Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos

Shaoxiang Chen[1⋆], Wenhao Jiang[2], Wei Liu[2], and Yu-Gang Jiang[1⋆⋆]

[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University
[2] Tencent AI Lab
{sxchen13, ygj}@fudan.edu.cn, cswhjiang@gmail.com, wl2223@columbia.edu

Abstract. Automatically generating sentences to describe events and temporally localizing sentences in a video are two important tasks that bridge language and videos. Recent techniques leverage the multimodal nature of videos by using off-the-shelf features to represent videos, but interactions between modalities are rarely explored. Inspired by the fact that there exist cross-modal interactions in the human brain, we propose a novel method for learning pairwise modality interactions in order to better exploit complementary information for each pair of modalities in videos and thus improve performances on both tasks. We model modality interaction in both the sequence and channel levels in a pairwise fashion, and the pairwise interaction also provides some explainability for the predictions of target tasks. We demonstrate the effectiveness of our method and validate specific design choices through extensive ablation studies. Our method turns out to achieve state-of-the-art performances on four standard benchmark datasets: MSVD and MSR-VTT (event captioning task), and Charades-STA and ActivityNet Captions (temporal sentence localization task).

Keywords: Temporal Sentence Localization · Event Captioning in Videos · Modality Interaction

1 Introduction

Neuroscience researches [5,3,15] have discovered that the early sensory processing chains in the human brain are not unimodal, information processing in one modality (*e.g.*, auditory) can affect another (*e.g.*, visual), and there is a system in the brain for modulating cross-modal interactions. However, modality interactions are largely overlooked in the research of high-level video understanding tasks, such as event captioning [71,57,58] and temporal sentence localization [17,34,8]. Both tasks involve natural language descriptions and are

⋆ Part of the work is done when the author was an intern at Tencent AI Lab.
⋆⋆ Corresponding author.

substantially more challenging than recognition tasks. Thus, it is crucial to utilize information from each of the available modalities and capture inter-modality complementary information to better tackle these tasks.

Recent event captioning methods [37,47,60,9,26] mostly adopt an encoder-decoder structure, where the encoder aggregates video features and the decoder (usually LSTM [24] or GRU [13]) generates sentences based on the aggregation results. The video features stem mainly from the visual appearance modality, which are usually extracted with off-the-shelf CNNs (Convolutional Neural Networks) [50,21,51,46] that are pre-trained to recognize objects and can output high-level visual representations for still images. Using features from the visual modality solely can generally work well on video event captioning. Recent works [73,37,9,44,11,43,41] suggest that further improvements can be obtained by additionally leveraging motion and audio representations. However, the limitation of these works is that the features from multiple modalities are simply concatenated without considering their relative importances or the high-level interactions among them, so the great potential of multiple modalities has not been fully explored. There exist a few works [28,25,76,69] that learn to assign importance weights to individual modalities via cross-modal attention in the encoder, but modality interactions are still not explicitly handled. Temporal sentence localization in videos is a relatively new problem [17]. Although various approaches [49,74,10] have been proposed and significant progresses have been made, this problem has not been discussed in a multimodal setting. Most recently, Rahman *et al.* [41] emphasized the importance of jointly considering video and audio to tackle dense event captioning, in which sentence localization is a subtask. Apart from the visual, motion, and audio modalities, utilizing semantic attributes is gaining popularity in recent methods [1,36,10,64] for both event captioning and sentence localization.

In order to better exploit multimodal features for understanding video contents, we propose a novel and generic method for modeling modality interactions that can be leveraged to effectively improve performances on both the sentence localization and event captioning tasks. Our proposed Pairwise Modality Interaction (PMI) explicitly models sequence-level interactions between each pair of feature sequences by using a channel-gated bilinear model, and the outputs of each interacting pair are fused with importance weights. Such a modeling provides some explainability for the predictions.

Our main contributions are as follows:

- We propose a novel multimodal interaction method that uses a Channel-Gated Modality Interaction model to compute pairwise modality interactions (PMI), which better exploits intra- and inter-modality information in videos. Utilizing PMI achieves significant improvements on both the video event captioning and temporal sentence localization tasks.
- Based on modality interaction within video and text, we further propose a novel sentence localization method that builds video-text local interaction for better predicting the position-wise video-text relevance. To the best of

our knowledge, this is also the first work that addresses sentence localization in a multimodal setting.
– Extensive experiments on the MSVD, MSR-VTT, ActivityNet Captions, and Charades-STA datasets verify the superiority of our method compared against state-of-the-art methods on both tasks.

2 Related Works

Temporal Sentence Localization Gao *et al.* [17] proposed the temporal sentence localization task recently, and it has attracted growing interests from both the computer vision and natural language processing communities. Approaches for this task can be roughly divided into two groups, *i.e.*, proposal-based methods and proposal-free methods. TALL [17] uses a multimodal processing module to fuse visual and textual features for sliding window proposals, and then predicts a ranking score and temporal boundaries for each proposal. NSGV [8] performs interaction between sequentially encoded sentence and video via an LSTM, and then predicts K proposals at each time step. Proposal-free methods usually regress the temporal boundaries. As the most representative one, ABLR [74] iteratively applies co-attention between visual and textual features to encourage interactions, and finally uses the interacted features to predict temporal boundaries.

Event Captioning The S2VT [57] method is the first attempt at solving video captioning using an encoder-decoder network, in which two layers of LSTMs [24] first encode the CNN-extracted video features and then predict a sentence word-by-word. Later works are mostly based on the encoder-decoder structure, and improvements are made for either the encoder or decoder. Yao *et al.* [71] applied temporal attention to the video features, which enables the encoder to assign an importance weight to each video feature during decoding, and this method is also widely adopted by the following works. Some works [37,4,12,78,61] tried to improve the encoder by considering the temporal structures inside videos. Another group of works [70,33,9] are focused on exploiting spatial information in video frames by applying a dynamic attention mechanism to aggregate frame features spatially. Utilizing multimodal (appearance, motion, and audio) features is also common in recent works, but only a few works [25,76,69,36] tried to handle the relative importances among different modalities using cross-modal attention. Most recently, some works [1,75,36] have proven that incorporating object/semantic attributes into video captioning is effective. As for the decoder, LSTM has been commonly used as the decoder for video captioning, and some recent attempts have also been made to using non-recurrent decoders such as CNN [7] or the Transformer [77] structure.

Modality Interaction There are some works trying to use self-attention to model modality interaction. Self-attention has been proven effective on both vision [65] and language [55] tasks. Its effectiveness in sequence modeling can be attributed to that it computes a response at one position by attending to all positions in a sequence, which better captures long-range dependencies. Au-

toInt [48] concatenates features from different modalities and then feeds them to a multi-head self-attention module for capturing interactions. For the referred image segmentation task, Ye *et al.* [72] introduced CMSA (Cross-Modal Self-Attention), which operates on the concatenation of visual features, word embeddings, and spatial coordinates to model long-range dependencies between words and spatial regions. DFAF [18] is a visual question answering (VQA) method, which applies self-attention for regional feature sequences and word embedding sequences to model inter-modality interactions, and also models intra-modality interactions for each sequence. We note that modality interaction is common in VQA methods, but they usually pool the multimodal feature sequences into a single vector using bilinear or multi-linear pooling [16,30,31,35]. And VQA methods are more focused on the interaction between visual and textual modalities, so they do not fully exploit the modality interactions within videos.

Compared to these existing methods, our proposed Pairwise Modality Interaction (PMI) has two distinctive features: (1) modality interactions are captured in a pairwise fashion, and information flow between each pair of modalities in videos is explicitly considered in both the sequence level and channel level; (2) the interaction does not pool the feature sequences (*i.e.*, temporal dimension is preserved), and the interaction results are fused by their importance weights to provide some explainability.

3 Proposed Approach

3.1 Overview

We first give an overview of our approach. As shown in Fig. 1, multimodal features are first extracted from a given video and then fed to a video modality interaction module, where a Channel-Gated Modality Interaction is performed for all pairs of modalities to exploit intra- and inter-modality information. The interaction results are tiled into a high-dimensional tensor and we then use a simple fully-connected network to efficiently compute the importance weights to transform this tensor into a feature sequence. This process to model pairwise modality interaction is abbreviated as PMI.

For sentence localization, the text features are also processed with modality interaction to exploit its intra-modality information. Then video and textual features are locally interacted in order to capture the complex association between these two modalities at each temporal location. Finally, a light-weight convolutional network is applied as the localization head to process the feature sequence and output the video-text relevance score and boundary prediction.

For video captioning, since the focus of this paper is to fully exploit multimodal information, we do not adopt a sophisticated decoder architecture and only use a two-layer LSTM with temporal attention on top of the video modality interaction. However, due to the superiority of PMI, state-of-the-art performances are still achieved. Note that video modality interaction can be used in either a sentence localization model or an event captioning model, but the models are trained separately.

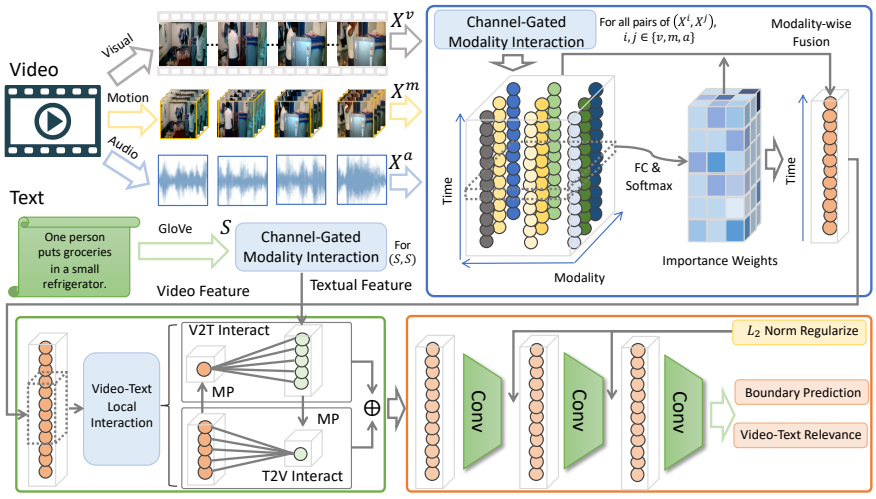

Fig. 1. The framework of our approach. The multimodal features from a video are processed with Channel-Gated Modality Interaction (see Fig. 2) for each pair of modalities, and a weighted modality-wise fusion is then executed to obtain an aggregated video feature(Blue box). **Note that this feature can also be used for video captioning, but the two tasks are not jointly trained**. For temporal sentence localization, the word embedding features also interact with themselves to exploit intra-sentence information, resulting in a textual feature. The video and textual features then interact locally at each temporal position (Green box), and the resulting feature is fed to a light-weight convolutional network with layer-wise norm regularization to produce predictions (Orange box). Each colored circle represents a feature vector.

3.2 Video Modality Interaction

Given an input video $\boldsymbol{V} = \{\boldsymbol{f}_i\}_{i=1}^{F}$, where $\boldsymbol{f}_i$ is the i-th frame, multimodal features can be extracted using off-the-shelf deep neural networks. In this paper, three apparent modalities in videos are adopted, which are visual modality, motion modality, and audio modality. Given features from these modalities, a sequence of features can be learned to represent the latent semantic modality[3]. The corresponding feature sequences from the above modalities are denoted by $\boldsymbol{X}^v = \{\boldsymbol{x}_n^v\}_{n=1}^{N}$, $\boldsymbol{X}^m = \{\boldsymbol{x}_n^m\}_{n=1}^{N}$, $\boldsymbol{X}^a = \{\boldsymbol{x}_n^a\}_{n=1}^{N}$, and $\boldsymbol{X}^l = \{\boldsymbol{x}_n^l\}_{n=1}^{N}$, respectively. The dimensionalities of the feature vectors in each modality are denoted as d_v, d_m, d_a, and d_l, respectively.

We propose to explicitly model modality interaction between a pair of feature sequences, denoted by $\boldsymbol{X}^p$ and $\boldsymbol{X}^q$, where $p \in \{a, m, v, l\}$ and $q \in \{a, m, v, l\}$. Note that p and q can be the same modality, and in that case, the interaction

[3] For fair comparison, we do not include this modality when comparing with state-of-the-art methods, but will demonstrate some qualitative results with the latent semantic modality. The corresponding learning method is placed in the Supplementary Material.

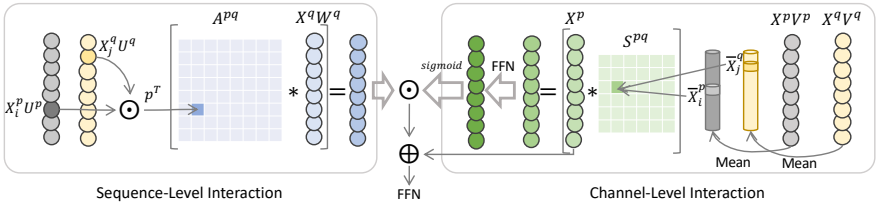

Fig. 2. Overview of Channel-Gated Modality Interaction. The Channel-Level Interaction results are used as a gating variable to modulate Sequence-Level Interaction results. Details are illustrated below in Eqs (1)-(6).

exploits intra-modality information. As shown in Fig. 2, the interaction can be formulated as

$$\mathrm{INT}(\boldsymbol{X}^p, \boldsymbol{X}^q) = \mathrm{FFN}\big(\mathrm{BA}(\boldsymbol{X}^p, \boldsymbol{X}^q) \odot \mathrm{CG}(\boldsymbol{X}^p, \boldsymbol{X}^q) \oplus \boldsymbol{X}^p\big), \tag{1}$$

where $\mathrm{BA}(\cdot)$ is the bilinear attention model that performs sequence-level modality interaction, $\mathrm{CG}(\cdot)$ is a channel gating mechanism based on the channel-level interaction and is used to modulate the sequence-level interaction output, a residual connection is introduced with $\oplus\boldsymbol{X}^p$, and $\mathrm{FFN}(\cdot)$ is a position-wise feed-forward network that projects its input into a lower dimension[4].

Sequence-Level Interaction We use a low-rank bilinear model to consider the interaction between each pair of elements in feature sequences $\boldsymbol{X}^p$ and $\boldsymbol{X}^q$:

$$\boldsymbol{A}^{pq}_{ij} = \boldsymbol{p}^T\big(\rho(\boldsymbol{X}^p_i\boldsymbol{U}^p) \odot \rho(\boldsymbol{X}^q_j\boldsymbol{U}^q)\big), \quad \boldsymbol{\mathcal{A}}^{pq}_{ij} = \mathrm{Softmax}_j(\boldsymbol{A}^{pq}_{ij}), \tag{2}$$

where $\boldsymbol{X}^p_i$ is the i-th element of $\boldsymbol{X}^p$, $\boldsymbol{X}^q_j$ is the j-th element of $\boldsymbol{X}^q$, and $\boldsymbol{U}^p \in \mathbb{R}^{d_p \times d}$ and $\boldsymbol{U}^q \in \mathbb{R}^{d_q \times d}$ are low-rank projection matrices ($d < min(d_p, d_q)$). $\odot$ denotes element-wise multiplication (Hadamard product), and ρ denotes ReLU non-linearity. $\boldsymbol{p} \in \mathbb{R}^d$ projects the element interaction into a scalar, so that $\boldsymbol{A}^{pq} \in \mathbb{R}^{N \times N}$ can be normalized into a bilinear attention map by applying column-wise softmax. Then the output of the bilinear model is

$$\mathrm{BA}(\boldsymbol{X}^p, \boldsymbol{X}^q) = \boldsymbol{\mathcal{A}}^{pq}(\boldsymbol{X}^q\boldsymbol{W}^q). \tag{3}$$

In the matrix multiplication of $\boldsymbol{\mathcal{A}}^{pq}$ and $\boldsymbol{X}^q\boldsymbol{W}^q$, a relative position embedding [42] is injected to make the sequence-level interaction to be position-aware.

Channel-Level Interaction In order to modulate the sequence-level interaction result, we devise a gate function based on fine-grained channel-level interaction. We first obtain a channel representation of $\boldsymbol{X}^p$ and $\boldsymbol{X}^p$ as

$$\overline{\boldsymbol{X}}^p = \mathrm{Mean}_n(\boldsymbol{X}^p\boldsymbol{V}^p), \quad \overline{\boldsymbol{X}}^q = \mathrm{Mean}_n(\boldsymbol{X}^q\boldsymbol{V}^q), \tag{4}$$

[4] Details about this FFN can be found in the Supplementary Material.

where Mean$(\cdot)$ is sequence-wise mean-pooling, and $\boldsymbol{V}^p, \boldsymbol{V}^q$ are used to project $\boldsymbol{X}^p$ and $\boldsymbol{X}^p$ to lower dimension for efficient processing. Similarly, we also compute a channel-to-channel attention map

$$\boldsymbol{S}_{ij}^{pq} = f_{chn}\big(\overline{\boldsymbol{X}}_i^p, \overline{\boldsymbol{X}}_j^q\big), \quad \boldsymbol{\mathcal{S}}_{ij}^{pq} = \text{Softmax}_i(\boldsymbol{S}_{ij}^{pq}), \tag{5}$$

where $f_{chn}(\cdot)$ is a function for computing channel-level interaction. Since each element in $\overline{\boldsymbol{X}}^p$ and $\overline{\boldsymbol{X}}^q$ is a scalar, we simply use $f_{chn}(a, b) = -(a-b)^2$. Then the output of the gate function is

$$\text{CG}(\boldsymbol{X}^p, \boldsymbol{X}^q) = \sigma\big(\text{FFN}(\boldsymbol{X}^p \boldsymbol{\mathcal{S}}^{pq})\big), \tag{6}$$

where σ is the Sigmoid function, so the output has values in $[0, 1]$.

Modality-Wise Fusion. Given M modalities, there will be M^2 pairs of interacting modalities, and they are tiled as a high-dimensional tensor $\boldsymbol{X}^{MI} \in \mathbb{R}^{N \times M^2 \times d}$. The information in $\boldsymbol{X}^{MI}$ needs to be further aggregated before feeding it to target tasks. Simple concatenation or pooling can achieve this purpose. Here, we consider the importance of each interacting result by using a position-wise fully-connected layer to predict importance weights:

$$\begin{aligned} \boldsymbol{e}_n &= \boldsymbol{X}_n^{MI} \boldsymbol{W}_n^a + \boldsymbol{b}_n^a, \quad \boldsymbol{\alpha}_n = \text{Softmax}_m(\boldsymbol{e}_n), \\ \widehat{\boldsymbol{X}}_n &= \sum\nolimits_{m=1}^{M^2} \boldsymbol{\alpha}_{nm} \boldsymbol{X}_{nm}^{MI}. \end{aligned} \tag{7}$$

Finally, the fusion result $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N \times d}$ is the modality-interacted representation of a video and is ready to be used in target tasks.

3.3 Sentence Localization

The sentence is represented as a sequence of word-embedding vectors $Y = \{\boldsymbol{w}_l\}_{l=1}^L$, which is also processed with the CGMI to exploit its intra-modality information, yielding a textual feature $\widehat{\boldsymbol{Y}}$. For sentence localization, it is crucial to capture the complex association between the video and textual modalities at each temporal location, and then predict each location's relevance to the sentence.

Video-Text Local Interaction Based on the above intuition, we propose Video-Text Local Interaction. For each temporal location $t \in [1, N]$ of $\widehat{\boldsymbol{X}}$, a local window $\widetilde{\boldsymbol{X}} = \{\widehat{\boldsymbol{X}}_n\}_{n=t-w}^{t+w}$ is extracted to interact with the textual feature $\widehat{\boldsymbol{Y}}$. As shown in Fig. 1, the local video-to-text interaction is modeled as

$$\boldsymbol{Z}_t^{xy} = \text{BA}(\text{Mean}(\widetilde{\boldsymbol{X}}), \widehat{\boldsymbol{Y}}), \quad \widehat{\boldsymbol{Z}}_t^{xy} = \text{MM}(\boldsymbol{Z}_t^{xy}, \text{Mean}(\widetilde{\boldsymbol{X}})). \tag{8}$$

Here instead of gating, we use a more efficient multimodal processing unit $\text{MM}(a, b) = \boldsymbol{W}^T[a||b||a \odot b||a \oplus b]$ to encourage further interaction of both modalities. Likewise, text-to-video interaction $\widehat{\boldsymbol{Z}}_t^{yx}$ is computed given $\widetilde{\boldsymbol{X}}$ and Mean$(\widehat{\boldsymbol{Y}})$, and then fused with the video-to-text interaction result

$$\boldsymbol{Z}_t = \widehat{\boldsymbol{Z}}_t^{xy} \oplus \widehat{\boldsymbol{Z}}_t^{yx}. \tag{9}$$

Localization Head We apply a light-weight convolutional network upon the video-text interacted sequence $\boldsymbol{Z}$ to produce predictions. Each layer can be formulated as

$$\boldsymbol{C}^k = Conv(\boldsymbol{C}^{k-1}||\text{Mean}(\widehat{\boldsymbol{Y}})), \tag{10}$$

where $k = 1, .., K$, and $\boldsymbol{C}_0 = \boldsymbol{Z}$. We apply Instance Normalization [54] and LeakyReLU [66] activation to each layer's output. Since we are computing the video-text relevance in a layer-wise fashion, we impose an ℓ_2 norm regularization on each layer's output to obtain a more robust feature

$$\text{Loss}_{norm} = \sum_{n=1}^{N}(||\boldsymbol{C}_n^k||_2 - \beta_k)^2, \tag{11}$$

where $||\cdot||$ is the ℓ_2 norm of a vector. The K-th layer output $\boldsymbol{C}^K$ has 1 output channel, which is normalized using Softmax, representing the Video-Text Relevance $\boldsymbol{r} \in [0, 1]^N$. Then a fully connected layer with two output units is applied to $\boldsymbol{r}$ to produce a boundary prediction $\boldsymbol{b} \in \mathbb{R}^2$. The loss for the predictions is

$$\text{Loss}_{pred} = \text{Huber}(\boldsymbol{b} - \hat{\boldsymbol{b}}) - \lambda_r \frac{\sum_n \hat{\boldsymbol{r}}_n \log(\boldsymbol{r}_n)}{\sum_n \hat{\boldsymbol{r}}_n}, \tag{12}$$

where $\hat{\boldsymbol{b}}$ is the ground-truth temporal boundary, Huber$(\cdot)$ is the Huber loss function, and $\hat{\boldsymbol{r}}_n = 1$ if n is in the ground-truth temporal region, otherwise $\hat{\boldsymbol{r}}_n = 0$. The overall loss is

$$\text{Loss}_{loc} = \text{Loss}_{pred} + \lambda_n \text{Loss}_{norm}, \tag{13}$$

where λ_n, λ_r are constant weights used to balance the loss terms.

3.4 Event Captioning

After the video modality interaction result is obtained, we use a standard bi-directional LSTM for encoding and a two-layer LSTM network with temporal attention [71] to generate sentences as in previous works [59,9,69]. The sentence generation is done in a word-by-word fashion. At every time step, a set of temporal attention weights is computed based on the LSTM hidden states and video features, which is then used to weighted-sum the video features into a single vector. This dynamic feature vector is fed to the LSTM with the previously generated word to predict the next word[5]. We would like to emphasize again that video modality interaction can be used as a basic video feature encoding technique for either sentence localization or event captioning, but we do not perform multi-task training for these two.

[5] Due to the space limit and that caption decoder is not the focus of this work, we omit formal descriptions here. We also move some experiments and analysis below to the Supplementary Material.

4 Experiments

In this section, we provide experimental analysis of our model design and present comparisons with the state-of-the-art methods on both temporal sentence localization and video captioning.

4.1 Experimental Settings

MSVD Dataset [6]. MSVD is a well-known video captioning dataset with 1,970 videos. The average length is 9.6 seconds, and each video has around 40 sentence annotations on average. We adopt the same common dataset split as in prior works [71,69,4]. Thus, we have 1,200 / 100 / 670 videos for training, validation, and testing, respectively.

MSR-VTT Dataset [68]. MSR-VTT is a large-scale video captioning dataset with 10,000 videos. The standard split [68] for this dataset was provided. Hence, we use 6,513 / 497 / 2,990 videos for training, validation, and testing, respectively, in our experiments. In this dataset, each video is associated with 20 sentence annotations and is of lenght 14.9 seconds on average.

ActivityNet Captions Dataset [32](ANet-Cap). ANet-Cap is built on the ActivityNet dataset [22] with 19,994 untrimmed videos (153 seconds on average). The standard split is 10,009 / 4,917 / 5,068 videos for training, validation, and testing, respectively. There are 3.74 *temporally localized* sentences per video on average. Since the testing set is not publicly available, we evaluate our method on the validation set as previous works [62,67].

Charades-STA Dataset [17]. Charades-STA is built on 6,672 videos from the Charades [45] dataset. The average duration of the videos is 29.8 seconds. There are 16,128 *temporally localized* sentence annotations, which give 2.42 sentences per video. The training and testing sets contain 12,408 and 3,720 annotations, respectively.

We evaluate the captioning performance of our method on MSVD and MSR-VTT with commonly used metrics, *i.e.,* BLEU [38], METEOR [14], and CIDEr [56]. ANet-Cap and Charades-STA are used to evaluate sentence localization performance. We adopt the same evaluation metric used by previous works [17], which computes "Recall@1,IoU=m" (denoted by $r(m, s_i)$), meaning the percentage of the top-1 results having IoU larger than m with the annotated segment of a sentence s_i. The overall performance on a dataset of N sentences is the average score of all the sentences $\frac{1}{N}\sum_{i=1}^{N} r(m, s_i)$.

Implementation Details. The sentences in all datasets are converted to lowercase and then tokenized. For the captioning task, randomly-initialized word embedding vectors of dimension 512 are used, which are then jointly fine-tuned with the model. For the sentence localization task, we employ the GloVe [40] word embedding as previous works. We use Inception-ResNet v2 [50] and C3D [52] to extract visual and motion features. For the audio features, we employ the MFCC (Mel-Frequency Cepstral Coefficients) on the captioning task and SoundNet [2] on the sentence localization task. We temporally subsample the feature sequences to length 32 for event captioning, and 128 for sentence localization. The bilinear

Table 1. Performance comparison of video modality interaction strategies on MSVD.

#	Method	B@4	M	C
0	Concat w/o Interact (Baseline)	45.28	31.60	62.57
1	Concat + Interact	46.24	32.03	66.10
2	Pairwise Interact + Concat Fusion	47.86	33.73	75.30
3	Pairwise Interact + Sum Fusion	51.37	34.01	78.42
4	Pairwise Interact + Weighted Fusion (ours)	54.68	36.40	95.17
5	Intra-modality Interactions only	49.92	34.76	88.46
6	Inter-modality Interactions only	47.30	32.72	70.20
7	(Intra+Inter)-modality (ours)	54.68	36.40	95.17

Table 2. Performances (%) of different localizer settings on the Charades-STA dataset.

#	PMI	VTLI	ℓ_2-Norm	IoU=0.3	IoU=0.5	IoU=0.7
0	✗	✗	✗	51.46	35.34	15.81
1	✓	✗	✗	53.22	37.05	17.36
2	✓	✓	✗	54.37	38.42	18.63
3	✓	✓	✓	55.48	39.73	19.27

attention adopts 8 attention heads, and the loss weights λ_r and λ_n are set to 5 and 0.001, respectively. In all of our experiments, the batch size is set to 32 and the Adam optimizer with learning rate 0.0001 is used to train our model.

4.2 Ablation Studies

Firstly, we perform extensive experiments to validate the design choices in our approach. We study the effect of different modality interaction strategies on the MSVD dataset, and the effects of sentence localizer components on the Charades-STA dataset. All experiments use Inception-ResNet v2 and C3D features.

On the MSVD dataset, we design 8 different variants and their performances are summarized in Table 1. In variant 0, which is a baseline, multimodal features are concatenated and directly fed to the caption decoder. Variant 1 treats the concatenated features as one modality and performs intra-modality interaction. In variants 2-4, PMI is performed and different fusion strategies are adopted. In variants 5-7, we study the ablation of intra- and inter-modality interactions.

Why pairwise? We perform modality interaction in a pairwise fashion in our model, and this is the main distinctive difference from existing methods [48,72], which employ feature concatenation. As shown in Table 1, while concatenating all modalities into one and performing intra-modality interaction can gain performance improvements over the baseline (#1 vs. #0), concatenating after pairwise interaction has a more significant advantage (#2 vs. #1). We also compare the effects of different aggregation strategies after pairwise interaction (#2-4), and weighted fusion (in PMI) yields the best result with a clear margin, which also indicates that the interactions between different modality pairs produce unique information of different importances.

Effect of inter-modality complementarity. We then inspect the intra- and inter-modality interactions separately. Table 1 (#5-7) shows that intra-modality interaction can already effectively exploit information in each modality compared to the baseline. Inter-modality complementarity alone is not sufficient

Table 3. Video captioning performances of our proposed PMI and other state-of-the-art multimodal fusion methods on the MSVD dataset. Meanings of features can be found in Table 4.

Method	Features	B@4	M	C
AF [25]	V+C	52.4	32.0	68.8
TDDF [76]	V+C	45.8	33.3	73.0
MA-LSTM [69]	G+C	52.3	33.6	70.4
MFATT [36]	R152+C	50.8	33.2	69.4
GRU-EVE [1]	IRV2+C	47.9	35.0	78.1
XGating [59]	IRV2+I3D	52.5	34.1	88.7
HOCA [28]	IRV2+I3D	52.9	35.5	86.1
PMI-CAP	V+C	49.74	33.59	77.11
PMI-CAP	G+C	51.55	34.64	74.51
PMI-CAP	R152+C	52.07	34.34	77.35
PMI-CAP	IRV2+C	54.68	36.40	95.17
PMI-CAP	IRV2+I3D	55.76	36.63	95.68

Table 4. Performances of our proposed model and other state-of-the-art methods on the MSVD and MSR-VTT datasets. R*, G, V, C, IV4, R3D, IRV2, Obj, and A mean ResNet, GoogLeNet, VGGNet, C3D, Inception-V4, 3D ResNeXt, Inception-ResNet v2, Object features, and audio features, respectively. Note that audio track is only available on MSR-VTT, and for fair comparison, we use the MFCC audio representation as [9,11]. Please refer to the original papers for the detailed feature extraction settings.

Dataset	MSVD				MSR-VTT			
Method	Features	B@4	M	C	Features	B@4	M	C
STAT [53]	G+C+Obj	51.1	32.7	67.5	G+C+Obj	37.4	26.6	41.5
M^3 [63]	V+C	51.78	32.49	-	V+C	38.13	26.58	-
DenseLSTM [79]	V+C	50.4	32.9	72.6	V+C	38.1	26.6	42.8
PickNet [12]	R152	52.3	33.3	76.5	R152	41.3	27.7	44.1
hLSTMat [47]	R152	53.0	33.6	73.8	R152	38.3	26.3	-
VRE [44]	R152	51.7	34.3	86.7	R152+A	43.2	28.0	48.3
MARN [39]	R101+R3D	48.6	35.1	92.2	R101+R3D	40.4	28.1	47.1
OA-BTG [75]	R200+Obj	**56.9**	36.2	90.6	R200+Obj	41.4	28.2	46.9
RecNet [60]	IV4	52.3	34.1	80.3	IV4	39.1	26.6	42.7
XGating [59]	IRV2+I3D	52.5	34.1	88.7	IRV2+I3D	42.0	28.1	49.0
MM-TGM [11]	IRV2+C	48.76	34.36	80.45	IRV2+C+A	44.33	29.37	49.26
GRU-EVE [1]	IRV2+C	47.9	35.0	78.1	IRV2+C	38.3	28.4	48.1
MGSA [9]	IRV2+C	53.4	35.0	86.7	IRV2+C+A	**45.4**	28.6	50.1
PMI-CAP	IRV2+C	54.68	**36.40**	**95.17**	IRV2+C	42.17	28.79	49.45
PMI-CAP	-	-	-	-	IRV2+C+A	43.96	**29.56**	**50.66**

for captioning, but it can be combined with intra-modality information to obtain a further performance boost, which again validates our design of pairwise interaction.

Effect of sentence localizer components. The PMI, video-text local interaction (VTLI), and ℓ_2-norm regularization are the key components of the sentence localization model. As can be observed from Table 2, incorporating each component consistently leads to a performance boost.

4.3 Comparison with State-of-the-Art Methods

Results on the Video Event Captioning Task. We abbreviate our approach as PMI-CAP for video captioning. To demonstrate the superiority of our proposed pairwise modality interaction, we first compare our method with

Table 5. Performances (%) of our proposed model and other state-of-the-art methods on the Charades-STA dataset. * means our implementation.

Method	IoU=0.3	IoU=0.5	IoU=0.7
Random	14.16	6.05	1.59
VSA-RNN [29]	-	10.50	4.32
VSA-STV [29]	-	16.91	5.81
MCN [23]	32.59	11.67	2.63
ACRN [34]	38.06	20.26	7.64
ROLE [35]	37.68	21.74	7.82
SLTA [27]	38.96	22.81	8.25
CTRL [17]	-	23.63	8.89
VAL [49]	-	23.12	9.16
ACL [19]	-	30.48	12.20
SAP [10]	-	27.42	13.36
SM-RL [64]	-	24.36	11.17
QSPN [67]	54.7	35.6	15.8
ABLR* [74]	51.55	35.43	15.05
TripNet [20]	51.33	36.61	14.50
CBP [62]	-	36.80	18.87
PMI-LOC (C)	**55.48**	**39.73**	**19.27**
PMI-LOC (C+IRV2)	56.84	41.29	20.11
PMI-LOC (C+IRV2+A)	58.08	42.63	21.32

Table 6. Performances (%) of our proposed model and other state-of-the-art methods on the ActivityNet Captions dataset.

Method	IoU=0.3	IoU=0.5	IoU=0.7
Random	12.46	6.37	2.23
QSPN [67]	45.3	27.7	13.6
TGN [8]	43.81	27.93	-
ABLR [74]	55.67	36.79	-
TripNet [20]	48.42	32.19	13.93
CBP [62]	54.30	35.76	17.80
PMI-LOC (C)	**59.69**	**38.28**	**17.83**
PMI-LOC (C+IRV2)	60.16	39.16	18.02
PMI-LOC (C+IRV2+A)	61.22	40.07	18.29

state-of-the-art methods that focus on the fusion of multimodal features for video captioning. For fair comparison, we use the same set of features as each compared method. As shown in Table 3, our PMI-CAP has outperformed all the compared methods when using the same features. The improvement in the CIDEr metric is especially significant, which is 10.8% on average. This shows that our pairwise modality interaction can really utilize multimodal features more effectively.

Table 4 shows the performance comparison on the MSVD and MSR-VTT datasets. We adopt the set of features commonly used by recent state-of-the-art methods [59,1,9], which are Inception-ResNet v2 and C3D for visual and motion modalities, respectively. Among the competitive methods, OA-BTG [75] utilizes object-level information from an external detector, and MARN [39] uses a more advanced 3D CNN to extract motion features. We do not exploit spatial information like MGSA [9] and VRE [44], or use a sophisticated decoder as hLSTMat [47] and MM-TGM [11], while we emphasize that PMI may be used along with most of these methods. Overall, our PMI-CAP achieves state-of-the-art performances on both MSVD and MSR-VTT.

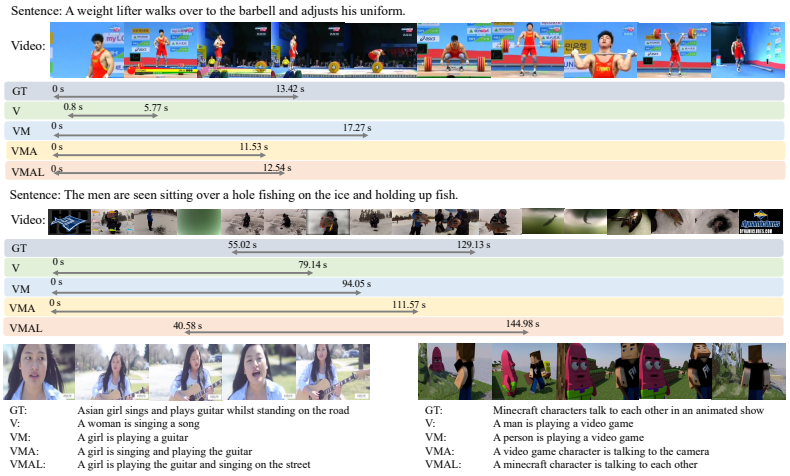

Fig. 3. Qualitative results of temporal sentence localization and event captioning. The results are generated using our model but with different combinations of modalities.

Results on the Sentence Localization Task. As previously introduced, current state-of-the-art methods for sentence localization haven't considered this problem in a multimodal setting and only use the C3D feature. Thus we present results with only C3D feature to fairly compare with these methods and also report performances under multimodal settings. Our approach is abbreviated as PMI-LOC for sentence localization. Table 5 shows results on the widely-used Charades-STA dataset. Our PMI-LOC outperforms all compared methods in all metrics. Further experiments with multimodal features show even higher localization accuracies, which verify the effectiveness of our modality interaction method. As shown in Table 6, on the large-scale ActivityNet Captions dataset, our method also achieves state-of-the-art performances.

4.4 Qualitative Results

We show some qualitative results in Figures 3, 4, and 5 to demonstrate the effectiveness of our modality interaction method and how it provides expainability to the final prediction of the target tasks. Note that in addition to the visual (V), motion (M), and audio (A) modalities, we also utilize the previously mentioned latent semantics (L) modality to comprehensively explore the video content.

Fig. 3 indicates that by utilizing more modalities, the model gets more complementary information through modality interaction and achieves better performance for both temporal sentence localization and event captioning. The event captioning examples in Fig. 4 show that each type of events has its modality interaction pattern. The sports video (top) has distinctive visual and motion patterns that are mainly captured by visual-motion modality interaction. The cooking video (middle) has unique visual cues and sounds made by kitchenware,

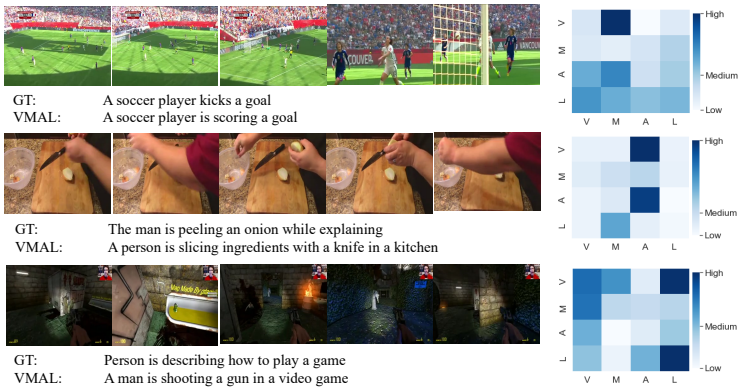

Fig. 4. Qualitative results of video event captioning with visualization of the modality importance weights.

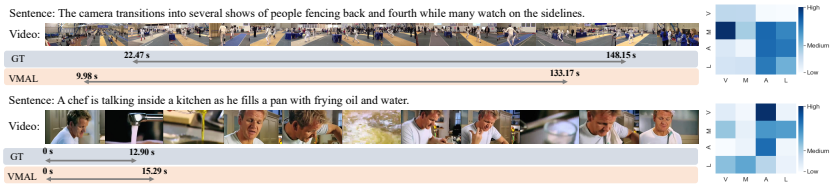

Fig. 5. Qualitative results of temporal sentence localization with visualization of the modality importance weights.

so the important interactions are between the visual and audio modalities and within the audio modality. For the animated video (bottom), latent semantics modality is important when the other modalities are not sufficient to capture its contents. Similar observations can also be made on the sentence localization examples in Fig. 5.

5 Conclusions

In this paper, we proposed pairwise modality interaction (PMI) for tackling the temporal sentence localization and event captioning tasks, and performed fine-grained cross-modal interactions in both the sequence and channel levels to better understand video contents. The extensive experiments on four benchmark datasets on both tasks consistently verify the effectiveness of our proposed method. Our future work will extend the proposed modality interaction method to cope with other video understanding tasks.

Acknowledgement

Shaoxiang Chen is partially supported by the Tencent Elite Internship program.

References

1. Aafaq, N., Akhtar, N., Liu, W., Gilani, S.Z., Mian, A.: Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: CVPR (2019)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NIPS (2016)
3. Baier, B., Kleinschmidt, A., Müller, N.G.: Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. Journal of Neuroscience **26**(47), 12260–12265 (2006)
4. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. In: CVPR (2017)
5. Calvert, G.A.: Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. Cerebral Cortex **11**(12), 1110–1123 (2001)
6. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: ACL (2011)
7. Chen, J., Pan, Y., Li, Y., Yao, T., Chao, H., Mei, T.: Temporal deformable convolutional encoder-decoder networks for video captioning. In: AAAI (2019)
8. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.: Temporally grounding natural sentence in video. In: EMNLP (2018)
9. Chen, S., Jiang, Y.: Motion guided spatial attention for video captioning. In: AAAI (2019)
10. Chen, S., Jiang, Y.: Semantic proposal for activity localization in videos via sentence query. In: AAAI (2019)
11. Chen, S., Chen, J., Jin, Q., Hauptmann, A.G.: Video captioning with guidance of multimodal latent topics. In: ACM MM (2017)
12. Chen, Y., Wang, S., Zhang, W., Huang, Q.: Less is more: Picking informative frames for video captioning. In: ECCV (2018)
13. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
14. Denkowski, M.J., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: WMT@ACL (2014)
15. Eckert, M.A., Kamdar, N.V., Chang, C.E., Beckmann, C.F., Greicius, M.D., Menon, V.: A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fmri connectivity analysis. Human brain mapping **29**(7), 848–857 (2008)
16. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP (2016)
17. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: ICCV (2017)
18. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C.H., Wang, X., Li, H.: Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: CVPR (2019)
19. Ge, R., Gao, J., Chen, K., Nevatia, R.: MAC: mining activity concepts for language-based temporal localization. In: WACV (2019)
20. Hahn, M., Kadav, A., Rehg, J.M., Graf, H.P.: Tripping through time: Efficient localization of activities in videos. arXiv preprint arXiv:1904.09936 (2019)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

22. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
23. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with natural language. In: ICCV (2017)
24. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
25. Hori, C., Hori, T., Lee, T., Zhang, Z., Harsham, B., Hershey, J.R., Marks, T.K., Sumi, K.: Attention-based multimodal fusion for video description. In: ICCV (2017)
26. Hu, Y., Chen, Z., Zha, Z., Wu, F.: Hierarchical global-local temporal modeling for video captioning. In: ACM MM (2019)
27. Jiang, B., Huang, X., Yang, C., Yuan, J.: Cross-modal video moment retrieval with spatial and language-temporal attention. In: ICMR (2019)
28. Jin, T., Huang, S., Li, Y., Zhang, Z.: Low-rank hoca: Efficient high-order cross-modal attention for video captioning. In: EMNLP-IJCNLP (2019)
29. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
30. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: NeurIPS (2018)
31. Kim, J., On, K.W., Lim, W., Kim, J., Ha, J., Zhang, B.: Hadamard product for low-rank bilinear pooling. In: ICLR (2017)
32. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: ICCV (2017)
33. Li, X., Zhao, B., Lu, X.: MAM-RNN: multi-level attention model based RNN for video captioning. In: IJCAI (2017)
34. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.: Attentive moment retrieval in videos. In: ACM SIGIR (2018)
35. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.: Efficient low-rank multimodal fusion with modality-specific factors. In: ACL (2018)
36. Long, X., Gan, C., de Melo, G.: Video captioning with multi-faceted attention. TACL **6**, 173–184 (2018)
37. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: CVPR (2016)
38. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
39. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.: Memory-attended recurrent network for video captioning. In: CVPR (2019)
40. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
41. Rahman, T., Xu, B., Sigal, L.: Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In: ICCV (2019)
42. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: NAACL-HLT (2018)
43. Shen, Z., Li, J., Su, Z., Li, M., Chen, Y., Jiang, Y., Xue, X.: Weakly supervised dense video captioning. In: CVPR (2017)
44. Shi, X., Cai, J., Joty, S.R., Gu, J.: Watch it twice: Video captioning with a refocused video encoder. In: ACM MM (2019)
45. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)

47. Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., Shen, H.T.: Hierarchical LSTM with adjusted temporal attention for video captioning. In: IJCAI (2017)
48. Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., Tang, J.: Autoint: Automatic feature interaction learning via self-attentive neural networks. In: CIKM (2019)
49. Song, X., Han, Y.: VAL: visual-attention action localizer. In: PCM (2018)
50. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
51. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
52. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
53. Tu, Y., Zhang, X., Liu, B., Yan, C.: Video description with spatial-temporal attention. In: ACM MM (2017)
54. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
56. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
57. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K.: Sequence to sequence - video to text. In: ICCV (2015)
58. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R.J., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: NAACL-HLT (2015)
59. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with POS sequence guidance based on gated fusion network. In: ICCV (2019)
60. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: CVPR (2018)
61. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning. In: CVPR (2018)
62. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: AAAI (2020)
63. Wang, J., Wang, W., Huang, Y., Wang, L., Tan, T.: M3: multimodal memory modelling for video captioning. In: CVPR (2018)
64. Wang, W., Huang, Y., Wang, L.: Language-driven temporal activity localization: A semantic matching reinforcement learning model. In: CVPR (2019)
65. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
66. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
67. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: AAAI (2019)
68. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016)
69. Xu, J., Yao, T., Zhang, Y., Mei, T.: Learning multimodal attention LSTM networks for video captioning. In: ACM MM (2017)
70. Yang, Z., Han, Y., Wang, Z.: Catching the temporal regions-of-interest for video captioning. In: ACM MM (2017)

71. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C.J., Larochelle, H., Courville, A.C.: Describing videos by exploiting temporal structure. In: ICCV (2015)
72. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR (2019)
73. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: CVPR (2016)
74. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI (2019)
75. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In: CVPR (2019)
76. Zhang, X., Gao, K., Zhang, Y., Zhang, D., Li, J., Tian, Q.: Task-driven dynamic fusion: Reducing ambiguity in video description. In: CVPR (2017)
77. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: CVPR (2018)
78. Zhu, L., Xu, Z., Yang, Y.: Bidirectional multirate reconstruction for temporal modeling in videos. In: CVPR (2017)
79. Zhu, Y., Jiang, S.: Attention-based densely connected LSTM for video captioning. In: ACM MM (2019)

Supplementary Material for: Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos

Shaoxiang Chen[1⋆], Wenhao Jiang[2], Wei Liu[2], and Yu-Gang Jiang[1⋆⋆]

[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University
[2] Tencent AI Lab
{sxchen13, ygj}@fudan.edu.cn, cswhjiang@gmail.com, wl2223@columbia.edu

1 Learning Latent Semantic Modality

Apart from the visual, motion, and audio modalities, which can be directly observed (apparent modalities), the latent semantics modality that carries high-level semantic information can be helpful for the language related tasks. We design a lightweight network to perform semantic attributes prediction using the sentence annotation provided by each dataset (either the video captioning or the sentence localization dataset). Note that this is a standalone task and the latent semantics modality is optional for our method.

The input of this network is the concatenation of all apparent modalities $\boldsymbol{X}^A = [\boldsymbol{X}^v||\boldsymbol{X}^m||\boldsymbol{X}^a]$, where $\boldsymbol{X}^A \in \mathbb{R}^{N\times(d_a+d_m+d_v)}$. We simply process $\boldsymbol{X}^A$ using bidirectional LSTMs and concatenate the hidden states of each LSTM:

$$\boldsymbol{X}^l = [\overrightarrow{\text{LSTM}}(\boldsymbol{X}^A)||\overleftarrow{\text{LSTM}}(\boldsymbol{X}^A)], \tag{1}$$

where $\boldsymbol{X}^l \in \mathbb{R}^{N\times 2d_{hid}}$, and $\overrightarrow{\text{LSTM}}(\cdot)$ and $\overleftarrow{\text{LSTM}}(\cdot)$ denote the LSTM networks that have d_{hid} units and process their input sequences in the forward and backward directions, respectively. $\boldsymbol{X}^l$ is then passed through a fully-connected layer with sigmoid activation to predict semantic attribute probabilities:

$$\boldsymbol{P} = \text{sigmoid}(\boldsymbol{X}^l\boldsymbol{W}_c + \boldsymbol{b}_c), \tag{2}$$

where $\boldsymbol{W}_c$ and $\boldsymbol{b}_c$ are parameters, $\boldsymbol{P} \in \mathbb{R}^{C\times N}$ collects the temporal semantic attributes, and C is the vocabulary size of predefined attributes.

To train this network, we construct labels from the sentence annotations in event captioning or sentence localization datasets. We first process the training sentences of a dataset, select the most frequent C words that are noun or verb, and lemmatize them to form an attribute vocabulary. Then each sentence can

⋆ Part of the work is done when the author was an intern at Tencent AI Lab.
⋆⋆ Corresponding author.

be converted to a one-hot label $\boldsymbol{l} \in \mathbb{R}^C$ according to whether its words are in the vocabulary, where $\boldsymbol{l}_c = 1$ indicates that attribute c is present in the sentence, otherwise $\boldsymbol{l}_c = 0$. The label $\boldsymbol{l}$ is broadcast to the N temporal locations to compute the cross entropy loss at each location:

$$\mathcal{L}_{ce} = -\frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{l}_c \ln \boldsymbol{P}_c + (1-\boldsymbol{l}_c)\ln(1-\boldsymbol{P}_c)), \tag{3}$$

where $\mathcal{L}_{ce} \in \mathbb{R}^N$. For the sentence localization task, the sentence annotations are usually available for temporal segments. To unify the loss representations, we construct a temporal mask $\boldsymbol{M}^{tcp} \in [0,1]^N$ defined as:

$$\boldsymbol{M}_i^{tcp} = \begin{cases} 1 & \text{if } i \in [s,e] \text{ and } rand(0,1) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $[s,e]$ is the temporal segment of the sentence annotation normalized to be in $[0,1]$. In event captioning it is safe to assume $s=0$ and $e=1$ since the videos are relatively short. Randomness is introduced in $\boldsymbol{M}^{tcp}$ to prevent overfitting. The final temporal semantic attributes prediction loss is computed as:

$$\mathcal{L}_{tcp} = \frac{1}{N}\mathcal{L}_{ce} \cdot \boldsymbol{M}^{tcp}, \tag{5}$$

where $\cdot$ is the dot product operator. From the above description, we can see that when the network learns to predict attributes, $\boldsymbol{X}^l$ carries rich information of latent semantics for every temporal location. Thus, it can be used to assist our target tasks through interacting with other modalities.

2 Feed-Forward Network (FFN)

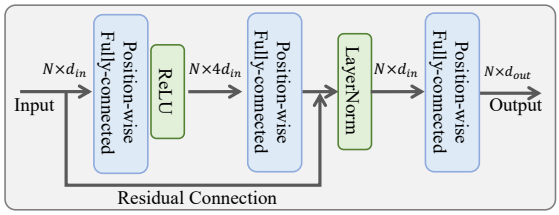

Fig. 1. The structure of the Feed-Forward Network (FNN).

As shown in Fig. 1, FFN is mainly composed of three position-wise fully-connected layers, each of which basically applies a fully-connected layer to each element of the input feature sequence with shared parameters. ReLU activation and layer normalization are applied to the first and second layers, respectively, and the initial input is connected to the second layer's output via a residual connection to encourage gradient flow. The output dimension d_{out} is decided according to the input ($d_{out} \leq d_{in}$).

3 More on Motivation

Our motivation is two-fold (Note that the Equations, Tables, and Figures mentioned in this section are all in the original paper):

(1) It is intuitive that both human and AI models understand events better via a combination of different sensory modalities, but the importances of different modalities vary among videos as well as among the moments inside a video. This motivates us to fuse the modality-interacted tensor by considering both the modality-wise and sequence-wise importances (Eq. (7)).

(2) Neuroscience researches have proven that information processing in one modality can affect another, which means that there are interactions among modalities and complementary information may communicate through such interactions. This motivates us to design sequence- and channel-level interactions for each pair of modalities. In the sequence-level interaction, each element from one sequence interacts with all the elements in the other sequence through the bilinear model (Eq. (2)). This fully-connected information flow between two modalities enables better utilization of complementary information than traditional fusion strategies as shown in Table 1. It is also widely accepted that different feature channels capture different information. Thus the goal of channel-level interaction is to emphasize important channels, which is realized by gating. The gate variable is computed via a channel-to-channel attention mechanism, and sequence-wise mean-pooling (Eq. (4)) is for reducing computation. The gating power is demonstrated by the experiment below in Section 6. The improvement brought by channel-level interaction is not as significant as sequence-level interaction, but it is indeed effective.

Based on the motivation, our goal is finding a better combination of modalities via fine-grained interaction. Attention is the building block we adopted to achieve this goal, because it is easy to understand and implement (also yields a clear framework). Finally, we have proven our modality interaction to be both effective and able to provide explainability (see Figs. 4 and 5).

4 Computational Complexity

Table 1. PMI-CAP's running times on one RTX 2080Ti GPU.

Mode	Memory	Time/batch
Train (batch size=32)	5939MB	0.38s
Infer (batch size=1)	1441MB	0.08s

The major computational cost is from sequence-level interaction, which mainly consists of feature projection and bilinear modeling (Eqs. (2) and (3) in the original paper). Assume that a pair of interacting feature sequences both have dimension $b \times n \times d$, where b and n stand for batch size and sequence length,

Table 2. Performances of PMI combined with other methods of target tasks.

Method	B@4	M	C
Masked Transformer [3]	47.49	32.43	77.35
Masked Transformer [3]+PMI	50.95	35.20	86.61
Method	IoU=0.3	IoU=0.5	IoU=0.7
ABLR [2]	53.55	37.47	16.21
ABLR [2]+PMI	55.26	39.52	16.88

Table 3. Performance comparison on the ActivityNet Captions dataset.

Method	B@4	M	C
vanilla-CAP (IRV2+I3D)	1.75	10.14	40.63
PMI-CAP (IRV2+I3D)	1.99	10.89	43.56
PMI-CAP (IRV2+I3D+A)	2.31	11.00	51.30
PMI-CAP/no-channel (IRV2+I3D)	2.00	10.52	43.06
2019 Rank-1 Intra-Event	3.91	11.96	49.56

respectively. Then the computational complexity is $O(bnd^2 + bn^2d)$. For short videos, since $n \ll d$, the complexity becomes $O(bnd^2)$ (mainly batch matrix multiplication) and is efficient to run on GPU. While for very long videos like TV shows the $O(bn^2d)$ term is dominant and the computational cost would grow quadratically with video length. Nonetheless, reducing the quadratic complexity for very long videos is out of scope of this work and is left for future work. Actual running times of PMI-CAP are shown in Table 1.

5 Compatibility with Other Models

We also test the effectiveness of our proposed PMI when combined with other types of architectures for event captioning or sentence localization. Note that the original methods used concatenated features [3] or single feature [2] as inputs, our implementations concatenate multimodal features for both methods. Results are presented in Table 2. For video captioning, we adopt the Masked-Transformer model [3] which is essentially different from RNN-based caption decoders. We use PMI to encode the multimodal features as its input and a substantial performance improvement over feature concatenation is obtained. We combine PMI with the state-of-the-art RNN-based sentence localization method ABLR [2] by inserting our PMI module between the feature extraction and Bi-LSTM feature encoding of ABLR, and a clear performance gain is also observed.

6 Captioning Performances on ActivityNet Captions

We further evaluate several variants of our PMI-CAP on the ActivityNet Captions and compare them with the 2019 ActivityNet captioning challenge winner [1], which used a more diverse set of features (e.g., objects and contexts) in

addition to the three common modalities. Following the official evaluation protocol, we compare the performances of captioning ground-truth event proposals on the validation set. The results are shown in Table 3. The vanilla-CAP method removes PMI and uses feature concatenation instead. The channel-level interaction and gating are disabled in the "no-channel" setting. As can be observed from the top four rows, our proposed method is consistently effective on ActivityNet. It is notable that our method can also achieve comparable performances with the challenge winner despite using fewer features.

References

1. Chen, S., Song, Y., Zhao, Y., Jin, Q., Zeng, Z., Liu, B., Fu, J., Hauptmann, A.: Activitynet 2019 task 3: Exploring contexts for dense captioning events in videos. arXiv preprint arXiv:1907.05092 (2019)
2. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI (2019)
3. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: CVPR (2018)