Low-Rank HOCA: Efficient High-Order Cross-Modal Attention for Video Captioning

Tao Jin, Siyu Huang, [*] **Yingming Li, and Zhongfei Zhang**
College of Information Science & Electronic Engineering
Zhejiang University, China
{jint_zju, siyuhuang, yingming, zhongfei}@zju.edu.cn

Abstract

This paper addresses the challenging task of video captioning which aims to generate descriptions for video data. Recently, the attention-based encoder-decoder structures have been widely used in video captioning. In existing literature, the attention weights are often built from the information of an individual modality, while, the association relationships between multiple modalities are neglected. Motivated by this observation, we propose a video captioning model with High-Order Cross-Modal Attention (HOCA) where the attention weights are calculated based on the high-order correlation tensor to capture the frame-level cross-modal interaction of different modalities sufficiently. Furthermore, we novelly introduce Low-Rank HOCA which adopts tensor decomposition to reduce the extremely large space requirement of HOCA, leading to a practical and efficient implementation in real-world applications. Experimental results on two benchmark datasets, MSVD and MSR-VTT, show that Low-rank HOCA establishes a new state-of-the-art.

1 Introduction

Video captioning has drawn much attention from natural language processing and computer vision researchers (Venugopalan et al., 2014; Bin et al., 2016; Ramanishka et al., 2016; Zanfir et al., 2016). As videos typically consist of multiple modalities (image, motion, audio, etc.), video captioning is actually a multimodal learning task. The abundant information underlying in the modalities is much beneficial for video captioning models. However, how to effectively learn the association relationships between different modalities is still a challenging problem.

In the context of deep learning based video captioning, multimodal attention mechanisms (Xu

---

[*]Corresponding author.

GT1: a referee coaches a wrestling match
GT2: boys are wrestling in front of a crowd

Figure 1: An example of video captioning, where "GT" denotes ground truth.

et al., 2017; Hori et al., 2017; Wang et al., 2018c) are shown to help deliver superior performance. Bahdanau Attention (Bahdanau et al., 2014) is widely used to calculate the attention weights according to the features of individual modalities, since Bahdanau Attention is originally proposed for machine translation which can be considered as a unimodal task. However, video captioning is a multimodal task and different modalities are able to provide complementary cues to each other when calculating the attention weights. For example, as shown in Fig. 1, a part of the video shows that two men are wrestling and another part shows that the referee coaches the match. With only the image modality, the model may pay equal attention to the frames with the competitors and referee. Considering the additional motion modality, the attention mechanism weighs more on one of them, resulting in a model with more focused attention.

Motivated by the above observations, in this paper we propose an attention mechanism called High-Order Cross-Modal Attention (HOCA) for video captioning, which makes a full use of the different modalities in video data by capturing their structure relationships. The key idea of HOCA is to consider the information of the other modalities when calculating the attention weights for one modality, different from Bahdanau Attention. In addition, we propose a low-rank version of HOCA which significantly reduces the com-

putational complexity of HOCA. Specifically, the attention weights of HOCA are computed based on the similarity tensor between modalities, fully exploiting the correlation information of different modalities at each time step. Given the fact that the space requirement of the high-order tensor increases exponentially in the number of modalities and inspired by tensor decomposition, we adopt a low-rank correlation structure across modalities into HOCA to enable a good scalability to the increasing number of modalities. Such improvement largely reduces the algorithm complexity of HOCA with good results in empirical study.

Our contributions can be summarized as:

(1) We propose High-Order Cross-Modal Attention (HOCA), which is a novel multimodal attention mechanism for video captioning. Compared with the Bahdanau Attention, HOCA captures the frame-level interaction between different modalities when computing the attention weights, leading to an effective multimodal modeling.

(2) Considering the scalability to the increasing number of modalities, we propose Low-Rank HOCA, which employs tensor decomposition to enable an efficient implementation of High-Order Cross-Modal Attention.

(3) Experimental results show that our method outperforms the state-of-the-art methods on video captioning, demonstrating the effectiveness of our method. In addition, the theoretical and experimental complexity analyses show that Low-Rank HOCA can implement multimodal correlation effeciently with acceptable computing cost.

2 Related Work

2.1 Attention Mechanism

The encoder-decoder structures have been widely used in sequence transformation tasks. Some models also connect the encoder and decoder through an attention mechanism(Bahdanau et al., 2014; Luong et al., 2015). In natural language processing (NLP), (Bahdanau et al., 2014) first proposes the soft attention mechanism to adaptively learn the context vector of the target keys/values. Such attention mechanisms are used in conjunction with a recurrent network. In the context of video captioning, there are many methods (Hori et al., 2017; Xu et al., 2017) using Bahdanau Attention to learn the context vector of the temporal features of video data, including the recent state-of-the-art method, hierarchically

aligned cross-modal attentive network (HACA) (Wang et al., 2018c). Such method ignores the structure relationships between modalities when computing the attention weights.

2.2 Video Captioning

Compared with image data, video data have rich multimodal information, such as image, motion, audio, semantic object, text. Therefore, the key is how to utilize the information. In the literature of video captioning, (Xu et al., 2017; Hori et al., 2017) propose the hierarchical multimodal attention, which selectively attends to a certain modality when generating descriptions. (Shen et al., 2017; Gan et al., 2017) adopt multi-label learning with weak supervision to extract semantic features of video data. (Wang et al., 2018b) proposes optimizing the metrics directly with hierarchical reinforcement learning. (Chen et al., 2017) extracts five types of features to develop the multimodal video captioning method and achieves promising results. (Wang et al., 2018a) performs the reconstruction of input video features by utilizing the output of the decoder, increasing the consistency of descriptions and video features. (Wang et al., 2018c) proposes the hierarchically aligned multimodal attention (HACA) to selectively fuse both global and local temporal dynamics of different modalities.

None of the methods mentioned above utilizes the interaction information of different modalities to calculate the attention weights. Motivated by this observation, we present HOCA and Low-Rank HOCA for video captioning. Different from the widely used Bahdanau Attention, our methods fully exploits the video representation of different modalities and their frame-level interaction information when computing the attention weights.

We introduce our methods in the following sections, where Section 3 introduces the details of Bahdanau Attention based multimodal video captioning (background), Section 4 gives the derivations of HOCA and Low-Rank HOCA, and the encoder-decoder structure which we propose for video captioning. Section 5 and 6 show the experimental settings and results of our methods.

3 Multimodal Video Captioning

Encoder-decoder structures combined with Bahdanau Attention are widely used in multimodal video captioning. Suppose that the num-

ber of input modalities is n, I_l denotes the features of l-th modality, the space of I_l is $\mathbb{R}^{d_l \times t_l}$ where t_l denotes the temporal length and d_l denotes the feature dimensions. The corresponding output is word sequence.

3.1 Encoder

The input video is fed to multiple feature extractors, which can be pre-trained CNNs for classification tasks such as Inception-Resnet-v2(Szegedy et al., 2017), I3D(Carreira and Zisserman, 2017), VGGish(Hershey et al., 2017), each extractor corresponds to one modality. The extracted features are sent to Bi-LSTM (Hochreiter and Schmidhuber, 1997) which has a capability to process sequence data, capturing the information from both forward and backward directions. The output of Bi-LSTM is kept.

3.2 Decoder

The decoder aims to generate word sequence by utilizing the features provided by the encoder and the context information. At time step t, the decoder output can be obtained with the input word y_t and the previous output h_{t-1},

$$h_t = \text{LSTM}(h_{t-1}, y_t) \quad (1)$$

We treat h_t as query q, the features of different modalities are allocated weights with Bahdanau Attention seperately as shown in Fig. 2(a). The attention weights $\alpha_t^{I_l} \in \mathbb{R}^{t_l}$ of l-th modality can be obtained as follows:

$$(\alpha_t^{I_l})_{r_l} = \frac{\exp\left\{(e_t^{I_l})_{r_l}\right\}}{\sum_{n=1}^{t_l} \exp\left\{(e_t^{I_l})_n\right\}} \quad (2)$$

$$(e_t^{I_l})_{r_l} = w_1^{\mathbb{T}}\tanh\left[W_1 h_t + U_1(I_l)_{r_l} + b_1\right] \quad (3)$$

where $(I_l)_{r_l}$ denotes the r_l-th time step of I_l and $(\alpha_t^{I_l})_{r_l}$ is the corresponding weight, we combine the attention weights and features, obtaining the context vector $\varphi_t(I_l)$ of l-th modality,

$$\sum_{r_l=1}^{t_l} (\alpha_t^{I_l})_{r_l} = 1 \quad (4)$$

$$\varphi_t(I_l) = \sum_{r_l=1}^{t_l} (\alpha_t^{I_l})_{r_l} (I_l)_{r_l} \quad (5)$$

the context vectors of other modalities can be obtained in the same way. We then integrate the context vectors to predict the word as follows:

$$p_t = \text{softmax}\left[W_p^h h_t + \sum_{i=1}^{n} W_p^{I_i} \varphi_t(I_i) + b_p\right] \quad (6)$$

where n is the number of modalities, p_t denotes the probability distribution of words at time step t.

4 HOCA and Low-Rank HOCA

Considering multimodal features, Bahdanau Attention and its variants process different modalities separately. In such situation, the interaction between different modalities is ignored. We propose HOCA and Low-Rank HOCA to excavate this information, in addition, the tensor decomposition used in Low-Rank HOCA reduces the complexity of the high-order correlation.

4.1 High-Order Cross-Modal Attention (HOCA)

Different modalities in video data are able to provide complementary cues for each other and should be integrated effectively. Inspired by the correlation matrix which is widely used in the natural language understanding (Seo et al., 2016), we use a high-order correlation tensor to model the interaction between different modalities. Fig. 2(b) shows the generalized form of the structure for n modalities, the core module is "tensor multiplication" defined by ourselves. After the nonlinear mapping similar to Eqn. 3 with query h_t, the features of the n modalities are in a d-dimensional common space. Note that, for convenience, we still use I_l to represent the features of the l-th modality (instead of $I_{l,t}$) and omit t which denotes the time step of the decoder in the following derivations of Section 4.1 and 4.2.

Let $\alpha_n^{I_l}$ denote the target attention weights of modality I_l. We obtain $\alpha_n^{I_l}$ through the high-order correlation tensor C_n between the n modalities. C_n can be obtained as below:

$$(C_n)_{r_1,...,r_n} = \mathbf{1}_d\left[(I_1)_{r_1} \circ (I_2)_{r_2} \circ ... \circ (I_n)_{r_n}\right]$$

$$= \mathbf{1}_d\left[\bigwedge_{i=1}^{n} (I_i)_{r_i}\right] \quad (7)$$

$$C_n = \bigotimes\left\{I_1, I_2, ..., I_n\right\} = \bigotimes_{i=1}^{n} I_i \quad (8)$$

where the $(r_1, ..., r_n)$-th entry of C_n is the inner-product of the r_1-th column (time step) of I_1, r_2-th column (time step) of I_2,...,r_n-th column (time step) of I_n. $\circ$ denotes element-wise multiplication and Λ denotes the element-wise multiplication $\circ$
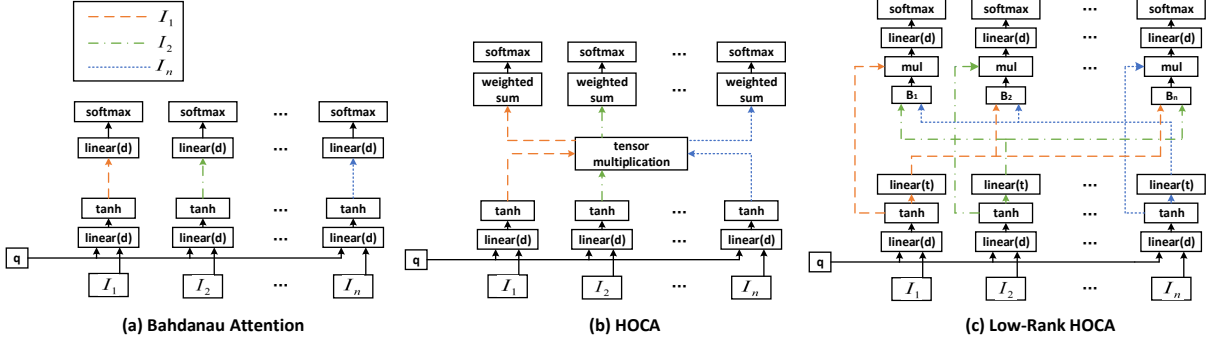
Figure 2: The attention mechanisms for multiple modalities where n denotes the number of modalities. (a) shows Bahdanau Attention for different modalities; I_l represents the l-th modality; linear(d) denotes the linear layer connected to the first dimension of I_l. (b) is the detailed structure of HOCA where the core module is "tensor multiplication" defined by ourselves. (c) is Low-Rank HOCA, where linear(t) denotes the linear layer in temporal (second) dimension; "mul" denotes multiplication operation; B_l is defined in Eqn. 16.

for a sequence of tensors. $\bigotimes$ is the tensor product operator and we use it to define a new operation $\bigotimes\{...,\}$ for the "tensor multiplication" of multiple matrices. $\mathbf{1}_d$ with space $\mathbb{R}^d$ consists of 1. We use the vectors $\mathbf{1}_d$ and $\mathbf{1}_{t_i}$ to denote the summation operation along the d and t_i dimensions. In this situation, Eqn. 7 and Eqn. 8 are equivalent. The attention weights of I_l can be calculated as below:

$$(\alpha_n^{I_l})_{r_l} = \frac{\exp\Big\{\sum\Big[W_{n-1}^{I_l} \circ (C_n^{I_l})_{r_l}\Big]\Big\}}{\sum_{o=1}^{t_l}\exp\Big\{\sum\Big[W_{n-1}^{I_l} \circ (C_n^{I_l})_o\Big]\Big\}} \quad (9)$$

where $(C_n^{I_l})_{r_l}$ is equal to $(C_n)_{:,:,...,r_l,...,:,:}$, which is an ($n$-1)-order tensor and denotes the correlation values between the r_l-th time step of I_l and different time steps of the other n-1 modalities. $W_{n-1}^{I_l}$ is also an (n-1)-order tensor, which has the same shape and denotes the relative importance of the correlation values. $\sum$ denotes the summation function for the high-order tensor. For simplicity, the "weighted sum" module can be considered as a linear layer in multiple dimensions.

4.2 Low-Rank HOCA

One of the main drawbacks of HOCA is the generation of high-order tensor, the size of the high-order tensor will increases exponentially with the number of modalities as $\prod_{i=1}^n t_i$, resulting in a lot of computation. Therefore, we implement the multimodal correlation between the different modalities in a more efficient way with low-rank approximation which has been widely used in the community of vision and language (Lei et al., 2015; Liu et al., 2018; Yu et al., 2017). Following the Eqn. 7 and 8, we rewrite $(C_n^{I_l})_{r_l}$ as[1]:

$$(C_n^{I_l})_{r_l} = \bigotimes\Big\{(I_l)_{r_l} \circ I_1, ..., I_{l-1}, I_{l+1}, ..., I_n\Big\}$$
$$= \bigotimes_{i=1,\neq l}^n (I_l)_{r_l} \star I_i \quad (10)$$

Element-wise multiplication operator $\circ$ is used for vector $(I_l)_{r_l}$ and matrix I_1. Each column of matrix I_1 multiplies vector $(I_l)_{r_l}$. $\star$ denotes that $(I_l)_{r_l}$ is multiplied ($\circ$) only when $i = 1$. We assume a low-rank factorization of the tensor $W_{n-1}^{I_l}$. Specifically, $W_{n-1}^{I_l}$ is decomposed into a sum of k rank 1 tensors,

$$W_{n-1}^{I_l} = \sum_{j=1}^k \bigotimes_{i=1,\neq l}^n w_j^{I_i} \quad (11)$$

where the space of $w_j^{I_i}$ is $\mathbb{R}^{1 \times t_i}$. Note that we set k to a constant value and use the recovered low-rank tensor to approximate $W_{n-1}^{I_l}$. The numerator in Eqn. 9 can be further derived as:

$$\sum\Big[W_{n-1}^{I_l} \circ (C_n^{I_l})_{r_l}\Big] = \sum\Big[\sum_{j=1}^k \bigotimes_{i=1,\neq l}^n w_j^{I_i}$$
$$\circ \bigotimes_{i=1,\neq l}^n (I_l)_{r_l} \star I_i\Big] \quad (12)$$

Since I_i is a matrix and $w_j^{I_i}$ is a vector, we directly multiply I_i and corresponding $w_j^{I_i}$[1]. Each row of the matrix I_i multiplies vector $w_j^{I_i}$,

$$\sum\Big[W_{n-1}^{I_l} \circ (C_n^{I_l})_{r_l}\Big] = \sum_{j=1}^k\Big[\sum \bigotimes_{i=1,\neq l}^n (I_l)_{r_l} \star I_i \circ w_j^{I_i}\Big] \quad (13)$$
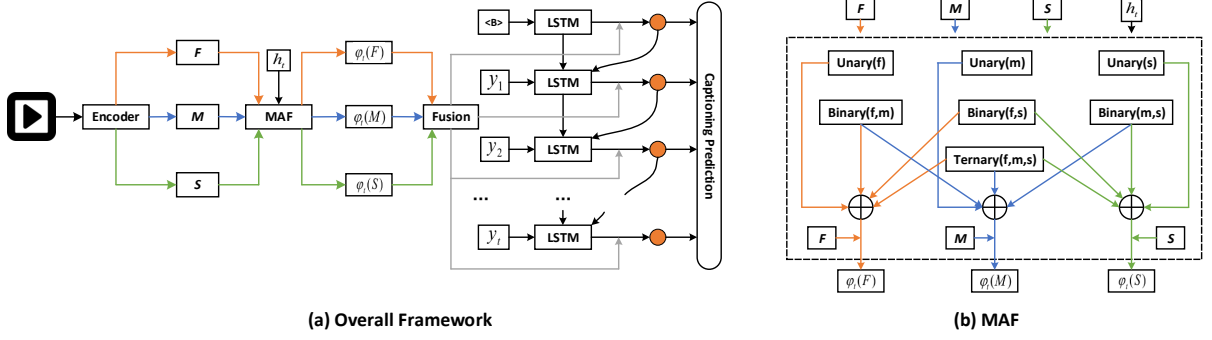
2004

(a) Overall Framework

(b) MAF

Figure 3: (a) is the overall framework of video captioning where F, M, S represent image, motion, and audio, respectively; h_t denotes the query at time step t. We predict words at all the time steps of the decoder with multiple attentive fusion (MAF) module. (b) is the detailed structure of MAF module, each modality has three types of attention weights (unary, binary, and ternary), where "Ternary" denotes that the number of modalities is 3 in HOCA and Low-Rank HOCA, "Binary" and "Unary" correspond to 2 and 1. The unary attention is equal to Bahdanau Attention. In addition, "Binary(f,m)" denotes the binary attention weights with image and motion. We utilize trainable parameters to determine the importance of these attention weights when integrating them.

where we first apply the tensor multiplication $\bigotimes$ to correlating the different time steps of all modalities. During the process, we sum in the d dimension to obtain the elements of the high-order tensor. Second, we sum (inner $\sum$) all the elements. For convenience, we change the operation order, we first sum in the temporal dimension, then sum in the d dimension[1]. Letting $(I_i')_j$ denote the global information of I_i with importance factor $w_j^{I_i}$,

$$(I_i')_j = (I_i \circ w_j^{I_i}) \mathbf{1}_{t_i} = I_i (w_j^{I_i})^{\mathbb{T}} \quad (14)$$

Eqn. 12 can be further derived as below:

$$\sum \left[W_{n-1}^{I_l} \circ (C_n^{I_l})_{r_l} \right] = \mathbf{1}_d \left[(I_l)_{r_l} \circ B_l \right] \quad (15)$$

$$B_l = \sum_{j=1}^{k} \bigwedge_{i=1, \neq l}^{n} (I_i')_j \quad (16)$$

Due to the different information carried by the elements of the feature, we use a linear layer w_{I_l} to replace $\mathbf{1}_d$,

$$(\alpha_n^{I_l})_{r_l} = \frac{\exp\left\{ w_{I_l} \left[(I_l)_{r_l} \circ B_l \right] \right\}}{\sum_{o=1}^{t_l} \exp\left\{ w_{I_l} \left[(I_l)_o \circ B_l \right] \right\}} \quad (17)$$

The detailed structure of Low-Rank HOCA is shown in Fig. 2(c).

---

[1] The details of the propositions are shown in the supplementary materials.

4.3 Complexity Analysis

We analyze the space complexity of Bahdanau Attention, HOCA, and Low-Rank HOCA in Fig. 2, focusing on the trainable variables and the output of each layer. For convenience, we start by calculating from the output of $tanh$ layer, since the front structures of three methods are same.

Bahdanau Attention The size of the trainable variable in second $linear(d)$ layer is $d \times 1$, and the size of the output is $\sum_{i=1}^{n} t_i$. The space complexity is $O(\sum_{i=1}^{n} t_i + d)$.

HOCA The size of the output of tensor multiplication is $\prod_{i=1}^{n} t_i$. The size of the trainable variables and corresponding output in $weighted\ sum$ layer is $\sum_{i=1}^{n} (\frac{\prod_{j=1}^{n} t_j}{t_i}) + \sum_{i=1}^{n} t_i$. The space complexity is $O(\prod_{i=1}^{n} t_i + \sum_{i=1}^{n} (\frac{\prod_{j=1}^{n} t_j}{t_i}) + \sum_{i=1}^{n} t_i)$.

Low-Rank HOCA The rank is set to k. The size of the trainable variable and corresponding output in $linear(t)$ layer is $\sum_{i=1}^{n} t_i \times k + n \times k \times d$. The size of the output in B_l and mul layer is $n \times d + \sum_{i=1}^{n} t_i \times d$. The size of the trainable variable and the corresponding output in second $linear(d)$ layer is $d + \sum_{i=1}^{n} t_i$. The space complexity is $O(\sum_{i=1}^{n} t_i \times (k + d + 1) + (n \times k + n + 1) \times d)$.

Bahdanau Attention and Low-Rank HOCA both scale linearly in the number of modalities while HOCA scales exponentially. Therefore, HOCA will have explosive complexity when the number of modalities is big, and Low-Rank HOCA can solve this problem effectively.

2005

Table 1: Evaluation results of our proposed models

Method	MSVD				MSR-VTT			
	BLEU4	ROUGE	METEOR	CIDEr	BLEU4	ROUGE	METEOR	CIDEr
HOCA-U	50.6	70.6	34.3	83.7	41.9	61.2	29.1	47.2
HOCA-B	49.6	70.8	34.3	84.3	42.2	61.1	29.1	47.6
L-HOCA-B	50.1	70.4	34.8	84.7	42.4	61.4	29.0	47.9
HOCA-T	50.9	71.2	34.8	84.8	42.4	61.5	29.0	48.1
L-HOCA-T	51.3	71.5	34.7	84.8	43.2	61.5	29.1	48.4
HOCA-UB	50.6	71.1	34.0	85.7	43.1	61.5	29.0	48.0
L-HOCA-UB	51.7	71.9	34.7	**86.1**	43.5	62.3	29.2	48.5
HOCA-UBT	52.3	71.9	**35.5**	85.3	43.9	62.0	29.0	49.3
L-HOCA-UBT	**52.9**	**72.0**	**35.5**	**86.1**	**44.6**	**62.6**	**29.5**	**49.8**

4.4 Video Captioning with HOCA and Low-Rank HOCA

In this section, we mainly introduce our encoder-decoder structure combined with high-order attention (HOCA and Low-Rank HOCA). As shown in Fig. 3(a), the features of different modalities, i.e. image(F), motion(M), audio(S) are extracted in the encoder. These features are sent to the decoder for generating words. The "MAF" module performs HOCA and Low-Rank HOCA for the features of different modalities with h_t. As shown in Fig. 3(b), each modality has three types of attention weights, i.e. unary, binary, ternary, which denote the different number (i.e. 1,2,3) of modalities applied to the HOCA and Low-Rank HOCA (note that unary attention is equal to Bahdanau Attention). In some cases, not all the modalities are effective, for example, the binary attention weights of image and motion are more accurate for the salient videos.

We use $\alpha_{t,3}^F$, $\alpha_{t,3}^M$, $\alpha_{t,3}^S$ to denote the ternary weights of three modalities at time step t, the calculation is shown as follows:

$$\alpha_{t,3}^F, \alpha_{t,3}^M, \alpha_{t,3}^S = \text{HOCA}(h_t, F, M, S) \quad (18)$$

where F,M,S denote the features of image, motion, audio, respectively, and HOCA can be replaced by Low-Rank HOCA. We obtain the binary and unary weights in the same way. For different attention weights, we utilize trainable variables to determine their importance. We take the image modality as an example, the fusion weights are obtained as follows:

$$\alpha_t^F = \text{softmax}(\theta_1\alpha_{t,1}^F + \theta_2\alpha_{t,2}^{F,M} + \theta_3\alpha_{t,2}^{F,S} + \theta_4\alpha_{t,3}^F) \quad (19)$$

where θ_{1-4} are trainable variables, $\alpha_{t,1}^F$ denotes the unary weights, $\alpha_{t,2}^{F,M}$ and $\alpha_{t,2}^{F,S}$ denote the binary weights with motion and audio, respectively. As we obtain the attention weights of three modalities, we can calculate the context vectors $\varphi_t(F)$, $\varphi_t(M)$, $\varphi_t(S)$ following the Eqn. 4 and 5.

To further determine the relative importance of multiple modalities, we perform a hierarchical attention mechanism for the context vectors in the "Fusion" module. The attention weights of image modality are calculated as follows:

$$\beta_t^F = \frac{\exp(e_t^F)}{\sum_{k=\{F,M,S\}} \exp(e_t^k)} \quad (20)$$

$$e_t^F = w_e^{\mathbb{T}} \tanh\left[W_e h_t + U_e \varphi_t(F) + b_e\right] \quad (21)$$

β_t^M and β_t^S are obtained in the same way. We integrate the context vectors and corresponding weights to predict word as follows:

$$p_t = \text{softmax}\left[W_p^h h_t + \sum_k \beta_t^k D_t(k) + b_p\right] \quad (22)$$

$$D_t(k) = W_p^k \varphi_t(k), k = \{F, M, S\} \quad (23)$$

The optimization goal is to minimize the cross-entropy loss function defined as the accumulative loss from all the time steps:

$$L = -\sum_{i=1}^{T} \log(p_t^*|p_1^*, ..., p_{t-1}^*, V) \quad (24)$$

where p_t^* denotes the probability of the ground-truth word at time step t, T denotes the length of description, V denotes the original video data.

5 Experimental Methodology

5.1 Datasets and Metrics

We evaluate video captioning on two standard datasets, MSVD (Chen and Dolan, 2011) and MSR-VTT (Xu et al., 2016), which are both provided by Microsoft Research, with several state-of-the-art methods. MSVD includes 1970 video clips. The time length of a video clip is about 10 to 25 seconds and each video clip is annotated with about 40 English sentences. MSR-VTT has 10000 video clips; each clip is annotated with 20 English sentences. We follow the commonly used protocol in the previous work and use four common metrics in the evaluation, including BLEU4, ROUGE, METEOR, and CIDEr.

5.2 Preprocessing and Experimental Setting

We sample video data to 80 frames for extracting image features. For extracting motion features, we divide the raw video data into video chunks centered on 80 sampled frames at the first step. Each video chunk includes 64 frames. For extracting audio features, we obtain the audio file from the raw video data with FFmpeg. For both datasets, we use a pre-trained Inception-ResNet-v2 (Szegedy et al., 2017) to extract image features from the sampled frames and we keep the activations from the penultimate layer. In addition, we use a pre-trained I3D (Carreira and Zisserman, 2017) to extract motion features from video chunks. We employ the activations from the last convolutional layer and implement a mean-pooling in the temporal dimension. We use the pre-trained VGGish (Hershey et al., 2017) to extract audio features. On MSRVTT, we also utilize the glove embedding of the auxiliary video category labels to initialize the decoder state.

The hidden size is 512 for all LSTMs. The attention layer size for image, motion, audio attention is also 512. The dropout rate for the input and output of LSTM decoder is 0.5. The rank is set to 1. In the training stage, we use Adam (Kingma and Ba, 2014) algorithm to optimize the loss function; the learning rate is set to 0.0001. In the testing stage, we use beam-search method with beam-width 5. We use a pre-trained word2vec embedding to initialize the word vectors. Each word is represented as a 300-dimension vector. Those words which are not in the word2vec matrix are initialized randomly. All the experiments are done on 4 GTX 1080Ti GPUs.

6 Experimental Results

6.1 Impact of Cross-Modal Attention

Table 2 shows the results of different variants of HOCA and Low-Rank HOCA. HOCA-U, HOCA-B, and HOCA-T denote the model with only unary, binary, and ternary attention, respectively. HOCA-UB and HOCA-UBT denote the models with original HOCA and more types of attention mechanisms. The prefix "L-HOCA" denotes the model with Low-Rank HOCA.

It is observed that the model with only one type of attention mechanism(U, B, or T) in the decoder achieves relatively bad results on both datasets. However, when we combine them, the performances are significantly improved on metrics, especially ROUGE and CIDEr. HOCA-UBT and L-HOCA-UBT with a combination of unary, binary, and ternary attention achieve relatively promising results on all the metrics. We argue that HOCA-UBT and L-HOCA-UBT can learn appropriate ratios of all types of attention mechanisms based on the specific video-description pairs, while other variants only focus on one or two types. In addition, the models of low-rank version (L-HOCA-UB and L-HOCA-UBT) have better metrics than the models of original version (HOCA-UB and HOCA-UBT). On the one hand, we utilize w_{I_l} to replace $\mathbf{1}_d$ in Eqn. 17, fully mining the different information carried by the elements of the feature, on the other hand, the low-rank approximation is effective.

Table 2: Evaluation results of video captioning, where B, R, M, C denote BLEU4, ROUGE, METEOR, CIDEr, respectively, and Ours denotes L-HOCA-UBT.

Dataset	Method	B	R	M	C
MSR-VTT	RecNet	39.1	59.3	26.6	42.7
	HRL	41.3	61.7	28.7	48.0
	Dense Cap	41.4	61.1	28.3	48.9
	HACA	43.5	61.8	**29.5**	49.7
	MM-TGM	44.3	-	29.3	49.2
	Ours	**44.6**	**62.6**	**29.5**	**49.8**
MSVD	SCN	50.2	-	33.4	77.7
	TDDF	45.8	69.7	33.3	73.0
	LSTM-TSA	52.8	-	33.5	74
	RecNet	52.3	69.8	34.1	80.3
	MM-TGM	48.7	-	34.3	80.4
	Ours	**52.9**	**72.0**	**35.5**	**86.1**

HOCA-U: a man is talking about a car
L-HOCA-UBT: a man is talking about the features of a car
GT: a guy talks about the features of the jeep Cherokee

HOCA-U: a woman is peeling a piece of raw meat
L-HOCA-UBT: a woman is peeling a shrimp
GT: a woman is peeling a shrimp

HOCA-U: a person is playing with an animal
L-HOCA-UBT: a person is playing with a cat
GT: the boy is playing the cat

HOCA-U: a person is playing a bike
L-HOCA-UBT: a man is demonstrating how to use a stroller
GT: a man displaying features on a stroller

Figure 4: Qualitative results of Low-Rank HOCA

6.2 Compared with the State-of-the-art

Table 2 shows the results of different methods on MSR-VTT and MSVD, including ours (L-HOCA-UBT), and some state-of-the-art methods, such as LSTM-TSA (Pan et al., 2017), TDDF (Zhang et al., 2017), SCN (Gan et al., 2017), MM-TGM (Chen et al., 2017), Dense Caption (Shen et al., 2017), RecNet (Wang et al., 2018a), HRL (Wang et al., 2018b),HACA(Wang et al., 2018c).

From Table 2, we find that Ours(L-HOCA-UBT) shows competitive performances compared with the state-of-the-art methods. On MSVD, L-HOCA-UBT has outperformed SCN, TDDF, Rec-Net, LSTM-TSA, MM-TGM, on all the metrics. In particular, L-HOCA-UBT achieves 86.1% on CIDEr, making an improvement of 5.7% over MM-TGM. On MSR-VTT, we have the similar observation, L-HOCA-UBT has outperformed RecNet, HRL, MM-TGM, Dense Caption, and HACA on all the metrics.

Table 3: Computing cost of different methods, where the "space" denotes the memory space requirement and the "training time" denotes the total time for training. We evaluate them on MSR-VTT. Note that the metrics belong to the whole model, not only the attention module.

Method	Space (G)	Training Time (s)
HOCA-U	4.1	16418
HOCA-UBT	9.7	37040
L-HOCA-UBT	5.6	24615

6.3 Computing Cost

The theoretical complexity of different attention mechanisms is illustrated in Section 4.3. In practice, we utilize the experimental settings mentioned above, the batch size and the maximum number of epochs are set to 25 and 100, respectively. the training time and memory space requirement are shown in Table 3. We can find that L-HOCA-UBT has smaller space requirement and less time cost than HOCA-UBT, in addition, the computing cost of L-HOCA-UBT is close to that of HOCA-U (Bahdanau Attention). The results demonstrate the advantage of Low-rank HOCA.

6.4 Rank Setting

We also evaluate the impact of different rank values. We show the results on MSVD in Fig. 5. The red and green lines represent HOCA-UBT and L-HOCA-UBT, respectively. We find that the CIDEr of L-HOCA-UBT has slight fluctuations as the rank changes and a small value of rank can achieve competitive results with high efficiency.
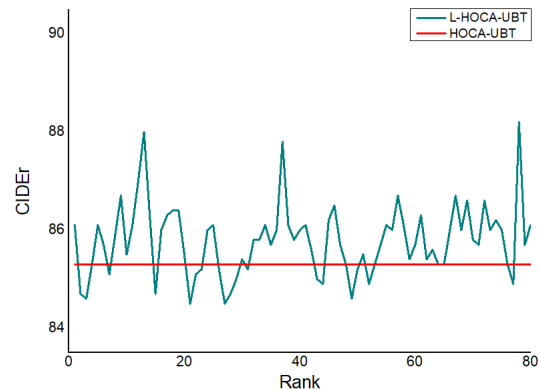

Figure 5: The performance of different rank values

6.5 Qualitative Analysis

Fig. 4 shows some qualitative results of our method. We simply compare the descriptions generated by HOCA-U and L-HOCA-UBT, respectively. GT represents "Ground Truth". Benefiting from the high-order correlation of multiple modalities, L-HOCA-UBT can generate more accurate descriptions which are close to GT.

7 Conclusion

In this paper, we have proposed a new cross-modal attention mechanism called HOCA for video captioning. HOCA integrates the information of the other modalities into the inference of attention weights of current modality. Furthermore, we have introduced the Low-Rank HOCA which has a good scalability to the increasing number of modalities. The experimental results on two standard datasets have demonstrated the effectiveness of our approach.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yi Bin, Yang Yang, Fumin Shen, Xing Xu, and Heng Tao Shen. 2016. Bidirectional long-short term memory for video description. In *ACM MM*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. 2017. Video captioning with guidance of multimodal latent topics. In *ACM MM*.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *CVPR*, volume 2.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *ICASSP*.

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

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *ICCV*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015. Molding cnns for text: non-linear, non-consecutive convolutions. *arXiv preprint arXiv:1508.04112*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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

Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *ACM MM*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

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

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018a. Reconstruction network for video captioning. In *CVPR*.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018b. Video captioning via hierarchical reinforcement learning. In *CVPR*.

Xin Wang, Yuan-Fang Wang, and William Yang Wang. 2018c. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msrvtt: A large video description dataset for bridging video and language. In *CVPR*.

Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. 2017. Learning multimodal attention lstm networks for video captioning. In *ACM MM*.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*.

Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2016. Spatio-temporal attention models for grounded video captioning. In *ACCV*.

Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. 2017. Task-driven dynamic fusion: Reducing ambiguity in video description. In *CVPR*.

A Supplemental Material

A.1 Proposition 1 for Eqn. 13 in the paper

Proposition 1: Suppose that we have n matrices, $I_1, I_2, ..., I_n$, and n vectors, $w_1, w_2, ..., w_n$. The space of I_l is $\mathbb{R}^{d \times t_l}$ and the space of w_l is $\mathbb{R}^{1 \times t_l}$. Then

$$\Big(\bigotimes_{i=1}^{n} I_i \Big) \circ \Big(\bigotimes_{i=1}^{n} w_i \Big) = \bigotimes_{i=1}^{n} I_i \circ w_i \qquad (25)$$

Proof: We use C_l and C_r to denote the left side and right side of the equation, respectively. We utilize the element-wise comparison in two tensors. Following Eqn. 7 and 8 in the paper, the $(r_1, ..., r_n)$-th entry of C_l is expressed as

$$(C_l)_{r_1,...,r_n} = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i} \Big] \cdot \Big[\bigwedge_{i=1}^{n} (w_i)_{r_i} \Big] \quad (26)$$

where $(I_i)_{r_i}$ is a vector which denotes r_i-th column of the I_i, $(w_i)_{r_i}$ is the r_i-th value of the vector. Since $(w_i)_{r_i}$ is a single element, we can directly multiply it with the corresponding vector $(I_i)_{r_i}$.

$$(C_l)_{r_1,...,r_n} = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i \circ w_i)_{r_i} \Big] \qquad (27)$$
$$= (C_r)_{r_1,...,r_n}$$

The proposition is proven and is used to convert Eqn. 12 to Eqn. 13 in the paper.

A.2 Proposition 2 for Eqn. 15 in the paper

Proposition 2: Suppose that we have n matrices, $I_1, I_2, ..., I_n$. The space of I_l is $\mathbb{R}^{d \times t_l}$. Then

$$\sum \Big(\bigotimes_{i=1}^{n} I_i \Big) = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i) \mathbf{1}_{t_i} \Big] \qquad (28)$$

here we use vectors $\mathbf{1}_d$ and $\mathbf{1}_{t_i}$ which consist of 1 to represent the summation operation for matrix I_i in d dimension and t_i dimensions, respectively.

Proof: We use v_l to denote the left side of the equation and v_r to denote the right side of the equation. We can express v_l as

$$v_l = \sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} C_{r_1,r_2,...,r_n} \qquad (29)$$

$$C = \bigotimes_{i=1}^{n} I_i \qquad (30)$$

Following Eqn. 7 and 8 in the paper, we can express $C_{r_1,r_2,...,r_n}$ as

$$C_{r_1,r_2,...,r_n} = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i} \Big] \qquad (31)$$

We apply Eqn. 32 to Eqn. 30,

$$v_l = \sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i} \Big]$$
$$= \mathbf{1}_d \Big[\sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} \bigwedge_{i=1}^{n} (I_i)_{r_i} \Big] \qquad (32)$$
$$= \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i) \mathbf{1}_{t_i} \Big] = v_r$$

The proposition is proven and is used to convert Eqn. 13 to Eqn. 15 in the paper.

A.3 Learning Curves

We show the learning curves of the CIDEr on the validation set in Fig. 6 and observe that the L-HOCA-UBT performs better than HOCA-UBT and HOCA-U when the training converges.
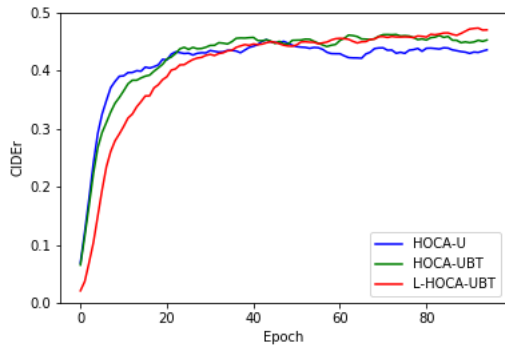
Figure 6: Learning curves of different methods on MSR-VTT, where the rank of L-HOCA-UBT is 1. Note that we use greedy search during training while beam search during testing, so the testing scores are higher.

A.4 Visualization of Attention Weights

We also perform visualization of the attention weights in multiple attentive fusion (MAF) module. As shown in Fig. 7, HOCA-UBT obtains a more accurate ratio of each modality than HOCA-U, i.e. for the word "man", HOCA-U obtains a higher score of motion modality, which violates human subjective understanding.

Figure 7: Visualization of the attention weights in multiple attentive fusion (MAF) module, the red bar denotes image modality, the green bar denotes motion modality, the blue bar denotes audio modality.