13]. Venugopalan *et al.* [14] first introduced this framework for video captioning. In their work, they first exploited a pre-trained CNN model to extract features from a sequence of frames. Then, they computed a mean pooling representation over all features and fed it into an LSTM decoder to generate a caption. However, simply averaging all the features often results in the confused representation of video content. To address this problem, Yao *et al.* [15] introduced the soft-attention mechanism [16] to focus on the key visual feature for generating the target word. After that, this attentive encoder-decoder has become a popular framework for video captioning.

CNN and Transformer-based Methods. Recently, Transformer-based methods have achieved state-of-the-art performances in various multi-model tasks [17]. Inspired by it, Chen *et al.* [18] has introduced this framework [19] for video captioning. In their work, a sequence of video frames were first fed into both 2-D and 3-D CNNs which then outputted corresponding appearance and motion features, respectively. Then, instead of using an LSTM, they attempted to utilize vanilla Transformer for sequential representation and devise two types of fusion blocks in decoder layers for combining different modalities effectively. Furthermore, Pan *et al.* [20] first extracted appearance, motion and object features by the 2-D CNN, 3-D CNN and region-based CNN (R-CNN), respectively. Then, they fed these features into a spatio-temporal graph to model their correlations. Finally, language decoding is performed through a Vanilla Transformer. Their works show that the use of Transformer will be a new direction for video captioning.

2.2. Enhancing the relation for attention

Enhancing the Inter-relation. The conventional attention mechanism often fails to attend to the relevant visual feature to generate the target word state. Prior researchers attempted to address this problem with various methods which can be classified into two dimensions. One is to devise complex attention models. The other is to exploit a memory module to bridge the gap between attended visual features and target word states. Specifically, on one hand, Hori *et al.* [21] proposed an attention-based fusion network to adaptively attend to different kinds of visual information to generate a caption, that is, nouns are generated based on appearance features and motion features are utilized to generate verbs. Yan *et al.* [22] proposed a spatial-temporal attention model which exploited both the spatial and temporal structures in a video, so each target word is generated based on not only the key frame, but also the key object or region in that frame. Zhao *et al.* [23] proposed a co-attention model which is composed of a visual attention module, a text attention module, and a balancing gate. During the generation procedure, the visual attention module is able to adaptively attend to the salient regions in each frame and the frames that are most correlated with the caption. The text attention module can automatically focus on the most relevant previously generated words or phrases. Gao *et al.* [2] devised an adaptive attention model that makes the decoder adaptively select the visual information or the language context information to generate the visual words or the non-visual words. Long *et al.* [24] proposed a multi-faceted attention network to flexibly attend to the relevant frames, regions, and semantic attributes to generate the target words. Tu *et al.* [1] proposed a textual-temporal attention model, where the textual attention model first selects key visual tags based on language context, because both of them belong to textual modality. Then, the temporal attention model attends to the key visual features under the guidance of the visual tags. Ryu *et al.* [25] proposed to align frames with the phrases of partially decoded caption, thus using the semantic groups as information unit to caption a video. On the other hand, Wang *et al.* [26] proposed a multi-modal memory model to learn the long-term visual-textual dependency and further guide attention process. However, all of these methods have a common feature that they only utilized unidirectional flow from vision to linguistics while not considering the reverse flow. On the contrary, our proposed attentive encoder-decoder with the HTG can additionally include the reverse flow by reversely modeling the relations of each target word state with respect to every attended visual feature.

Enhancing the intra-relation. To the best of our knowledge, previous works in video captioning have not considered modeling the intra-relations among all of the attention results. But in image captioning, Qin *et al.* [27] have built the relations between adjacent attention results. In their work, they concatenated the previous attention result and the current hidden state of the LSTM decoder to calculate the current attention result, which is able to embed visual information of the past. Different from their work, we consider using a homogeneous graph to capture the intra-relations among all of the attention results, which is able to obtain the attention information from the past and future during the current attention process. Through this manner, the attention results will be relevant to the others and the attention process will suit the attention habits of human beings.

2.3. Graph learning in captioning

Using graph learning to explore the visual relations among the objects in an image has been widely used in image captioning. For instance, Wang *et al.* [28] leveraged a graph neural network to implicitly model the visual relations between objects or regions in an image. Inspired by the related works in image captioning, the idea of graphical representation has been attracting attentions in video captioning. Pan *et al.* [20] first proposed a spatio-temporal graph network to exploit object interactions in a video and then performed the graph convolution network to update the graph representation. Zhang *et al.* [29] captured detailed temporal dynamics for the salient objects in a video via a bidirectional temporal graph, and learned discriminative spatio-temporal video representations by performing object-aware local feature aggregation on object regions. However, all of their methods only 1) exploited graph learning in the encoder stage and 2) learned the graph based on the homogeneous visual features. Different from them, our graph learning 1) is performed in the attention process and 2) learned the graphs based on not only the homogeneous visual features, but also the heterogeneous features between linguistic word states and visual features.

2.4. Summary

Our work aims to devise a relation-aware attention mechanism by enhancing the intra- and inter-relation for the conventional attention mechanism in video captioning. For enhancing the inter-relation, different from [1,2,21,24–26], which both used unidirectional flow from attended visual features to target word states, we additionally include reverse flow by utilizing a heterogeneous graph to reversely model the inter-relation of each word state with regard to every attended visual features. In terms of enhancing the intra-relation, to the best of our knowledge, we are the first work in video captioning to model the relations among all of the attention results by using a homogeneous graph.

3. Methodology

Our overall captioning framework is shown in Fig. 2, in which the proposed two kinds of graphs, namely linguistics-to-vision

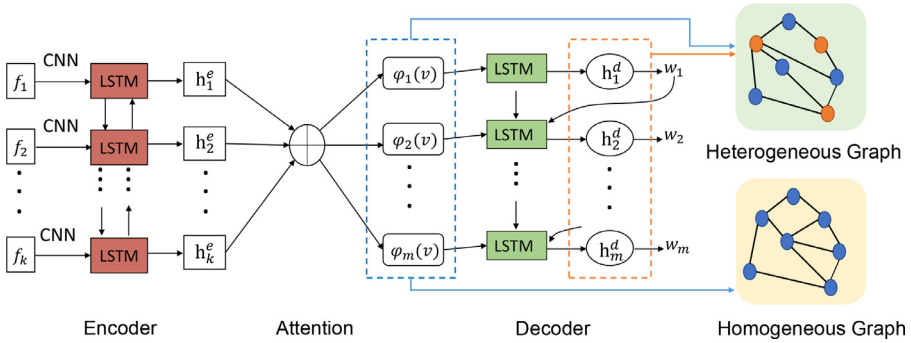

Fig. 2. The overall framework of our proposed method. We couple a linguistics-to-vision heterogeneous graph and a vision-to-vision homogeneous graph with an attentive encoder-decoder. The details of two graphs are shown in Figs. 3 and 4, respectively.

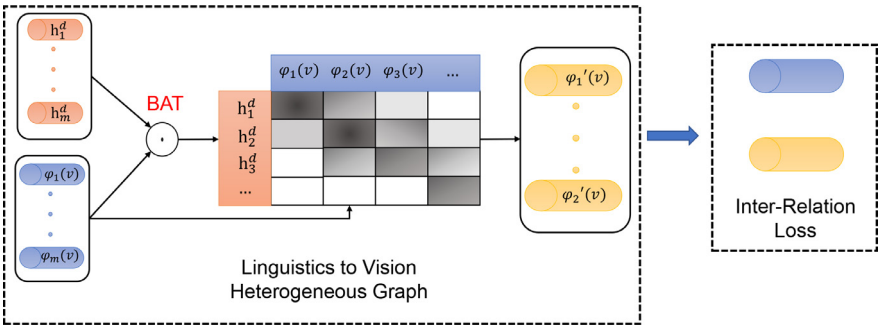

Fig. 3. The architectures of the linguistics-to-vision heterogeneous graph (HTG) and inter-relation loss (better viewed in color), where BAT means the bilinear attention mechanism.

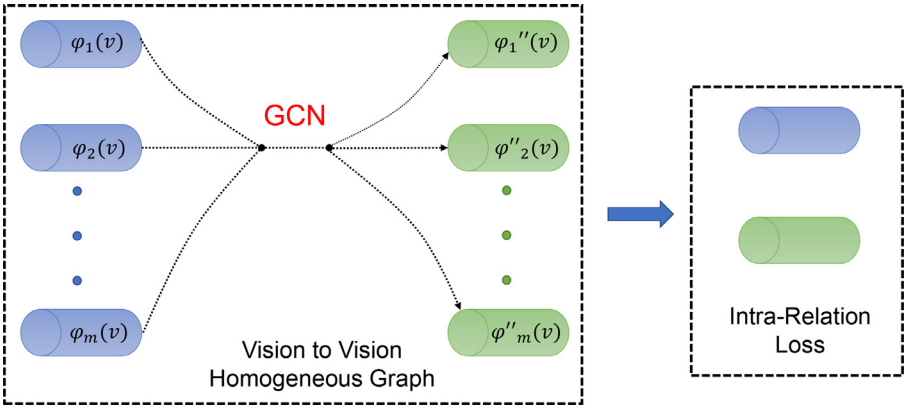

Fig. 4. The architectures of the vision-to-vision homogeneous graph (HMG) and the intra-relation loss (better viewed in color), where GCN means the graph convolution network.

heterogeneous graph (HTG) and vision-to-vision homogeneous graph (HMG), are coupled with the conventional attentive encoder-decoder. In this section, we first briefly review the attentive encoder-decoder in Section 3.1 and then elaborate the HTG and HMG in Section 3.2.

3.1. Attentive encoder-decoder

As illustrated in Fig. 2, the attentive encoder-decoder consists of a bidirectional LSTM as the encoder and a unidirectional LSTM as the decoder, which is similar to [30]. Hence, we first breily review the basic structure of an LSTM, which is defined as:

$$\begin{aligned} i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i), \\ f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f), \\ o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o), \\ g_t &= tanh(W_f h_{t-1} + U_f x_t + b_g), \\ c_t &= c_{t-1} \odot f_t + i_t \odot g_t, \\ h_t &= o_t \odot tanh(c_t), \end{aligned} \tag{1}$$

where x_t is the input of the LSTM and h_t is the output of the LSTM. c_t is the memory state of the LSTM. σ is a *sigmoid* activation function and $\odot$ refers to element-wise multiplication. W_*, U_*, and b_* are the parameters to be learned. For simplicity, we refer to the operation procedure of an LSTM encoder and an LSTM decoder with the following notations, respectively:

$$h_i^e = LSTM^e(h_{i-1}^e, x_i), \quad h_t^d = LSTM^d(h_{t-1}^d, x_i'), \tag{2}$$

where x_i and x_i' represent different inputs to the LSTM encoder and LSTM decoder, respectively.

3.1.1. Encoder

A given video is first uniformly sampled as a sequence of video frames $F = \{f_1, \ldots, f_k\}$. Then, a pre-trained CNN model is utilized to extract visual features one by one from this sequence, which is denoted as $v = \{v_1, v_2, \ldots, v_k\}$. Finally, to keep the temporal relationships between adjacent visual features, we use a bidirectional LSTM (BiLSTM) to encode these visual features.

A BiLSTM consists of forward and backward of an LSTM's. First, the forward of it $\overrightarrow{f}$ reads the input sequence of visual features $v = \{v_1, v_2, \ldots, v_k\}$ as it is ordered (from v_1 to v_k) and calculates a sequence of forward hidden states $\left(\overrightarrow{h}_1, \cdots, \overrightarrow{h}_k\right)$. The backward of it $\overleftarrow{f}$ reads the sequence in the reverse order (from v_k to v_1), resulting in a sequence of backward hidden states $\left(\overleftarrow{h}_1, \cdots, \overleftarrow{h}_k\right)$. Then, we obtain an encoded visual feature by concatenating the forward hidden state $\overrightarrow{h}_i$ and the backward one $\overleftarrow{h}_i$, *i.e.*, $h_i^e = \left[\overrightarrow{h}_i, \overleftarrow{h}_i\right]$. Finally, a sequence of visual features $v = \{v_1, v_2, \ldots, v_k\}$ are encoded as $h^e = \{h_1^e, h_2^e, \ldots, h_k^e\}$.

3.1.2. Attention-based decoder

Video captioning aims to translate from k encoded visual features $\{h_1^e, h_2^e, \ldots, h_k^e\}$ into the word sequence $\{w_1, w_2, \ldots, w_m\}$ of length m. The distribution of the output word sequence w.r.t the input feature sequence is:

$$p\left(w_1, \ldots, w_m | h_1^e, \ldots, h_k^e\right) = \prod_{t=1}^{m} p\left(w_t | h_t^d\right), \tag{3}$$

where h_t^d is the hidden state of the LSTM decoder at the time step t, which is calculated by the previous hidden state h_{t-1}^d, previously generated word embedding $E[w_{t-1}]$ and the currently attended visual feature $\varphi_t(v)$:

$$h_t^d = \mathrm{LSTM}^d\left([E[w_{t-1}], \varphi_t(v)], h_{t-1}^d\right), \tag{4}$$

where [,] stands for tensor concatenation. In our work, the attended visual feature $\varphi_t(v)$ at time step t is calculated via the conventional attention mechanism:

$$\varphi_t(v) = \sum_{i=1}^{k} \alpha_i^{(t)} h_i^e. \tag{5}$$

The attention weight $\alpha_i^{(t)}$ reflects the relevance of the i-th encoded visual feature given the previous hidden state h_{t-1}^d which summarizes the information of the previously generated words, and is computed as:

$$\alpha_i^{(t)} = softmax(w \tanh\left(W_a h_{t-1}^d + U_a h_i^e + b_a\right)), \tag{6}$$

where w, W_a, U_a, and b_a are learned parameters.

3.2. The proposed graph learning

In this section, we will elaborate the proposed linguistics-to-vision heterogeneous graph (HTG) and vision-to-vision homogeneous graph (HMG). The architectures of two kinds of graphs are illustrated in Figs. 3 and 4.

3.2.1. Linguistics-to-vision heterogeneous graph (HTG)

The architecture of the HTG is shown in Fig. 3, which mainly consists of a bilinear attention mechanism (BAT) and an inter-relation loss. The HTG aims to enhance the inter-relations between attended visual features and corresponding target word states by reversely modeling the relations in between. Specifically, given the attended visual features set $\varphi(v) = \left\{\varphi_j(v)\right\}_{j=1}^{m}$ and the generated word state (*i.e.*, hidden state of the LSTM decoder) set $\mathcal{H} = \{h_r\}_{r=1}^{m}$, where $\varphi_j(v) \in \mathbb{R}^{d_v}$ and $h_r \in \mathbb{R}^{d_h}$, we seek to construct a linguistics-to-vision heterogeneous graph $G_T = \{\varphi(v), \mathcal{H}, \zeta, A\}$. ζ are a set of graph edges to learn and they indicate whether each word state is relevant or not relevant to one of attended visual features. $A \in \mathbb{R}^{m \times m}$ is the corresponding heterogeneous adjacency weighted matrix which represents the relation weights of each word state with respect to every attended visual feature. In order to compute A, we utilize BAT to calculate the accumulative weights of word state nodes with respect to attended visual feature nodes, which is formulated as:

$$A_{rj} = \frac{\exp\left(A'_{rj}\right)}{\sum_{rj} \exp\left(A'_{rj}\right)}, \quad A'_{rj} = h_r W \varphi_j(v)^T, \tag{7}$$

where $A_{rj} \in A$ is a scalar to indicate the relation between h_r and $\varphi_j(v)$. $W \in \mathbb{R}^{d_h \times d_v}$ is a trainable weighted matrix. The $A \in \mathbb{R}^{m \times m}$ is normalized by using a $softmax$ at each location. After the heterogeneous adjacency weighted matrix is computed, we will obtain the reversely attended visual features by:

$$\varphi'(V) = A\varphi(v), \tag{8}$$

In this way, different heterogeneous node representation can adaptively propagate to each other. Once the reversely attended feature $\varphi'(v) = \left\{\varphi'_j(v)\right\}_{j=1}^{m}$ are yielded, the inter-relation loss is defined as follows:

$$\mathcal{L}_T = \sum_{j=1}^{m} \left\|\varphi_j(v) - \varphi'_j(v)\right\|_2^2. \tag{9}$$

By minimizing the defined inter-relation loss, we expect each attended visual feature is more relevant to its corresponding target word state.

3.2.2. Vision-to-vision homogeneous graph (HMG)

The architecture of the HMG is shown in Fig. 4. It is built based on the graph convolution network (GCN) [31] and is utilized to capture the intra-relations among all of the attention results. To be more specific, we seek to construct a vision-to-vision homogeneous graph $G_M = \{\varphi(v), \xi, B\}$, where each node $\varphi_j(v) \in \varphi(v)$ corresponds to an attended visual feature and $\varphi_j(v) \in \mathbb{R}^{d_v}$. ξ are a set of graph edges which indicate whether each attended visual feature is relevant or not relevant to the others. $B \in \mathbb{R}^{m \times m}$ is the corresponding homogeneous adjacency matrix and can be seen as how much information each attention result obtains from the past and future attention results. We define the adjacency matrix for the homogeneous graph as follows:

$$B = \mathrm{softmax}_r\left(\varphi(v)\varphi(v)^T\right) + I_d, \tag{10}$$

where I_d indicates the identity matrix and $softmax_r$ indicates we make $softmax$ operation across the row direction. Finally, We apply the GCN to update this graph, so the independently attended visual features $\varphi(v)$ are updated to relatedly attended visual features $\varphi''(v)$:

$$\begin{aligned} M &= B\varphi(V), \\ \varphi''(V) &= \tanh\left(w_f * M + b_f\right) \odot \sigma\left(w_c * M + b_c\right), \end{aligned} \tag{11}$$

where $w_f \in \mathbb{R}^{1 \times d_v \times d_v}$, $w_c \in \mathbb{R}^{1 \times d_v \times d_v}$, $b_f \in \mathbb{R}^{d_v}$, and $b_c \in \mathbb{R}^{d_v}$ are the trainable parameters. '$*$' indicates the convolutional operation. '$\odot$' indicates element-wise product. σ indicates the *sigmoid* non-linear activate function. Each row of the matrix M represents a node's feature vector, which integrates the information of its neighboring node features. $\varphi''(v) \in \mathbb{R}^{m \times d_v}$ indicates the output of the GCN. Finally, the intra-relevance loss is defined as follows:

$$\mathcal{L}_M = \sum_{j=1}^{m} \left\|\varphi_j(v) - \varphi''_j(v)\right\|_2^2. \tag{12}$$

Through minimizing this loss, we encourage current attention process to obtain more useful attention information from the past and future.

3.3. Training procedure

The training procedure of our proposed method consists of three stages. The first stage is to train the attentive encoder-decoder by optimizing the negative log-likelihood:

$$\mathcal{L}_{\mathrm{NLL}} = -\sum_{t=1}^{m} \log p\left(w_t | h_t^d\right), \tag{13}$$

where w_t and h_t^d are the target word and the hidden state of the LSTM decoder at time step t, respectively. m is the total length of a video caption.

In the second stage, after the attentive encoder-decoder converges, we respectively train each single graph with pre-trained attentive encoder-decoder. One is the attentive encoder-decoder with the linguistics-to-vision heterogeneous graph (HTG), which objective function is defined as:

$$\mathcal{L}_1 = \mathcal{L}_{\mathrm{NLL}} + \lambda_T \mathcal{L}_T. \tag{14}$$

Another is the attentive encoder-decoder with the vision-to-vision homogeneous graph (HMG), which objective function is defined as:

$$\mathcal{L}_2 = \mathcal{L}_{\mathrm{NLL}} + \lambda_M \mathcal{L}_M. \tag{15}$$

Here, λ_T and λ_M are trade-off parameters to balance the contributions from the each graph and the attentive encoder-decoder, which are discussed in Section 4.4. In the third stage, these trade-off parameters are also used to train the attentive encoder-decoder with the combination of these two graphs, *i.e.,* global graph, which objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \mathcal{L}_1 + \mathcal{L}_2. \tag{16}$$

4. Experiments

4.1. Datasets and evaluation metrics

The Microsoft Video Description Corpus (MSVD). MSVD [8] has 1970 video clips. Each video clip is provided with about 41 human annotated sentences. Following [15], the dataset can be divided into a training set of 1200 video clips, a validation set of 100 clips, and a test set consisting of the rest of 670 clips.

MSR Video to Text (MSR-VTT). MSR-VTT [9] consists of 10,000 video clips from 20 general categories. Each video clip is provided with 20 human annotated sentences. We use the official spilt with 6513 videos for training, 497 for validation and 2990 for testing.

Evaluation Metrics. We use four standard metrics to evaluate the quality of generated captions, *i.e.*, BLEU-4 [32], METEOR [33], ROUGE-L [34] and CIDEr [35]. BLEU-4 has used for corpus level comparisons over which 4-gram matches exist. METEOR can generate an alignment according to exact token matching to judge the word correlation between candidate and reference sentences. ROUGE-L uses a measure based on the Longest Common Subsequence (LCS), which is a set words shared by two sentences which occur in the same order. CIDEr is especially designed for the captioning task to capture human judgment of consensus. We obtain all the results in this paper based on the Microsoft COCO evaluation server [36].

4.2. Implementation details

For the ground truths on the both datasets, we remove the punctuations, split them with blank space and convert all words into lowercase. The maximum vocabulary size is set to 13,010 on MSVD and 23,000 on MSR-VTT. Especially, on the both datasets, we has added four special token signs, *i.e.*, $< UNK >$, $< PAD >$, $< START >$, and $< STOP >$, to indicate "unknown words", "keeping the length of all sentences the same", "start of sentence", and "end of sentence", respectively. When generating the first word, we need a zeroth word w_0 (*i.e.*, $< START >$) to indicate "start of sentence". For the last word w_m, we denote it as $< STOP >$. This is necessary because the model needs to know when to stop decoding during inference. The maximum length of sentences is set to 30 on the both datasets. Each word in the sentence is represented as one-hot vector (1-of-g coding), where g denotes the number of words in the vocabulary. To reduce parameter size, we project the one-hot representations into a low dimensional word vector space with a word embedding matrix E, which size is set to 512 on the both datasets.

In terms of visual feature extraction, we sample 28 equally-spaced frames for each video and feed them into a pre-trained Inception-V4 [37] to extract visual features, so the dimension of each feature is 1536. The hidden state size of both LSTM encoder and LSTM decoder are set to 1024 on the both datasets.

Model Training. In the training phase, on the both datasets, we use Adam optimizer [38] and set the mini-batch size as 64. On MSVD, in the first stage, we set the learning rate as 2×10^{-4}. In the second and third stages, we set the learning rate as 5×10^{-5}. On MSR-VTT, in the first stage, we set the learning rate as 1×10^{-4}. We set the learning rate as 1×10^{-5} and 9×10^{-6} in the second and third stages. Besides, we set dropout regularization in the rate of 0.4 in all layers and cip graients element wise as 10.0. Since CIDEr [35] is a consensus-based evaluation metric and especially designed for captioning task, the highest score of CIDEr on the validation set is used as a metric to choose the best model for testing.

Inference. During inference, the ground truths are not provided, we need to input the start sign word $< START >$ to start decoding and feed the previously generated word to the LSTM decoder at each time step until the end sign word $< STOP >$ is reached. At each time step, the straightforward option would be to choose the word with the maximum score after $softmax$ and use it to predict the next word. But this is not optimal because the rest of the sequence hinges on previous word. If that choice isn't the best, everything that follows is sub-optimal. Therefore, we use beam search with size 5 to choose the sequence that has the highest overall score in 5 candidate sequences. Both training and inference are implemented with PyTorch on an RTX 2080 Ti GPU.

Table 1
Ablation studies on MSVD. All values are reported as percentage (%).

	BLEU-4	METEOR	ROUGE-L	CIDEr
Base	47.5	33.3	69.5	77.1
HTG	**49.2**	34.2	70.8	83.4
HMG	49.0	**34.3**	70.9	83.8
HTG+HMG	**49.2**	**34.3**	**71.0**	**84.0**

4.3. Ablation studies

In order to verify the effectiveness of the HTG, HMG and the global graph (HTG+HMG), we set the following comparative experiments: (1) only using an attentive encoder-decoder; (2) the attentive encoder-decoder with the HTG; (3) the attentive encoder-decoder with the HMG; (4) the attentive encoder-decoder with the HTG+HMG. For convenience, we respectively denote them as Base, HTG, HMG, and HTG+HMG hereinafter. The trade-off parameters for λ_T and λ_M defined in Eqs. (14) and (15) are set to 0.2 and 0.4 on MSVD; 0.3 and 0.1 on MSR-VTT, which are analysed in the next section. The experimental results are shown in Tables 1 and 2.

From Tables 1 and 2, on the both datasets, we observe that: 1) each single graph and their combination both significantly improve the Base model for all metrics; 2) similar scores are obtained

Table 2
Ablation studies on MSR-VTT. All values are reported as percentage (%).

	BLEU-4	METEOR	ROUGE-L	CIDEr
Base	38.4	27.3	59.3	44.5
HTG	**39.1**	27.4	**59.6**	45.3
HMG	**39.1**	27.4	**59.6**	45.3
HTG+HMG	**39.1**	**27.5**	**59.6**	**45.4**

Table 3
Study on the Trade-off Parameters λ_T and λ_M on MSVD and MSR-VTT. All values are reported as percentage (%).

MSVD				MSR-VTT			
HTG		HMG		HTG		HMG	
λ_T	CIDEr	λ_M	CIDEr	λ_T	CIDEr	λ_M	CIDEr
0.0	77.14	0.0	77.14	0.0	44.50	0.0	44.50
0.01	83.26	0.01	83.29	0.01	45.21	0.01	45.21
0.1	83.41	0.1	83.26	0.1	45.24	0.1	**45.25**
0.2	**83.43**	0.2	83.56	0.2	45.26	0.2	45.23
0.3	83.39	0.3	83.63	0.3	**45.27**	0.3	45.17
0.4	83.36	0.4	**83.76**	0.4	45.26	0.4	45.17
0.5	83.17	0.5	83.52	0.5	45.26	0.5	45.19

for the single HTG and the single HMG, but the best performances are achieved when coupling both the HTG and HMG with the Base model. It shows that: 1) it is effective that we improve the conventional attention mechanism to a relation-aware attention mechanism by learning a linguistics-to-vision heterogeneous graph and a vision-to-vision homogeneous graph during the attention process; 2) although enhancing either inter-relation or intra-relation is useful for the conventional attention mechanism, the best proposal is to enhance the conventional attention mechanism from the viewpoint of mutual reinforcement between inter-relation and intra-relation. The reason is that by jointly learning both relations, the attend visual features will be more relevant to the target word states, and the attention mechanism is capable of using the past and future attention information to guide the current attention process. This suits the attention habits of human beings.

4.4. Study on the trade-off parameters λ_T and λ_M

In this section, we will discuss the effect of λ_T and λ_M, which are defined in Eqs. (14) and (15), respectively. λ_T is to balance the contributions from the proposed HTG and the attentive encoder-decoder, while λ_M is to balance the contributions from the proposed HMG and the attentive encoder-decoder. We will discuss these two parameters on MSVD and MSR-VTT, respectively.

For λ_T and λ_M, with different values on the both datasets, we obtain the CIDEr scores in Table 3. From both tables, We observe that: 1) adding the two kinds of relation losses ($\lambda_* > 0$) did improve the performances of the conventional attentive encoder-decoder; 2) as the values of λ_T and λ_M are increasing, the performances both decrease, for the whole model will focus too much on the graph learning part but ignore the supervision signal from ground truth; 3) the results with different trade-off parameters for HTG and HMG on the both datasets are very close, which indicates that the superior performance benefits from the proposed graph learning strategy, instead of heavily relying on the specific parameters. According to Table 3, we find 0.2 and 0.4 to be the perfect values for λ_T and λ_M on MSVD, respectively. On MSR-VTT, we find 0.3 and 0.1 to be the perfect values for λ_T and λ_M, respectively.

4.5. The study of model's sensitivity to video quality

Previous studies validate the effectiveness of the proposed graph learning strategy, and analyze the proper trade-off parameters for the two kinds of graphs. To explore the generalization ability of the HMG and HTG, we further study their sensitivity to the different video qualities based on the above trade-off parameters, that is, 0.2 and 0.4 for λ_T and λ_M on MSVD; 0.3 and 0.1 for λ_T and λ_M on MSR-VTT. To this end, we change the video quality by representing the content of each video with different visual modalities, *i.e.,* only appearance features extracted by ResNet-152 [39], only motion features extracted by C3D [40], and the context features (appearance+motion).

Table 4
The Study of Model's Sensitivity to Video Quality Representing by ResNet-152.

R152	MSVD				MSR-VTT			
Method	B-4	M	R	C	B-4	M	R	C
Base	45.2	33.3	69.0	74.7	38.3	27.4	59.1	44.1
HTG	47.9	34.2	69.7	76.2	38.7	27.7	59.5	45.4
HMG	**48.5**	34.3	69.9	76.4	38.8	27.7	59.5	**45.5**
HTG+HMG	48.4	**34.7**	**70.2**	**77.7**	**39.1**	**27.8**	**59.6**	**45.5**

Table 5
The Study of Model's Sensitivity to Video Quality Representing by C3D.

C3D	MSVD				MSR-VTT			
Method	B-4	M	R	C	B-4	M	R	C
Base	45.6	32.1	69.4	75.1	39.4	27.4	59.9	45.7
HTG	**47.6**	**32.4**	**69.9**	80.0	40.1	**27.7**	**60.2**	46.9
HMG	47.5	32.3	**69.9**	81.1	40.1	27.6	**60.2**	46.8
HTG+HMG	**47.6**	**32.4**	**69.9**	**81.7**	**40.2**	**27.7**	**60.2**	**47.0**

Table 6
The Study of model's Sensitivity to Video Quality Representing by ResNet-152 and C3D.

R152+C3D	MSVD				MSR-VTT			
Method	B-4	M	R	C	B-4	M	R	C
Base	50.6	34.1	71.6	86.7	40.7	27.9	60.7	46.7
HTG	51.6	35.1	**72.8**	89.0	40.7	28.2	60.8	47.6
HMG	51.3	35.0	72.7	90.1	40.9	28.1	60.8	47.8
HTG+HMG	**52.7**	**35.2**	**72.8**	**91.4**	**42.1**	**28.4**	**61.6**	**48.9**

The experimental results are shown in Tables 4, 5, and 6. We observe that with different video quality, the each single graph and global graph all achieve the consistently better results over the base model. And the best results are achieved by the global graph. The above observations validate that the proposed method has a good generalization ability, *i.e.,* 1) it is robust to the impact of video quality; 2) the trade-off parameters are generic. Besides, we find that the overall performance of appearance modality and motion modality is similar, but the best performance is achieved by using both modalities. This indicates that appearance and motion features are both important for representing video content. Appearance features represent frame information such as object and background scene, while motion features summarize the action information of consecutive frames, so each modality can supplement each other and form an omni-representation for video content.

4.6. The study of model's sensitivity to video length

Besides video quality, we further explore how sensitive the proposed global graph (HTG+HMG) is to the length of videos. To this end, we first sample 24, 26, 28, and 30 equally-spaced frames for each video, respectively. Then, these different length frames are represented by the concatenation of ResNet-152 and C3D. Fig. 5 illustrates the results of the proposed method on MSVD and MSR-VTT with different video length. Each sub-figure shows the scores of different length with the same metric and each color shows the scores of the same length with different metrics. We can observe that 1) the overall performance with different video length is

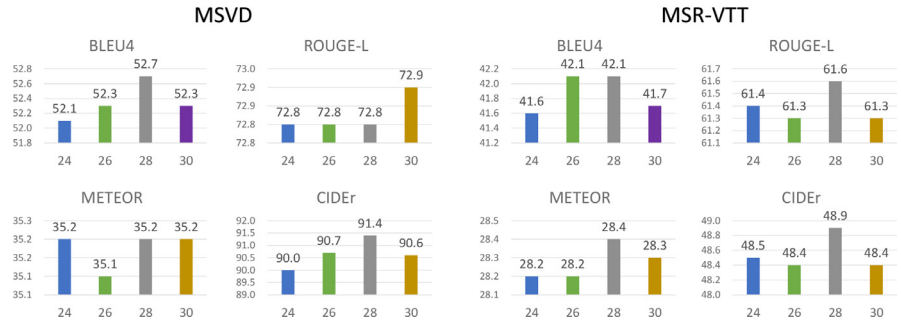

Fig. 5. The Study of model's sensitivity to different video length on MSVD and MSR-VTT, *i.e.*, 24, 26, 28, and 30. The video features are extracted from ResNet-152 and C3D.

Table 7
Comparison with the State-of-the-art Methods Enhancing the Inter-relation for Attention on MSR-VTT, where V, G, R152, FR, MR and C denote VGG, GoogleNet, ResNet-152, Faster R-CNN, Mask R-CNN, and C3D, respectively. The symbol "-" indicates such metric is unreported.

Method	BLEU-4	METEOR	ROUGE	CIDEr
AF [21](V+C) (2017)	39.4	25.7	-	40.4
MFATT [24] (R152+C) (2018)	39.1	26.7	-	-
CAM-RNN [23] (V) (2019)	37.7	26.7	58.5	38.3
hLSTMat [2] (R152+C) (2020)	38.7	26.8	-	41.9
STAT [22] (G+C+FR) (2020)	37.9	26.8	-	44.0
SGN [25](V) (2021)	37.8	27.0	58.3	41.9
TTA [1](R152+C+MR) (2021)	**41.4**	27.7	**61.1**	46.7
M^3 [26](V+C)(2018)	38.1	26.6	-	-
HTG (V)	37.8	27.0	58.5	42.4
HTG (R152+C)	40.7	**28.2**	60.8	**47.6**

Table 8
Comparison with the Methods Exploiting Graph Learning on MSR-VTT, where R101, R152, R200, FR, MR, I and C denote ResNet-101, ResNet-152, ResNet-200, Faster R-CNN, Mask R-CNN, I3D and C3D, respectively. B-4, M, R-L and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr. The symbol "-" indicates such metric is unreported.

Method	B-4	M	R-L	C
OA w/ ForTG [29] (R200+MR) (2019)	**40.8**	26.9	-	45.1
OA w/ BackTG [29] (R200+MR) (2019)	**40.8**	27.3	-	45.3
HTG (R152)	38.7	**27.7**	**59.5**	45.4
HMG (R152)	38.8	**27.7**	**59.5**	**45.5**
OA-BTG [29] (R200+MR) (2019)	41.4	28.2	-	46.9
STG (R101+I+FR) [20] (2020)	40.5	28.3	60.9	47.1
HTG+HMG (R152+C)	**42.1**	**28.4**	**61.6**	**48.9**

close; 2) the best performances are obtained on the both datasets when setting length as 28 for each video. This indicates that 1) HTG+HMG is robust to the different video length; 2) the video length should be proper, because when the length is shorter, some temporal information might be missing. When the length is longer, on the contrary, some redundant information might be noises to impact the representation of video content.

4.7. Comparison with the methods enhancing the inter-relation for attention

Since there is no any video captioning work exploring to enhance the intra-relation for attention according to our survey, we will compare our proposed HTG with the video captioning methods enhancing the inter-relation for attention. In Section 2.2, we have qualitatively made a comparative discussion between our proposed method and [1,2,21–26]. In this section, we will quantitatively compare our proposed HTG with them on MSR-VTT. For a fair comparison with these methods, we implemented the proposed HTG with encoding features of VGG, ResNet-152, and C3D. The comparison results are reported in Table 7.

From Table 7, when only using single visual feature (e.g., VGG), we can observe that the HTG outperforms or is on par with all the methods on the four metrics. When exploiting multiple features, our method still achieves best results on METEOR and CIDEr, especially improving the CIDEr score significantly, in particular with increases of 17.8% over AF, 13.6% over hLSTMat, and 8.2% over STAT, respectively. Compared with recent methods TTA additional using visual tags to enhance visual-textual alignment, our method does not rely on any external knowledge and also achieves comparative results on BLEU-4 and ROUGE-L. Compared to these methods, which either devised complex attention models or leveraged a memory model to guide the attention, our method is mainly based on the conventional attentive encoder-decoder, and additionally includes reverse inflow from target word states to attended visual features so as to enhance their inter-relation. The experimental results validate this motivation is intuitive and effective.

4.8. Comparing with the methods exploiting graph learning

Since our proposed HTG and HMG are both based on graph learning, we compared them with the two recent methods which exploited graph learning. In Section 2.3, we have qualitatively made a comparative discussion between our proposed method and them, *i.e.*, object-aware bidirectional temporal graph (OA-BTG) [29] and spatial-temporal graph (STG) [20]. In this section, we will quantitatively compare our proposed graphs with them on MSR-VTT. Especially, OA-BTG made ablation studies, where one is object-aware forward temporal graph (OA w/ ForTG) and the other is object-aware backward temporal graph (OA w/ BackTG). Hence, we also compare each proposed single graph with them.

The comparison results are reported in Table 8. We can see that 1) our proposed HTG and HMG are both better than OA w/ ForTG and OA w/ backTG on the three metrics; 2) compared to the full model of OA-BTG and STG, our proposed HMG+HTG also achieves highest scores on all the metrics, especially improving the CIDEr score significantly, in particular with an increase of 4.3% over OA-BTG and 3.8% over STG, respectively. As our introduced in Section 2.3, both OA-BTG and STG exploited graph learning in the encoding stage, while our proposed graph learning is performed to improve the attention process. As we know, the main challenge of video captioning is to align the target word states with encoded visual features due to the intrinsic gap between linguistics and vision. The experimental results validate that the proposed graph learning methods can effectively augment the attention model, so as to improve the quality of generated captions.

4.9. Comparing with other state-of-the-art methods

In this section, we compare our method with some newly published state-of-the-art methods. Their major approaches can be grouped to two categories: the CNN and Transformer-based

Table 9
Comparison with the State-of-the-art Methods on MSVD, where V, R-152, IRV2, C, IV4, and I denote VGG, ResNet-152, Inception-ResNet-V2, C3D, InecpctionV4 and I3D, respectively. The symbol "-" indicates such metric is unreported.

Method	B-4	M	R-L	C
Att-TVT [18] (R152+I) (2018)	53.0	34.7	71.7	80.8
TDConvED [41] (R152) (2019)	**53.3**	33.8	-	76.4
GRU-EVE$_{hft+sem}$ [13] (IRV2+C+YOLO) (2019)	47.9	35.0	71.5	78.1
LG-DenseLSTM [42](V+C) (2019)	50.4	32.9	69.9	72.6
RecNet$_{local}$ [43] (IV4) (2019)	52.3	34.1	69.8	80.3
SAAT [44] (IRV2+C) (2020)	46.5	33.5	69.4	81.0
BP-LSTMs [45] (R152) (2020)	42.9	32.0	68.3	62.2
STA-FG [46] (R152+C) (2020)	52.7	34.5	-	-
SHAN [47] (IRV2+I) (2022)	50.9	35.1	72.2	91.3
SMAN [48] (IRV2+C+F) (2022)	50.0	34.8	71.5	84.7
HTG+HMG (IV4)	49.2	34.3	71.0	84.0
HTG+HMG (R152)	48.4	34.7	70.2	77.7
HTG+HMG (R152+C)	52.7	**35.2**	**72.8**	**91.4**

Table 10
Comparison with the State-of-the-art Methods on MSR-VTT, where V, R152, IRV2, C, IV4, I, N, YL, and FR denote VGG-16, ResNet-152, Inception-ResNet-V2, C3D, InecpctionV4, I3D, NasNet, YOLO and Faster R-CNN, respectively. The symbol "-" indicates such metric is unreported.

Method	B-4	M	R-L	C
Att-TVT [18] (N+I) (2018)	40.1	27.9	59.6	47.7
TDConvED [41] (R152) (2019)	39.5	27.5	-	42.8
GRU-EVE$_{hft+sem}$ [13] (IRV2+C+YL) (2019)	38.3	**28.4**	60.7	48.1
LG-DenseLSTM [42](V+C) (2019)	37.6	26.2	-	42.0
RecNet$_{local}$ [43] (IV4) (2019)	39.1	26.6	59.3	42.7
SAAT [44] (IRV2+C+FR) (2020)	40.5	28.2	60.9	**49.1**
BP-LSTMs [45] (R152) (2020)	36.6	27.0	58.7	40.5
STA-FG [46] (R152+C) (2020)	40.8	27.4	-	-
SHAN [47] (IRV2+I) (2022)	39.7	28.3	60.4	49.0
SMAN [48] (IRV2+C+F) (2022)	38.6	27.6	59.6	45.0
HTG+HMG (IV4)	39.0	27.5	59.6	45.4
HTG+HMG (R152)	39.1	27.6	59.6	45.4
HTG+HMG (R152+C)	**42.1**	**28.4**	**61.6**	48.9

method (1) and the CNN and LSTM-based methods (2 - 10). Due to previous methods using different CNN model to extract visual features and for a relatively fair comparison, we respectively leverage several typical CNN models to extract appearance features (Inception-V4 and ResNet-152) and motion features (C3D). The compared methods are as follows:

(1) Att-TVT [18] which proposed to use the transformer for video captioning.

(2) TDConvED [41] which aimed to fully employ convolutions in both encoder and decoder networks.

(3) GRU-EVE$_{hft+sem}$ [13] which embedded rich temporal dynamics in visual features by hierarchically applying Short Fourier Transform to CNN features of the whole video.

(4) LG-DenseLSTM [42] which proposed a dense LSTM for video captioning.

(5) RecNet$_{local}$ [43] which proposed to augmented the conventional attentive encoder-decoder with an LSTM-based local reconstructor.

(6) SAAT [44] which proposed a syntax-aware action targeting module.

(7) BP-LSTMs [45] which proposed an architecture comprising two LSTM layers and a word selection module.

(8) STA-FG [46] which proposed to a hierarchical decoder with semantic temporal attention and multi-fusion mechanism.

(9) SHAN [47] which proposed a syntax-guided hierarchical attention network to utilize semantic and syntax clues to integrate visual and sentence-context features for video captioning.

(10) SMAN [48] which proposed a stacked multimodal attention network to extends the decoder into a multi-layer stacked attention network, with textual and visual historical information explicitly learned.

Results on MSVD. We report the results on MSVD in Table 9. One one hand, compared to the methods using the single visual feature (*i.e.* 2, 5, 7), our HTG+HMG (IV4) achieves the better METEOR, ROUGE-L and CIDEr scores than RecNet$_{local}$, especially improving the CIDEr score with 4.6%; the HTG+HMG (R152) outperforms TDConvED (R152) and BP-LSTMs (R152) on three and all of the metrics. Especially compared with BP-LSTMs (R152), the HTG+HMG (R152) has the increases of 24.9% on CIDEr scores. On the other hand, compared to the other methods using multiple kind of visual features, our HTG+HMG (R152+C) also outperforms them in terms of METEOR, ROUGE-L and CIDEr. Especially, compared to Att-TVT, GRU-EVE$_{hft+sem}$, LG-DenseLSTM, SAAT, and SMAN, the HTG+HMG (R152+C) achieves the improvements of 13.1%, 17.0%, 25.9%, 12.8%, and 7.9% on CIDEr. This superior performance mainly benefits by the proposed graphs that can learn the intra- and inter-relation for the attention model, thus 1) supporting proper semantic alignment between target word states and attended visual features; leveraging the attention information from the past and future to guide the current attention process. Besides, Att-TVT is a Transformer-based method which outperforms us on BELU-4, so incorporating our proposed graph learning with Transformer will be a new research direction.

Results on MSR-VTT. We report the results on MSR-VTT in Table 10. We first compare the HTG+HMG (IV4) with RecNet$_{local}$ (IV4). We can find that HTG+HMG achieves better METEOR, ROUGE-L and CIDEr scores and is on par with RecNet$_{local}$ on BLEU-4, especially improving the CIDEr score with 6.3%. Besides, the HTG+HMG (R152) outperforms TDConvED (R152) except BLEU-4 and BP-LSTMs (R152) on all metrics. Especially compared with BP-LSTMs (R152), the HTG+HMG (R152) has the increases of 12.1% on CIDEr scores. Next, compared to the methods using multiple kinds of features, on one hand, our HTG+HMG (R152+C) outperforms all the methods on BLEU-4, METEOR, and ROUGE-L metrics, and is slight lower than SAAT (48.9 vs. 49.1). On the other hand, compared to the methods using three kinds of features (*i.e.*,3, 6, 10), our method also achieves the best scores on most metrics. Besides, compared to MSVD, this dataset is more challenge, because it includes large scale video-caption pairs, diverse video content and captions. Hence, it is beneficial to obtain better results when using multiple kinds of features extracted from videos.

4.10. Qualitative analysis

Caption generated by different kinds of graphs. We have quantitatively evaluated the proposed HMG, HTG and HMG+HTG on two datasets, as shown in Tables 1 and 2. However, these evaluations are mainly based on the automatic evaluation metrics, so we cannot intuitive observe whether the captions generated by them are good or not. To this end, in Fig. 6, we show the four examples with videos and captions generated by human-annotated ground truths, a conventional attentive encoder-decoder (Base) as well as our proposed methods, *i.e.*, HMG, HTG, HTG+HMG. From the first video of Fig. 6, we can intuitively see that the caption generated by the Base can describe the main video content, but it still lacks a detail, *i.e.*,"dog's tail", which is helpful to generate a more vivid caption. For the second video, the "cosmetic brush" is wrongly understood as a "microphone" by the Base, so it generates a totally irrelevant caption concerning this video. For the third video, as the Base cannot attend to the correct part which the decoder really expects, a certain product is wrongly described as a "video segment". By contrast, the HMG, HTG, and their combination are able to generate the accurate captions via enhancing the intra-relation and/or inter-relation for attention. Furthermore, it is interesting to observe that in the fourth video, the main object

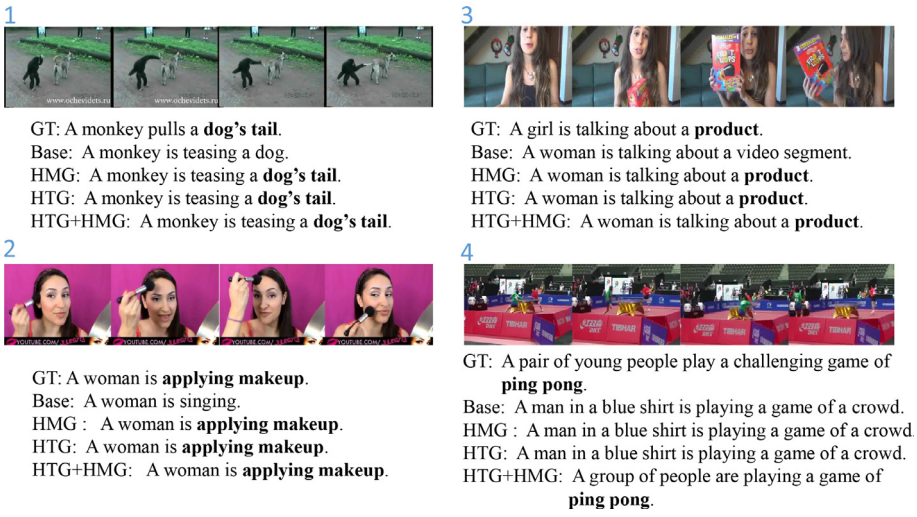

Fig. 6. Four examples from the test sets of MSVD and MSR-VTT, which involve human-annotated ground truth captions (GT) and the captions generated by the conventional attentive encoder-decoder (Base), HMG, HTG, HTG+HMG, respectively.

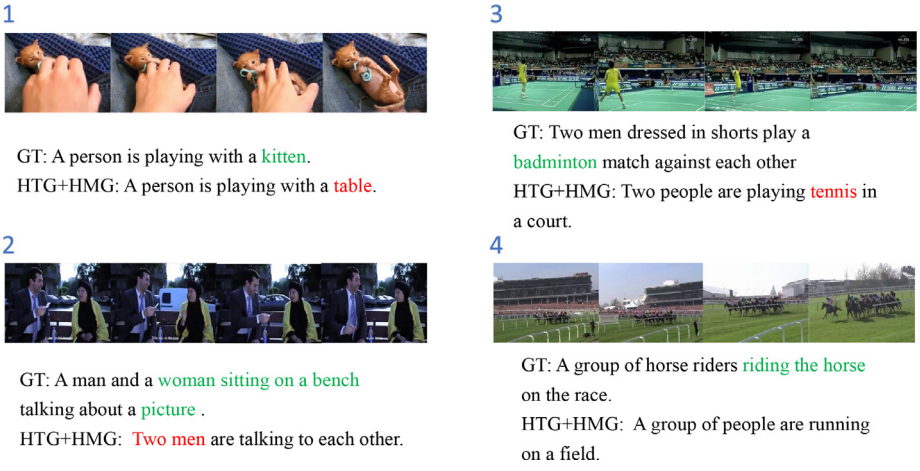

Fig. 7. Four failure examples obtained from HTG+HMG on the test set of MSVD and MSR-VTT. GT refers to ground truth. The red words in the captions generated by HTG+HMG indicate that it fails to describe the corresponding details in the ground truth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

"ping-pong" is tiny and moving quickly. Hence, it is not enough to generate an accurate caption via only enhancing the inter-relation or intra-relation for attention individually. In this case, it is necessary to enhance both of them for attention. In a word, coupling our proposed graphs with the attentive encoder-decoder can effectively improve the quality of generated video captions.

Failure Examples. We also find that in some cases, the proposed method fails to generate detailed or even yield inaccurate sentence to caption the given videos. Fig. 7 visualize four failure cases on the test set of MSVD and MSR-VTT dataset. We can observe that for the 1st video, the tiny but key object "kitten" is wrongly recognized as "table"; in the 2nd video, the model cannot recognize that one of person is a "woman" and the talked object "picture"; the main event "badminton" match is wrongly described as "tennis" match in the 3rd case; in the 4th example, there are a clutter of horses in a grass field, and these "horses" are ignored by the proposed method. Our conjecture is that the proposed model mainly uses the attention model to select key visual features at frame level rather than object level. As such, although the proposed HTG+HMG can enhance the alignment between attended visual features and word states, and use past and future attention results to regularize current attention, it has inadequate ability to focus on and describe the fined-grained objects. Hence, we will explore to incorporate the proposed graph learning into more fine-grained attention model, such as spatial-temporal, textual-temporal, and multi-level attention [1,12,23].

5. Conclusion

This paper proposes to improve the conventional attention mechanism to a relation-aware attention mechanism with two kinds of graph learning, namely the vision-to-vision homogeneous graph (HMG) and the linguistics-to-vision heterogeneous graph (HTG). The HMG aims to capture the intra-relations among all of the attend visual features. The HTG aims to enhance the inter-relations between target word states and attended visual features. Therefore, the proposed relation-aware attention mechanism not only exploits the attention information from the past and future to guide the current attention process, but also makes each attended visual feature more relevant to the corresponding target word state. Experimental results show that our proposed method not only significantly improves the conventional attentive encoder-decoder, but also achieves the state-of-the-art results on two well-known datasets. In the future, we will try to couple both graphs with fine-grained attention models to focus on the object-level information, so as to further boost the performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work was supported by National Key Research and Development Plan (Grant Nos. 2019QY1801, 2019QY1802, 2019QY1800), National Natural Science Foundation of China (Grant Nos. 61972186, 61732005, U21B2027, 61866020), Yunnan high-tech industry development project (Grant No. 201606), Yunnan provincial major science and technology special plan projects (Grant No. 202103AA080015, 202002AD080001-5), Yunnan Basic Research Project (Grant Nos. 202001AS070014), and Reserve Talents for Academic and Technological Leaders in Yunnan Province (Grant No. 202105AC160018).

References

[1] Y. Tu, C. Zhou, J. Guo, S. Gao, Z. Yu, Enhancing the alignment between target words and corresponding frames for video captioning, Pattern Recognit 111 (2021) 107702.
[2] L. Gao, X. Li, J. Song, H.T. Shen, Hierarchical LSTMs with adaptive attention for visual captioning, IEEE Trans Pattern Anal Mach Intell 42 (5) (2020) 1112–1131.
[3] J.H. Lim, C.S. Chan, K.W. Ng, L. Fan, Q. Yang, Protect, show, attend and tell: empowering image captioning models with ownership protection, Pattern Recognit 122 (2022) 108285.
[4] Z. Yang, P. Wang, T. Chu, J. Yang, Human-centric image captioning, Pattern Recognit 126 (2022) 108545.
[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, In Neural Computation, (1997) 1735–1780.
[6] C.J. Howard, A.O. Holcombe, Unexpected changes in direction of motion attract attention, Attention, Perception, & Psychophysics 72 (8) (2010) 2087–2095.
[7] L. Itti, P.F. Baldi, Bayesian surprise attracts human attention, in: NeurIPS, 2006, pp. 547–554.
[8] D.L. Chen, W.B. Dolan, Collecting highly parallel data for paraphrase evaluation, in: ACL, 2011, pp. 190–200.
[9] J. Xu, T. Mei, T. Yao, Y. Rui, MSR-VTT: a large video description dataset for bridging video and language, in: CVPR, 2016, pp. 5288–5296.
[10] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, Translating video content to natural language descriptions, in: ICCV, 2013, pp. 433–440.
[11] R. Xu, C. Xiong, W. Chen, J.J. Corso, Jointly modeling deep video and compositional text to bridge vision and language in a unified framework, in: AAAI, 2015, pp. 2346–2352.
[12] Y. Tu, X. Zhang, B. Liu, C. Yan, Video description with spatial-temporal attention, in: ACM MM, 2017, pp. 1014–1022.
[13] N. Aafaq, N. Akhtar, W. Liu, S.Z. Gilani, A. Mian, Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning, in: CVPR, 2019, pp. 12487–12496.
[14] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, in: NAACL-HLT, 2015, pp. 1494–1504.
[15] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: ICCV, 2015, pp. 4507–4515.
[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[17] Y. Tu, L. Li, L. Su, S. Gao, C. Yan, Z.-J. Zha, Z. Yu, Q. Huang, I2transformer: Intra- and inter-relation embedding transformer for TV show captioning, IEEE Trans. Image Process. 31 (2022) 3565–3577.
[18] M. Chen, Y. Li, Z. Zhang, S. Huang, Tvt: Two-view transformer network for video captioning, in: ACML, 2018, pp. 847–862.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017, pp. 5998–6008.
[20] B. Pan, H. Cai, D.-A. Huang, K.-H. Lee, A. Gaidon, E. Adeli, J.C. Niebles, Spatio-temporal graph for video captioning with knowledge distillation, CVPR, 2020.
[21] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J.R. Hershey, T.K. Marks, K. Sumi, Attention-based multimodal fusion for video description, in: ICCV, 2017, pp. 4193–4202.
[22] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Stat: spatial-temporal attention mechanism for video captioning, IEEE Trans Multimedia 22 (1) (2020) 229–241.
[23] B. Zhao, X. Li, X. Lu, CAM-RNN: co-attention model based RNN for video captioning, IEEE Trans. Image Process. 28 (11) (2019) 5552–5565.
[24] X. Long, C. Gan, G. de Melo, Video captioning with multi-faceted attention, Transactions of the Association for Computational Linguistics 6 (2018) 173–184.
[25] H. Ryu, S. Kang, H. Kang, C.D. Yoo, Semantic grouping network for video captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 2514–2522.
[26] J. Wang, W. Wang, Y. Huang, L. Wang, T. Tan, M3: Multimodal memory modelling for video captioning, in: CVPR, 2018, pp. 7512–7520.
[27] Y. Qin, J. Du, Y. Zhang, H. Lu, Look back and predict forward in image captioning, in: CVPR, 2019, pp. 8367–8375.
[28] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, Learning visual relationship and context-aware attention for image captioning, Pattern Recognit 98 (2020) 107075.
[29] J. Zhang, Y. Peng, Object-aware aggregation with bidirectional temporal graph for video captioning, CVPR, 2019.
[30] R. Pasunuru, M. Bansal, Multi-task video captioning with video and entailment generation, ACL, 2017.
[31] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, ICLR (2017).
[32] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318.
[33] S. Banerjee, A. Lavie, Meteor: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL, 2005, pp. 65–72.
[34] C.-Y. Lin, Rouge: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[35] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: CVPR, 2015, pp. 4566–4575.
[36] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. Zitnick, Microsoft coco captions: data collection and evaluation server, arXiv preprint arXiv:1504.00325 (2015).
[37] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, 2017, pp. 4278–4284.
[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[40] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? in: CVPR, 2018, pp. 6546–6555.
[41] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, T. Mei, Temporal deformable convolutional encoder-decoder networks for video captioning, in: AAAI, 2019, pp. 8167–8174.
[42] Y. Zhu, S. Jiang, Attention-based densely connected LSTM for video captioning, in: ACM MM, 2019, pp. 802–810.
[43] W. Zhang, B. Wang, L. Ma, W. Liu, Reconstruct and represent video contents for captioning via reinforcement learning, IEEE Trans Pattern Anal Mach Intell (2019). 1–1
[44] Q. Zheng, C. Wang, D. Tao, Syntax-aware action targeting for video captioning, in: CVPR, 2020, pp. 13096–13105.
[45] M. Nabati, A. Behrad, Video captioning using boosted and parallel long short-term memory networks, Comput. Vision Image Understanding 190 (2020) 102840.
[46] L. Gao, X. Wang, J. Song, Y. Liu, Fused GRU with semantic-temporal attention for video captioning, Neurocomputing 395 (2020) 222–228.
[47] J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha, Q. Huang, Syntax-guided hierarchical attention network for video captioning, IEEE Trans. Circuits Syst. Video Technol. 32 (2) (2022) 880–892.
[48] Y. Zheng, Y. Zhang, R. Feng, T. Zhang, W. Fan, Stacked multimodal attention network for context-aware video captioning, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2022) 31–42.

Yunbin Tu He received the B.S. degree in Automation from Hangzhou Dianzi University, and the M.S. degree in Pattern Recognition and Intelligent System from Kunming University of Science and Technology. He is currently pursuing the Ph.D. degree from the School of Computer Science and Technology, University of Chinese Academy of Sciences. His research interests include multimedia content analysis, especially for video and change captioning.

Chang Zhou He received the bachelor's degree in Electrical engineering and its automation from China University of Mining and Technology. He is studying for a M.S degrees in control engineering from Tsinghua University. His research interests include image processing and computer vision.

Junjun Guo He graduated with Ph.D. degree in engineering from Xi'an Jiaotong University. He is working as a lecturer of Kunming University of Technology. His main research interest is Multi-model information fusion.

Huafeng Li He received the M.S. degree in applied mathematics major from Chongqing University in 2009 and obtained his Ph.D. degree in control theory and control engineering major from Chongqing University in 2012. He is currently a professor at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include image processing, computer vision, and information fusion.

Shengxiang Gao She received the M.S. degree in Pattern Recognition and Intelligent System and the Ph.D. degree in Control Engineering from Kunming University of Science and Technology in 2005 and 2016, respectively. She is currently associate professor in School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her main research interests include machine learning, nature language processing and machine translation.

Zhengtao Yu He received his Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in2005. He is currently a professor in the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language processing, information retrieval and machine learning.