This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

I^{2}Transformer: Intra- and Inter-relation Embedding Transformer for TV Show Captioning

Yunbin Tu, Liang Li, Li Su, *Member, IEEE,* Shengxiang Gao, Chenggang Yan, Zhengjun Zha, Zhengtao Yu, Qingming Huang, *Fellow, IEEE*

***Abstract*—TV show captioning aims to generate a linguistic sentence based on the video and its associated subtitle. Compared to purely video-based captioning, the subtitle can provide the captioning model with useful semantic clues such as actors' sentiments and intentions. However, the effective use of subtitle is also very challenging, because it is the pieces of scrappy information and has semantic gap with visual modality. To organize the scrappy information together and yield a powerful omni-representation for all the modalities, an efficient captioning model requires understanding video contents, subtitle semantics, and the relations in between. In this paper, we propose an Intra- and Inter-relation Embedding Transformer (I^{2}Transformer), consisting of an Intra-relation Embedding Block (IAE) and an Inter-relation Embedding Block (IEE) under the framework of a Transformer. First, the IAE captures the intra-relation in each modality via constructing the learnable graphs. Then, IEE learns the cross attention gates, and selects useful information from each modality based on their inter-relations, so as to derive the omni-representation as the input to the Transformer. Experimental results on the public dataset show that the I^{2}Transformer achieves the state-of-the-art performance. We also evaluate the effectiveness of the IAE and IEE on two other relevant tasks of video with text inputs, *i.e.*, TV show retrieval and video-guided machine translation. The encouraging performance further validates that the IAE and IEE blocks have a good generalization ability. The code will be released in the future.**

***Index Terms*—TV Show captioning, video and subtitle, intra-relation embedding, inter-relation embedding, Transformer.**

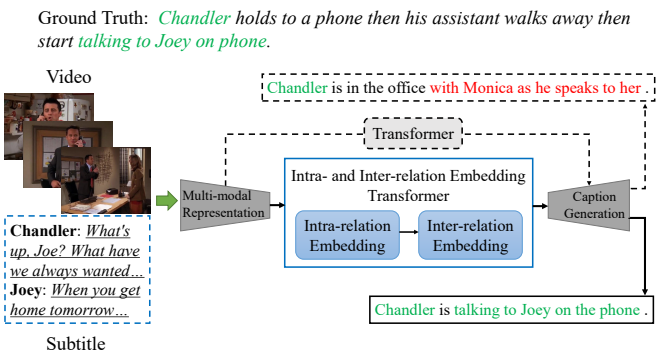


Fig. 1. An example of TV show caption generation with and w/o the proposed I^{2}Transformer. The dotted line denotes the caption generated by the state-of-the-art MMT [7] method, and the solid line denotes the caption generated by our proposed method. The words in green and red color respectively denote correct and incorrect words with respect to the ground truth words.

I. Introduction

AUTOMATICALLY describing TV show videos and their associated subtitles is a new research direction in the community of image captioning [1], [2], [3] and video captioning [4], [5], [6]. Compared to purely video-based captioning [4], [5], [6], on one hand, video+subtitle captioning is able to generate high-level linguistic sentences rather than simple descriptions of visual content, because subtitles can provide a captioning model with some implicit but very useful semantic clues to explain actors' sentiments and intentions. In this way, video+subtitle captioning is able to give the viewers better experiences when they browse and retrieval the video content. On the other hand, video+subtitle captioning also faces the challenges of 1) organizing the pieces of scrappy information together from subtitles and 2) bridging the semantic gap among different modalities so as to learn an omni-representation of both video and subtitle for caption generation. Thus, an efficient captioning model should understand video contents, subtitle semantics, and, most importantly, the relations in between.

As the pioneer work for this task, Lei *et al.* [7] first represented each video and its subtitle by appearance, motion, and text modality, respectively. Then, they directly concatenated

The work was supported by National Natural Science Foundation of China (Grant Nos. 61761026, 61972186, 61732005, 61762056, 61771457, 61732007), National Key Research and Development Plan of China (Grant Nos. 2018YFE0303104, 2019QY1802, 2019QY1801, and 2019QY1800), Youth Innovation Promotion Association of Chinese Academy of Sciences (Grant No. 2020108), Yunnan high-tech industry development project (Grant No. 201606), Yunnan provincial major science and technology special plan projects (Grant No. 202002AD080001-5), and Yunnan Basic Research Project (Grant Nos. 202001AS070014, 2018FB104). (*Corresponding author: Liang Li; Shengxiang Gao.*)

Yunbin Tu, Shengxiang Gao, and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with the Yunnan Provincial Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: tuyunbin1995@foxmail.com; gaoshengxiang.yn@foxmail.com; ztyu@hotmail.com).

Liang Li is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: liang.li@ict.ac.cn)

Li Su is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: suli@ucas.ac.cn).

Chenggang Yan is with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: cgyan@hdu.edu.cn).

Zhengjun Zha is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230052, China (e-mail: zhazj@ustc.edu.cn).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China, with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the Peng Cheng Laboratory, Shen Zhen 518057, China (e-mail: qmhuang@ucas.ac.cn).

all modalities as the input to a vanilla Transformer [8] for caption generation. Later, Li *et al* [9] concatenated appearance and motion modalities as visual modality. Then, they modeled the inter-relation between visual and text modality by cross-attention mechanism [8]. However, in both works, the intra-relation in each modality and inter-relations in cross-modalities are ignored or insufficiently learned during the fusion of video and subtitle. In this case, it is difficult for the captioning model to understand the semantics of each modality and the semantic interactions in between.

For a video, there may exist irrelevant information in certain frames, but its main content is supposed to focus on one event. To this end, it is necessary to learn the intra-relation of semantics for appearance and motion modalities, which would help distinguish irrelevant information from the main content. For subtitles, they mainly consist of dialogues which are always scrappy information, so the captioning model would be difficult to learn its meaning. Yao *et al.* [10] have shown that each word representation can be induced from other words in a sentence. Hence, it is beneficial to summarize the main content of subtitles by learning semantic relations between words.

Besides, simple feature concatenation cannot make full use of the inter-relations in cross modalities [11], [12]. Thus, learning the inter-relations in cross modalities is also important. In a video, static entities and dynamic actions can be represented by appearance and motion modality. If directly concatenating them, the semantic interactions between entities and actions will be ignored. In terms of subtitles, they mainly consist of dialogues of actors, which are very useful information to convey actors' intentions and sentiments. Hence, each modality not only has the specific meaning itself, but also can be as supplementary information for the others. For instance, as shown in Fig. 1, the visual entity "person" and action "talking" can be easily represented by appearance and motion modality, respectively. As for the key detail "phone" which is too tiny to recognize, it can be inferred by the clue from their dialogues. In this case, if the inter-relations in cross modalities cannot be fully exploited, the captioner will generate an inaccurate caption (in the dotted line box of Fig. 1). Hence, the inter-relations in cross modalities play crucial roles in thoroughly understanding the video content. To this end, it is necessary to build the intra-relation in each modality and inter-relations in cross modalities when obtaining the omni-representation for both video and subtitle.

In this paper, we propose a novel Intra- and Inter-relation Embedding Transformer (I^2Transformer) for TV show captioning (TVC). The model consists of three blocks: an Intra-relation Embedding block (IAE), an Inter-relation Embedding block (IEE) and a standard Transformer. Specifically, in the IAE, we first formulate every frame in a video and every word in a subtitle to be a node, so we can gain three kinds of node representations, *i.e.*, appearance, motion, and text nodes. Then, we build three learnable fully-connected graphs to learn the semantic relations between node representations in each modality. With the aid of graph convolutional networks [13], the node representation of each modality can be enhanced during the process of relation reasoning. In the IEE, we first design the cross attention gates via the sigmoid activation function to determine the relevance among different modalities. Then, based on them, we mine useful information from each modality to generate the omni-representation for video and subtitle. Finally, the learned omni-representation is fed into the Transformer for caption generation. When the intra- and inter-relations are embedded, this omni-representation can clearly represent entities and actions in videos, as well as intentions and sentimental information in subtitles.

The contributions of this work are summarized as follows:

- We propose a novel I^2Transformer for generating the TV show caption, where the omni-representation for both video and subtitle is learned in the fusion process of multiple modalities.
- Both IAE and IEE blocks are introduced to learn the intra-relation in video and subtitle, as well as the inter-relations in between. This is beneficial to understand the semantics of each modality and the semantic interactions in cross modalities.
- Extensive experiments on the TVC dataset show that our approach can achieve state-of-the-art performance. Meanwhile, we also evaluate the effectiveness of IAE and IEE blocks for the TV show retrieval (TVR) and video-guided machine translation (VMT) tasks, and the experimental results demonstrate that they have a good generalization ability.

The remainder of this paper is organized as follows. In Section II, we first review some relevant research works. In Section III, we introduce the overall framework of our proposed method and describe each block in detail. In Section IV, we elaborate the used datasets, evluation metrics, experimental procedure, quantitative analysis, and qualitative analysis. In Section V, we come to the conclusion and make a discussion about future research of this work.

II. Related Works

In this section, we will first briefly review the previous works for video captioning and video+subtitle captioning (TV show captioning). Then, we will introduce the captioning works focusing on multi-modal fusion. Finally, we introduce the use of graph neural networks in captioning.

A. Video Captioning

The task of video captioning has been flourishing in the community of multi-modal learning these years, because it connects computer vision [14], [15] and natural language processing [8], [16] which are two important applications in Artificial Intelligence.

The methods for video captioning can be classified into two dimensions: (1) template-based methods and (2) encoder-decoder-based methods.

1) Templated-based Methods: In the early years, the template-based method [17], [18] is a common strategy for video captioning. Concretely, this pipeline first exploits different kinds of classification methods to predict a set of visual concepts, such as objects, relationships, and attributes. Then, based on the pre-defined sentence template and basic grammar

rules (*e.g.*, subjects-verbs-objects), a video caption is yield by organizing these pre-detected concepts. The advantage of this method is intuitive. However, its disadvantage is also apparent. Due to the limitation of pre-defined templates, this method is inflexible to generate diverse and meaningful sentences.

2) Encoder-decoder-based Methods: This kind of method [19], [20], [21], [22], [23], [24], [25] is inspired by the idea of neural machine translation [16] to translate a short video into a sentence. Generally, a pre-trained convolutional neural network (CNN) is first used to extract video features from corresponding a sequence of video frames. Then, either the LSTM-based recurrent neural network [26] or Transformer [8] is leveraged as the encoder to model the temporal dependencies of features (by the LSTM) or feature correlations (by the Transformer), and as the decoder to transform encoded features into target words. Recently, dense video captioning [27], [28], [29], [30] has attracted more and more attentions. Instead of describing short videos, this task aims to detect multiple events that occur in a long video, and describe each event with a natural language sentence. However, both tasks belong to purely video-based captioning and do not use associated subtitles that are very useful information to explain actors' intentions.

B. Video+Subtitle Captioning

Recently, a new task called TV show captioning (TVC) [7] mitigates the limitation of previous tasks by reserving the subtitle of each video. Compared to purely video-based captioning, TVC is more challenging, because it requires simultaneously dealing with video and subtitle, which are totally different modalities. Especially for subtitles, on one hand, they are able to convey much implicit information such as actors' intentions and sentiments, which can not only augment the visual modality, but also improve the cognition ability of the captioning model. On the other hand, subtitles originally derive from actors' dialogues, so they are a set of scrappy text information and need to reorganize.

The methods of Lei *et al.* [7] and Li *et al.* [9] are the pioneers for this task. Following the encoder-decoder framework of conventional video captioning, both methods leveraged a pre-trained CNN to extract video features, a pre-trained or trainable word embedding model to represent subtitle information, and a standard Transformer to encode and decoder the omni-representation of both video and subtitle. However, Lei *et al.* [7] and Li *et al* [9] mainly projected different modalities into the same dimensional space and then concatenated them into an omni-representation. In their approaches, the intra-relation in each modality and inter-relations in cross-modalities are ignored or insufficiently modeled. Therefore, the information of each modality cannot be fully mined, and the semantic gap among different modalities still exists. In contrast, to better address the TVC, this paper focuses on mining the intra-relation in each input modality and inter-relations in between when fusing different modalities.

C. Multi-Modal Fusion

The main challenge in video captioning is how to fuse different kinds of modalities effectively. Therefore, multi-modal fusion strategies are gaining popularity for video captioning. On one hand, early works [11], [12] focused on the fusion of visual modalities. Specifically, Zhang *et al.* [11] presented a task-driven dynamic fusion method that learns to utilize specific task status to dynamically choose different kinds of visual modalities (*i.e.,* appearance, motion and their combination). Wang *et al.* [12] devised a cross gating strategy to capture the inter-relation between the appearance and motion modalities. On the other hand, previous works [31], [32], [33], [34] explored to fuse both visual and audio modalities. Concretely, Hori *et al.* [31] and Rahman *et al.* [34] both presented an attention-based multi-modal fusion model to integrate both audio and video information. Wang *et al.* [33] proposed a hierarchical encoder-decoder to fuse both global and local contexts of each modality. Hao *et al.* [32] devised three kinds of multi-modal fusion approaches to fuse both visual and audio modalities. Compared to the aforementioned methods, there are two apparent differences in our work. First, we focus on the multi-modal fusion based on the video and subtitle, *i.e.*, appearance, motion and text modalities. Second, the aforementioned methods mainly built inter-relations in cross modalities, while not fully exploring the intra-relations in each modality. By comparison, we aim to model the both relations in the fusion of video and subtitle.

D. Captioning with Graph Neural Networks

Recently, graph neural networks (GNNs) have been introduced in image or video captioning tasks and shown effectiveness for capturing relations between different nodes. For image captioning, Wang *et al.* [35] exploited GNNs to implicitly build relations between object nodes in an image. Yao *et al.* [36] first constructed graphs between the object nodes in an image based on their spatial and semantic connections. Then, they exploited graph convolutional networks (GCNs) to refine each node representation. For the purely video-based captioning task, Zhang *et al.* [37] and Pan *et al.* [38] proposed similar methods of building an object relational graph to learn their spatial and temporal relations, and then using GCNs to update the graph representations. Different from their works only building graph structure in a single modality, in this paper, we mainly focus on how to extend GNNs to learn semantic relations in each modality of video and subtitle for TV show captioning.

III. Methodology

In this section, we introduce the proposed intra- and inter-relation embedding Transformer (I^2Transformer). Its highlight is to learn the intra-relation in each modality and inter-relations in cross modalities when fusing them, as shown in Fig. 2. We begin with an introduction to the multi-modal representation of video and subtitle. Then we elaborate the intra-relation embedding block and the inter-relation embedding block. Finally, we describe the caption generation based on the Transformer.

A. Multi-modal representation

In our work, a given video is represented by appearance and motion modality, respectively. Specifically, a video is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

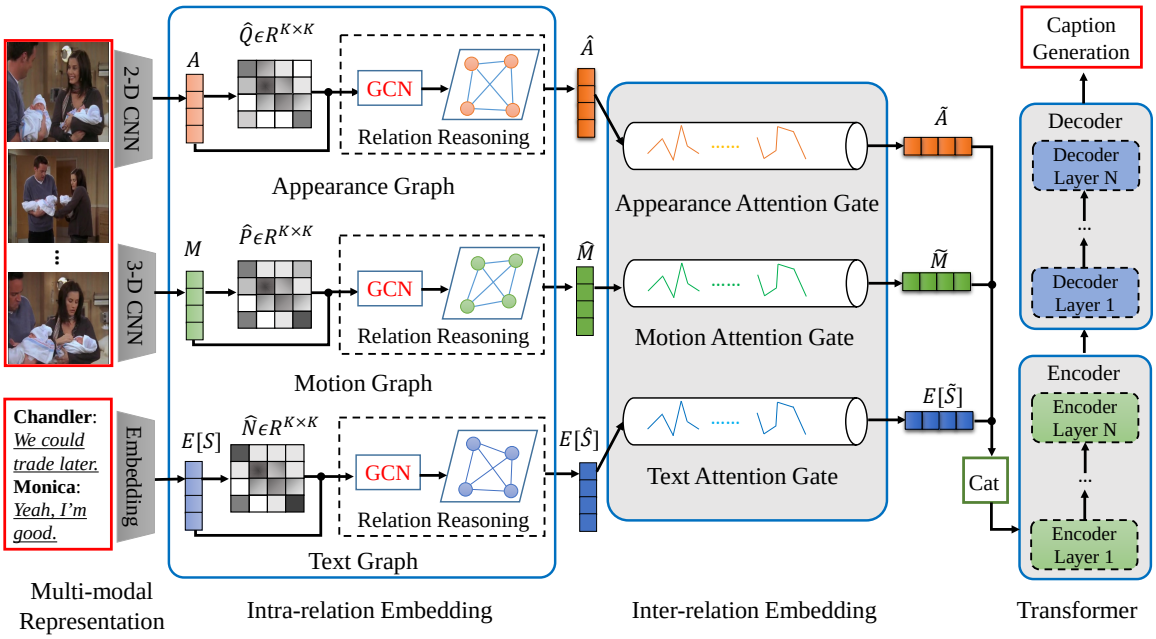


Fig. 2. The architecture of our intra- and inter-relation embedding Transformer (I^2Transformer). It consists of an intra-relation embedding block, an inter-relation embedding block and Transformer. "Cat" refers to the concatenation operation. More details about inter-relation embedding block and the Transformer are shown in Fig. 3 and Fig. 4, respectively.

first uniformly sampled as k frames $F = \{f_1, ..., f_k\}$. Then, a pre-trained 2-D CNN is utilized to extract appearance features one by one from k frames, which are denoted as $A = \{a_1, a_2, ..., a_k\}$ where each $a_i \in \mathbb{R}^{da}$. Next, we exploit 3-D CNN to extract each motion feature from a short-range consecutive frames. We denote them as $M = \{m_1, m_2, ..., m_k\}$ where each $m_j \in \mathbb{R}^{dm}$. Please note that in a video, the number of motion features is shorter than appearance features, we pad zeros in the shortage part to keep the same length as appearance features. In terms of subtitles $S = \{s_1, ..., s_{k'}\}$, they are represented by the word embedding features (text modality) with a trainable word embedding matrix E. We denote them as $E[S] = \{E[s_1], ..., E[s_{k'}]\}$ where each $E[s_v] \in \mathbb{R}^{ds}$.

For convenient computation and fair comparison, we follow the pioneer work [7] to set three kinds of features as the same length $K = k + k'$. For appearance and motion features, the shortage part is padded with zeros, respectively. For word features, the shortage part is padded with the embedding features of a special token "VID".

B. Intra- and Inter-relation Embedding

As illustrated in Fig. 2, the intra-relation embedding block (IAE) first leverages three learnable fully-connected graphs to build the intra-relation in appearance, motion and text modality, respectively. Then, the inter-relation embedding block (IEE) exploits cross attention gates to learn inter-relations among the modalities and select useful information from each modality to generate an omni-representation, which is able to make each modality build the inter-relation with other modalities and embed cross-modal information as supplement during multi-modal semantic interactions.

1) Intra-relation Embedding (IAE):

Appearance Graph. In order to model the semantic relations in the appearance modality, we construct an appearance graph for a set of appearance features and then use it to update them. Specifically, given K appearance features, each feature is regarded as a node. Let $A \in \mathbb{R}^{K \times da}$ denote K appearance nodes with da dimensional feature. We denote $\hat{Q} \in \mathbb{R}^{K \times K}$ as the relation coefficient matrix among K nodes, which is defined as follows:

$$Q = \phi(A) \cdot \phi(A)^\top, \tag{1}$$

$$\phi(A) = \mathrm{ReLU}(AW_a + b_a), \tag{2}$$

where $W_a \in \mathbb{R}^{da \times d}$ and $b_a \in \mathbb{R}^d$ are the transformation matrices and bias terms. Then, Q is normalized to make the sum of edges, connecting to the same node, equal to 1:

$$\hat{Q} = \mathrm{softmax}(Q, \mathrm{dim} = 1), \tag{3}$$

where softmax(, dim=1) denotes performing softmax at the second dimension of the input. $\hat{Q}$ represents how much information each appearance node obtains from the surrounding nodes.

Afterward, we utilize the GCN to perform relation reasoning, so the original appearance features A are updated to $\hat{A}$:

$$\hat{A} = \mathrm{ReLU}\left[\left(\phi(A) + \hat{Q} \cdot \phi(A)\right) W_{a'} + b_{a'}\right], \tag{4}$$

where $W_{a'} \in \mathbb{R}^{d \times d}$ and $b_{a'} \in \mathbb{R}^d$ are the parameters to be learned.

Motion Graph. A motion graph is devised to model the intra-relation in motion modality. Specifically, given K motion features, each feature is considered as a node. Let $M \in \mathbb{R}^{K \times dm}$ denote K motion nodes with dm dimensional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

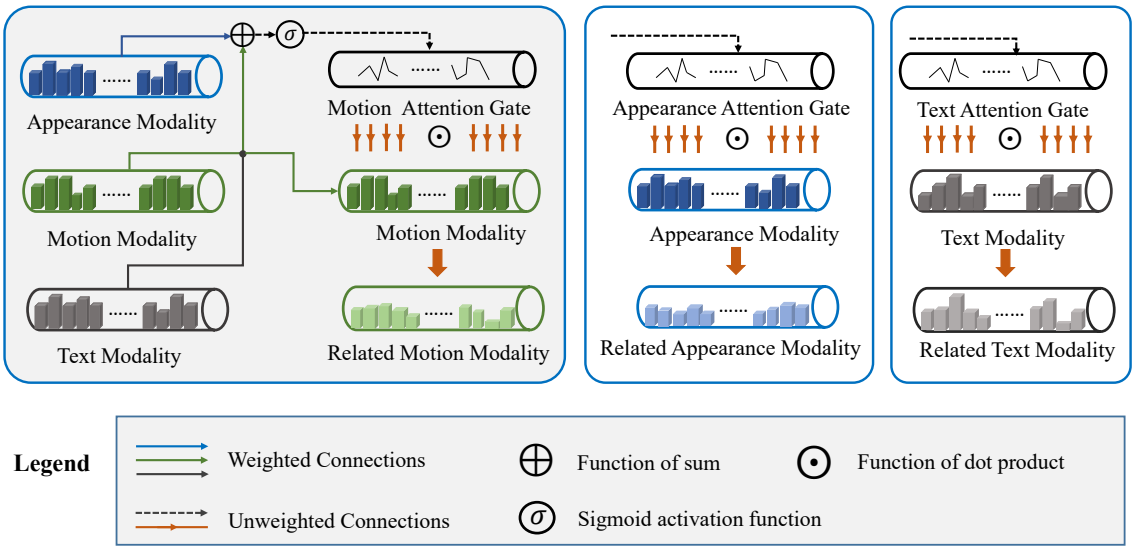


Fig. 3. The details of the proposed inter-relation embedding block. We illustrate how to learn a motion attention gate from all the independent modalities. The appearance and text attention gates are learned with the similar operation but with different parameters.

feature and $\hat{P} \in \mathbb{R}^{K\times K}$ denotes the relation coefficient matrix among K nodes. $\hat{P}$ is defined as follows:

$$P = \varphi(M) \cdot \varphi(M)^\top, \tag{5}$$

$$\varphi(M) = \mathrm{ReLU}(MW_m + b_m), \tag{6}$$

where $W_m \in \mathbb{R}^{dm\times d}$ and $b_m \in \mathbb{R}^d$ are the transformation matrices and bias terms. Then, P is normalized via the softmax function:

$$\hat{P} = \mathrm{softmax}(P, \mathrm{dim} = 1). \tag{7}$$

Afterward, we utilize the GCN to perform relation reasoning, so the original motion features M are updated to $\hat{M}$:

$$\hat{M} = \mathrm{ReLU}\left[\left(\varphi(M) + \hat{P} \cdot \varphi(M)\right) W_{m'} + b_{m'}\right], \tag{8}$$

where $W_{m'} \in \mathbb{R}^{d\times d}$ and $b_{m'} \in \mathbb{R}^d$ are the parameters to be learned.

Text Graph. The associated subtitle of each video can convey the actors' sentiments and intentions so as to help the model generate the high-level linguistic sentence. However, the effective use of subtitle is very challenging, because it is the pieces of scrappy information. In practice, two words in a sentence usually have certain relations and each word can be induced from all the other words [39], [10]. Thus, we customize a text graph to learn the intra-relations among subtitle words. Specifically, given K word embedding features, we consider each of them to be a node. Let $E[S] \in \mathbb{R}^{K\times ds}$ denote K word nodes with ds dimensional feature and $\hat{N} \in \mathbb{R}^{K\times K}$ denotes the relation coefficient matrix among K nodes. $\hat{N}$ is defined as follows:

$$N = \psi(E[S]) \cdot \psi(E[S])^\top, \tag{9}$$

$$\psi(E[S]) = \mathrm{ReLU}(E[S]W_s + b_s), \tag{10}$$

where $W_s \in \mathbb{R}^{ds\times d}$ and $b_s \in \mathbb{R}^d$ are the transformation matrices and bias terms. Then, N is normalized by the softmax function:

$$\hat{N} = \mathrm{softmax}(N, \mathrm{dim} = 1). \tag{11}$$

Next, the GCN is leveraged to perform relation reasoning, so the original word features $E[S]$ are updated to $E[\hat{S}]$:

$$E[\hat{S}] = \mathrm{ReLU}\left[\left(\psi(E[S]) + \hat{N} \cdot \psi(E[S])\right) W_{s'} + b_{s'}\right], \tag{12}$$

$W_{s'} \in \mathbb{R}^{d\times d}$ and $b_{s'} \in \mathbb{R}^d$ are the parameters to be learned.

2) Inter-relation Embedding (IEE): In this block, an appearance, a motion and a text attention gate, *i.e.,* α, β, and γ are designed to determine the relevance of different modalities, as shown in Fig. 3. To be more specific, each attention gate is derived from all the independent modalities via the similar non-linear transformation but with different parameters, which is computed respectively as:

$$\begin{aligned}
\alpha &= \sigma\left(\hat{A}W_{a1} + \hat{M}W_{m1} + E[\hat{S}]W_{t1} + b_1\right), \\
\beta &= \sigma\left(\hat{A}W_{a2} + \hat{M}W_{m2} + E[\hat{S}]W_{t2} + b_2\right), \\
\gamma &= \sigma\left(\hat{A}W_{a3} + \hat{M}W_{m3} + E[\hat{S}]W_{t3} + b_3\right),
\end{aligned} \tag{13}$$

where σ is the sigmoid activation function. $W_* \in \mathbb{R}^{d\times d}$ and $b_* \in \mathbb{R}^d$ are the parameters to be learned. The values of each attention gate indicate the relevance of this modality with respect to other modalities. Then, the related representations $\tilde{A}$, $\tilde{M}$, and $E[\tilde{S}]$ are yielded via applying each attention gate to the independent modality $\hat{A}$, $\hat{M}$, and $E[\hat{S}]$ using element-wise multiplication, respectively:

$$\tilde{A} = \alpha \odot \hat{A}, \qquad \tilde{M} = \beta \odot \hat{M}, \qquad E[\tilde{S}] = \gamma \odot E[\hat{S}]. \tag{14}$$

Through this manner, each modality is encouraged to build the inter-relation with other modalities and embed cross-modal information as supplement during multi-modal semantic interactions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

After inter-relations in cross-modalities are built, we fuse three kinds of related representations into the omni-representation by a fully connected layer:

$$X = \left[\tilde{A}, \tilde{M}, E[\tilde{S}]\right] W_{c1} + b_{c1}, \tag{15}$$

where [,] denotes the concatenation operation. $W_{c1} \in \mathbb{R}^{3d \times d}$ and $b_{c1} \in \mathbb{R}^{d}$ are parameters to be learned.

C. Transformer-based Caption Generation

Our encoder and decoder are based on the vanilla transformer [8]. We first briefly review this framework which is shown in Fig. 4. The core of it is the scaled dot-product attention. Concretely, given a query matrix $Q \in \mathbb{R}^{T_q \times d_k}$, key matrix $K \in \mathbb{R}^{T_v \times d_k}$ and value matrix $V \in \mathbb{R}^{T_v \times d_v}$, the attention result is computed as:

$$\text{Attention } (Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}, \text{dim} = 1\right) V. \tag{16}$$

The multi-head attention is built upon the scaled dot-product attention. It consists of h different "heads", where each head is independent and computed in parallel. For each head, the attention result is

$$\text{head }_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \tag{17}$$

Afterward, the multi-head attention operation is to concatenate all the heads, which is defined as:

$$\text{MultiHead } (Q, K, V) = \text{ Concat }_{i=1...h}\left(\text{ head }_i\right) W^O. \tag{18}$$

This attention mechanism can be used for two kinds of purposes, such as self-attention where query, key, and value matrix are all the same, and cross-attention where the query matrix is different from the key and value matrix. In our method, we also use multi-head attention to update the learned omni-representation, as discussed later. Furthermore, the output of each attention layer x is fed into a feed-forward layer (FFN) which utilizes a non-linear transformation:

$$\text{FFN}(\boldsymbol{x}) = \text{GELU}\left(\boldsymbol{x}W_{f1} + \boldsymbol{b}_{f1}\right) W_{f2} + \boldsymbol{b}_{f2} \tag{19}$$

1) Encoding Stage: In this stage, we employ the multi-head self-attention and LayerNorm (LN) [40] operations to encode the omni-representation X obtained from the IAE and IEE. Besides, residual connections are [41] also used to help avoid the vanishing gradient problem during the training phase. In the self-attention layer, the query, key and value matrix are all the X and we update X with $\hat{X}$ by:

$$\hat{X} = \text{ LN } (X + \text{ MultiHead } (X, X, X)). \tag{20}$$

Then, the $\hat{X}$ is inputed to the feed-forward layer:

$$\tilde{X} = \text{ LN}((\hat{X} + \text{FFN}(\hat{X})). \tag{21}$$

In our method, following [7], [9], the encoder consists of a stack of N identical layers. At the l-th encoder layer, the multi-head attention mechanism takes the output of the last layer as inputs and performs self-attention to model the intra-relation in the omni-representation. The output of the attention layer is then projected by a feed-forward layer. After the above operations, the intra-relation of this fused representation $\tilde{X}$ is further enhanced.

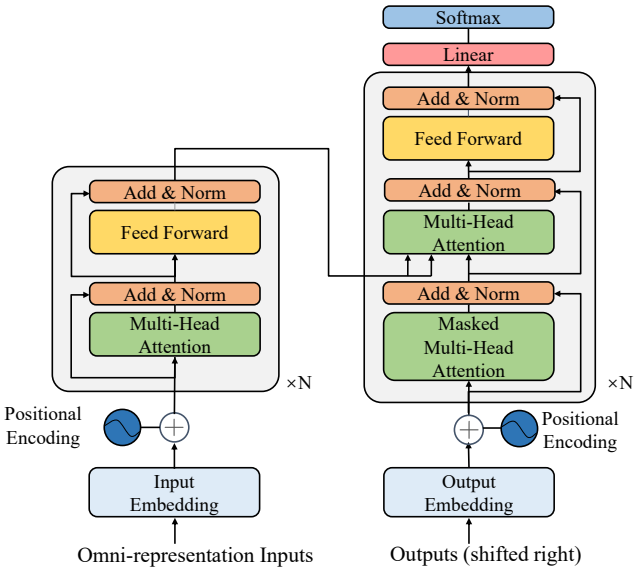


Fig. 4. The architecture of the used Transformer. The input is the learned omni-representation of both video and subtitle, and the output is the high-level linguistic sentence.

2) Decoding Stage: Our decoder also consists of a stack of N identical layers. At the l-th decoder layer, the masked self-attention layer, which is used to prevent the model from seeing future words, first takes the caption word embedding features $E[W] = \{E[w_1], ..., E[w_m]\}$ as the inputs and projected them into query, key, and value matrix. The operation of this layer is defined as:

$$E[\hat{W}] = \text{ LN } (E[W] + \text{ MultiHead } (E[W], E[W], E[W])). \tag{22}$$

Then, in the cross-attention layer, the query matrix is the $E[\hat{W}]$, and the key and value matrix are the output of the last encoder layer, *i.e.*, $\tilde{X}$. The attention operation is defined as:

$$\hat{H} = \text{ LN } (E[\hat{W}] + \text{ MultiHead } (E[\hat{W}], \tilde{X}, \tilde{X})). \tag{23}$$

Afterward, the $\hat{H}$ is passed to a feed-forward layer:

$$\tilde{H} = \text{ LN}((\hat{H} + \text{FFN}(\hat{H})). \tag{24}$$

Finally, The probability distribution of target words will be calculated via a single hidden layer:

$$\tilde{W} = \text{softmax}(\text{GELU}\left(\tilde{H}W_p + \boldsymbol{b}_p\right) W_q + \boldsymbol{b}_q) \tag{25}$$

IV. Experiments

A. Datasets

TV show Captioning Dataset (TVC). TVC [7] has 21,793 video clips from 6 long-running TV shows. Each video in the training set is paired with a subtitle and two ground truth captions, while each video in the validation and test sets is paired with a subtitle and four ground truth captions. We use the official split with 17,435 videos for training, 2,179 for validation and 1,089 for testing.

TV show Retrieval Dataset (TVR). TVR [7] is the first task for using video+subtile to perform text-based video-moment retrieval. Given a query sentence, the model is need

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE TVC VALIDATION AND ONLINE TEST-PUBLIC SET. ALL THE METHODS UTILIZE BOTH VIDEOS AND SUBTITLES. R, I, SF, SUB ARE SHORT FOR RESNET-152, I3D, SLOWFAST, AND SUBTITLE. B-4, M, R, C ARE SHORT FOR BLEU-4, METEOR, ROUGE-L, AND CIDER. THE SYMBOL "-" INDICATES SUCH RESULTS ARE UNREPORTED IN THE ORIGNAL PAPER.

Method	Pre-training	Validation				Test			
		B-4	M	R	C	B-4	M	R	C
HERO (R+SF+Sub) (2020) [9]	7.6M	12.25	17.54	34.10	50.46	12.35	17.64	34.16	49.98
MMT (R+I+Sub) (2020) [7]	None	10.53	16.61	32.35	44.39	10.87	16.91	32.81	45.38
HERO (R+SF+Sub) (2020) [9]	None	10.75	16.42	32.72	43.62	-	-	-	-
I^2Transformer (R+I+Sub)	None	**11.46**	**16.83**	**33.02**	**47.21**	**11.73**	**17.10**	**33.30**	**48.07**
(Relative Improvements%↑)		**(6.6%↑)**	**(1.3%↑)**	**(0.9%↑)**	**(6.4%↑)**	**(7.9%↑)**	**(1.1%↑)**	**(1.5%↑)**	**(5.9%↑)**

to not only retrieve the most relevant video from the video corpus, but also localize the relevant moment in the retrieved video. TVR contains 109K queries from 21.8K videos and is split into 80% train, 10% validation, and 5% test-public.

Video-guided Machine Translation Dataset (VMT). VMT [42] is based on VATEX dataset, which contains over 41,250 videos and 825,000 captions in both English and Chinese. Among the captions, there are over 206,000 English-Chinese parallel translation pairs. We use the official split with 25,991 videos for training, 3,000 for validation and 6,000 for testing.

B. Evaluation Metrics

For TVC, we use four standard metrics to evaluate the quality of generated sentences, *i.e.*, BLEU-4 [43], METEOR [44], ROUGE-L [45] and CIDEr [46]. BLEU-4 is widely utilized for corpus level comparisons over which 4-gram matches exist. METEOR is able to generate an alignment based on exact token matching so as to judge the word correlation between candidate and reference sentences. ROUGE-L utilizes the measure according to the Longest Common Sub-sequence (LCS), which is a group of words shared by two sentences in the same order. CIDEr is recently proposed and especially designed for the captioning task to capture human judgment of consensus. We get all the results in this paper according to the Microsoft COCO evaluation server [47].

For TVR, following [7], [9], we utilize average recall at K (R@K) over all queries as the metric. The prediction is correct if: (1) predicted video matches the ground truth; (2) predicted span has high overlap with the ground truth, where temporal intersection over union (IoU) is used to measure overlap. For VMT, following [42], BLEU-4 is used as the metric to evaluate the quality of translated sentences. All result values in our tables are reported as percentage (%).

C. Implementation Details

Video Feature Extraction. Appearance features are represented by 2048D ResNet-152 [41] pre-trained on the Imagenet dataset [48], and motion features are represented by 1024D I3D [49] pre-trained on the Kinetics-600 dataset [50]. The frame number is set as 20, 100, and 32 on TVC, TVR, and VATEX, respectively.

Words Processing. For the subtitles and ground truth captions on TVC, the maximum length is set to 30 and 20, respectively. For the subtitles and query sentences on TVR, the maximum length is set to 50 and 30, respectively. For the source and target sentences on VATEX, the maximum length of them are both set as 40. The size of word embedding for each word is set to 300 and 512 on TVC and VATEX, respectively. For TVR, following [7], the words in subtitles and queries are represented by 768D RoBERTa [51].

Model Setting. The hidden size of overall model is set to 768, 256, and 512 on TVC, TVR, and VATEX, respectively. The layers of graphs and attention gates are set to 2. In the Transformer, for fair comparison, following the previous work [7], we set the layers of both encoder and decoder to 2, and the number of attention heads to 12.

Model Training. (1) TVC: In the training phase, we set the mini-batch size as 128 and the learning rate is set to 1×10^{-4}. Moreover, we set dropout regularization in the rate of 0.1 and implement element-wise gradients clipping at 1. The maximum training iteration is set as 50 epochs. We use Adam optimizer [52] to minimize the negative log-likelihood loss:

$$L(\theta) = -\sum_{t=1}^{m} logp(w_t|w_{<t}, A, M, E[S], \theta), \qquad (26)$$

where θ are the parameters of a video captioning model and m is the length of a caption. Especially, CIDEr [46] is designed for captioning task, so the highest score of CIDEr on the validation set is used as a metric to choose the best model for testing. For inference, greedy decoding strategy is used to generate target captions.

(2) TVR and VMT: On both TVR and VMT, the mini-batch size is set to 128 and 512. The learning rate is set to 1×10^{-4} and 1×10^{-3}. Adam is also used as the optimizer.

For three datasets, we train the model with PyTorch [53] on an RTX 2080 Ti GPU. GPU memory cost is about 9GB on TVC, 6GB on TVR, and 10GB on VATEX. The training time is around 7 hours on three datasets.

D. The performance comparison on TVC

1) Comparing with state-of-the-art Methods: In this dataset, we compare the proposed I^{2}Transformer with two state-of-the-art methods, MMT [7] and HERO [9], on both validation and test-public sets. The experimental results are shown in Table I.

Compared to MMT, when using the same features, we can observe that I^{2}Transformer outperforms MMT on both split sets in terms of all metrics, in particular with increases of 8.8% and 7.9% on BLEU-4, as well as 6.4% and 5.9% on CIDEr. This benefits from that intra-relations and inter-relation are

embedded into all modalities during multi-modal semantics interactions. Hence, the learned omni-representation is able to not only summarize the useful information from each modality, but also make each modality as supplementary information for other modalities. For HERO, without pre-training, it achieves performance comparable to MMT, but inferior to ours in terms all metrics on validation set. When using additional 7.6M multi-modal samples for pre-training (requiring about 3 weeks with 16 Nvidia V100 GPUs), HERO does only slightly better than our model. The comparative results show the effectiveness of our method. Besides, I^{2}Transformer achieves similar performance on the validation and test sets, which show its good generalization ability.

2) The Performance with Different Input Modality: On the online test-public set of TV show dataset, we evaluate our proposed I^{2}Transformer with different input modalities.

TABLE II
PERFORMANCE COMPARISON ON THE TVC ONLINE TEST-PUBLIC SET, WITH DIFFERENT INPUT MODALITY, WHERE A, M, T DENOTE THE APPEARANCE, MOTION, AND TEXT MODALITY, RESPECTIVELY.

Method	B-4	M	R	C
MMT (T)	6.33	13.92	27.73	33.76
I^{2}Transformer (T)	**6.69**	**14.28**	**28.23**	**34.57**
MMT (A+M)	9.98	15.23	30.44	36.07
I^{2}Transformer (A+M)	**10.15**	**15.55**	**31.13**	**36.09**
MMT (A+M+T)	10.87	16.91	32.81	45.38
I^{2}Transformer (A+M+T)	**11.73**	**17.10**	**33.30**	**48.07**

The experimental results are shown in Table II. We can observe that the I^{2}Transformer with different modalities all achieved significant improvements over the MMT. In the MMT, the input of the transformer is computed by linear transformation and concatenation, which neglects the intra-relation in each modality and inter-relations in cross modalities. For instance, appearance features (A) can represent entities and motion features (M) can represent actions. If directly concatenating them, the semantic interactions between entities and actions will be ignored. Instead, our method can respectively build semantic relations in videos and subtitles, as well as semantic interactions in between. Therefore, the learned omni-representation is capable of clearly representing the entities and actions in videos, as well as intentions and sentimental information in subtitles. Besides, as we discussed above, the information of subtitle (T) is helpful, but how to fuse it with two kinds of video modalities is a big challenge in this task. We can observe that although the improvement of I^{2}Transformer (A+M) is not much better than the MMT (A+M), the improvement is significant when inputs are A+M+T. This indicates that the proposed method can learn an intra- and inter-relation embedded omni-representation for video and subtitle, which is the main contribution of this paper.

E. Ablation Studies

In this section, we make ablation analyses for coupling the proposed IAE, IEE and their combination with the transformer. For convenience, we denote them as IAE-Trans, IEE-Trans, and I^{2}Transformer, respectively. Please note that since the ground truth captions are not provided in the test-public set, all the ablation studies are conducted based on the validation set and in this set, the results of compared MMT [7] are reproduced via their released code [1].

1) The Evaluation of the Intra-relation Embedding Transformer: In order to validate the effectiveness of the proposed IAE-Trans, we conduct the two kinds of experiments based on the number of input modality: (1) single modality; (2) multi-modalities. For the former, we utilize single appearance, motion, and text modality to generate target captions, respectively. For the latter, we utilize multiple kinds of modalities to generate target captions. The compared MMT [7] only uses linear transformation to process each single modality, and multiple modalities are fused by concatenation.

TABLE III
ABLATION ANALYSIS OF THE INTRA-RELATION EMBEDDING TRANSFORMER (IAE-TRANS) ON THE TVC VALIDATION SET, WHERE A, M, T DENOTE APPEARANCE, MOTION AND TEXT MODALITY, RESPECTIVELY.

Method	B-4	M	R	C
MMT (A)	8.66	**14.78**	29.32	33.44
IAE-Trans (A)	**9.03**	14.75	**29.63**	**33.58**
MMT (M)	9.05	14.46	29.35	31.86
IAE-Trans (M)	**9.48**	**14.69**	**29.77**	**32.25**
MMT (T)	5.98	13.76	27.46	33.18
IAE-Trans (T)	**6.37**	**14.02**	**27.88**	**33.74**
MMT (A+M)	9.75	15.21	30.42	36.16
IAE-Trans (A+M)	**10.07**	**15.58**	**30.86**	**36.82**
MMT (A+T)	9.59	16.23	31.51	42.71
IAE-Trans (A+T)	**10.32**	**16.49**	**32.20**	**44.31**
MMT (M+T)	10.55	16.37	32.10	43.92
IAE-Trans (M+T)	**10.56**	**16.38**	**32.20**	**44.70**
MMT (A+M+T)	10.52	16.46	32.19	44.60
IAE-Trans (A+M+T)	**11.16**	**16.83**	**33.02**	**46.33**

The experimental results are shown in Table III. We can observe that for both single and multi-modal inputs, augmenting the Transformer with the IAE achieves significant improvements, which indicates that the IAE indeed has the ability of capturing the intra-relation about each modality. Please note that to keep the same length of three kinds of features, we pad zeros in the shortage part of video and special tokens "VID" in the shortage part of subtitle. There is no doubt that the padding parts in the both video and subtitle features would be the noise for the feature representation. We can find that when using the same input modality with the same padding strategy, equipping the standard Transformer with the proposed IAE significantly outperforms the MMT, which validates the IAE can learn useful information from each modality and overcome the influence of the irrelevant padding parts by measuring the relevance of feature information within the same modality. Moreover, we have some interesting observations, that is, 1) captioning models with appearance or motion modality perform better than those using text modality on the most metrics; 2) using both video and subtitle information is much better than only using video information. These observations indicate that 1) appearance and motion modalities can represent explicit information such as objects and actions, while text modality is able to convey

[1] https://github.com/jayleicn/TVCaption

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

implicit information such as actors' intentions and sentiments; 2) subtitle information is an essential part to improve the cognition-level of the captioning model and beneficial for generating high-quality captions.

TABLE IV
ABLATION ANALYSIS OF THE INTER-RELATION EMBEDDING TRANSFORMER (IEE-TRANS) ON THE TVC VALIDATION SET.

Method	B-4	M	R	C
MMT (A+M)	9.75	15.21	30.42	36.16
IEE-Trans (A+M)	**10.10**	**15.45**	**30.88**	**36.45**
MMT (A+T)	9.59	16.23	31.51	42.71
IEE-Trans (A+T)	**10.01**	**16.33**	**31.74**	**43.04**
MMT (M+T)	**10.55**	16.37	**32.10**	43.92
IEE-Trans (M+T)	10.52	**16.47**	**32.10**	**44.76**
MMT (A+M+T)	10.52	16.46	32.19	44.60
IEE-Trans (A+M+T)	**10.89**	**16.78**	**32.77**	**45.85**

2) The Evaluation of the Inter-relation Embedding Transformer: In order to validate the effectiveness of the proposed IEE, first, similar to previous work [7], we use linear transformation to process each modality. Second, instead of directly fusing multiple kinds of modalities via simple concatenation, we first leverage the IEE to model inter-relations in cross modalities and then fuse three related modalities via concatenation.

The experimental results are shown in Table IV. We can observe that 1) the IEE-Trans outperforms the MMT; 2) the improvement achieved by the IEE-Trans is lower than that of the IAE-Trans. The above observations indicate that 1) it is significant to model inter-relations in cross modalities, because the representation of each modality can be augmented by the cross-modal information during the relations learning process [54], [55]; 2) intra-modal relation should be built first, because only if the information of each modality is fully exploited, can it be as useful supplementary information during multi-modal semantic interactions.

3) Evaluation of Intra- and Inter-relation Embedding Transformer: In this part, we evaluate the performance of the I^{2}Transformer, which first utilizes the IAE to build the intra-relation of each modality and then leverages the IEE to build the inter-relations in cross modalities.

TABLE V
ABLATION ANALYSIS OF THE INTRA- AND INTER-RELATION EMBEDDING TRANSFORMER (I^{2}TRANSFORMER) ON THE TVC VALIDATION SET.

Method	B-4	M	R	C
MMT (A+M)	9.75	15.21	30.42	36.16
I^{2}Transformer (A+M)	**10.10**	**15.57**	**31.10**	**36.85**
MMT (A+T)	9.59	16.23	31.51	42.71
I^{2}Transformer (A+T)	**10.46**	**16.46**	**32.13**	**44.40**
MMT (M+T)	10.55	16.37	32.10	43.92
I^{2}Transformer (M+T)	**11.03**	**16.49**	**32.40**	**44.62**
MMT (A+M+T)	10.52	16.46	32.19	44.60
I^{2}Transformer (A+M+T)	**11.46**	**16.83**	**33.02**	**47.21**

The experimental results are shown in Table V. We can observe that when building the relations of each modality and cross modalities simultaneously, the best performances are achieved. Furthermore, comparing Table II with Table V, we can observe that when using all the appearance, motion and text modalities, the I^{2}Transformer achieves similar performances on the validation and test sets. This proves that our method has good generalization ability, which benefits from the intra- and inter-relation embedding omni-representation.

4) Learning Curves: In order to obtain a more intuitive view of each component and the full model capacity, the learning curves of the CIDEr scores on the validation set are shown in Fig. 5. Four model performances are presented: MMT, IAE-Trans, IEE-Trans, and I^{2}Transformer. All of the models are trained with the same input modalities, *i.e.*, appearance, motion, and text. We can observe that the I^{2}Transformer performs better consistently than the others and achieves the best CIEDr score. The learning curves validate the effectiveness of learning intra-relation in each modality and inter-relation among different modalities when modeling the omni-representation of both video and subtitle.

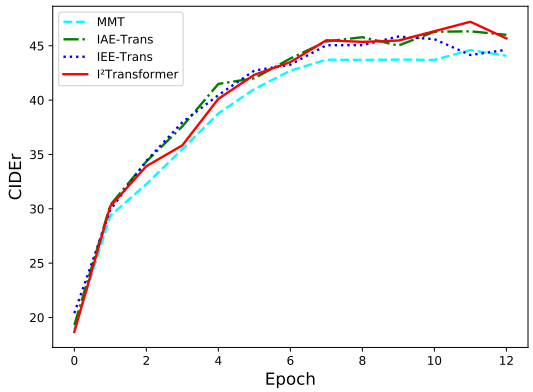


Fig. 5. Learning curves of the CIDEr scores on the validation set of TVC.

F. Performance comparison on TVR

In order to validate the generalization ability of the proposed IAE and IEE, we evaluate them on TVR, which is a relevant multi-channel videos (video+subtitle) task. We will give more details about how we apply the proposed IAE and IEE to perform TV show retrieval. For video and subtitle, we first use proposed IAE to learn the intra-relation in appearance, motion, and text modality, respectively. Then, the IEE is performed to learn the inter-relations among three modalities, and select the related information from each modality. Finally, we obtain the subtitle representation $H^s \in \mathbb{R}^{l_s \times d_s}$, and concatenate appearance with motion modalities to form the video representation $H^v \in \mathbb{R}^{l_v \times d_v}$. For queries in TVR, we first use IAE to learn semantic relations between words for summarizing its main content, which is denoted as $H^q \in \mathbb{R}^{l_q \times d_q}$ and $H^q = \{\mathbf{h_r^q}\}_{r=1}^{l_q}$. Then, since a query can be relevant to either the video or its paired subtitle, following [7], the query is decomposed into two vectors:

$$a_r^m = \frac{\exp\left(\mathbf{w}_m^T \mathbf{h}_r^q\right)}{\sum_{k=1}^{l_q} \exp\left(\mathbf{w}_m^T \mathbf{h}_k^q\right)}, \mathbf{q}^m = \sum_{r=1}^{l_q} a_r^m \mathbf{h_r^q}. \quad (27)$$

Here, m can indicate either video (v) or subtitle (s). $w^m \in \mathbb{R}^{d_q}$ and $q^m \in \mathbb{R}^{d_q}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

TABLE VI
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TVR VALIDATION AND ONLINE TEST-PUBLIC SET. ALL THE METHODS UTILIZE BOTH VIDEOS AND SUBTITLES. R, I, SF, SUB ARE SHORT FOR RESNET-152, I3D, SLOWFAST, AND SUBTITLE. THE SYMBOL "-" INDICATES SUCH RESULTS ARE UNREPORTED IN THE ORIGINAL PAPER.

Method	Pre-training	Validation				Test			
		R@1	R@5	R@10	R@100	R@1	R@5	R@10	R@100
HERO (R+SF+Sub) (2020) [9]	7.6M	5.13	-	16.26	24.55	6.21	-	19.34	36.66
XML (R+I+Sub) (2020) [7]	None	2.62	6.39	9.05	22.47	3.32	9.46	13.41	30.52
HERO (R+SF+Sub) (2020) [9]	None	2.98	-	**10.65**	18.42	-	-	-	-
I^{2}Transformer (R+I+Sub)	None	**3.04**	**7.25**	9.84	**24.34**	**3.53**	**9.86**	**13.94**	**31.00**
(Relative Improvements%↑)		**(2.0%↑)**	**(13.5%↑)**	(7.6%↓)	**(8.3%↑)**	**(6.3%↑)**	**(4.2%↑)**	**(4.0%↑)**	**(1.6%↑)**

After obtained H^q, H^v, and H^s, we exploit a convolutional start-end detector with two 1D convolution filters [7] to compute the start (st) and end (ed) probabilities in a video moment:

$$S_{\text{st}} = \text{Conv 1D}_{\text{st}}\left(S_{\text{query-clip}}\right), \quad S_{\text{ed}} = \text{Conv1D}_{\text{ed}}\left(S_{\text{query-clip}}\right), \tag{28}$$

where $S_{\text{query-clip}} \in \mathbb{R}^l$ is the query-based matching scores and it is computed by:

$$S_{\text{query-clip}} = \frac{1}{2}\left(H_1^v \mathbf{q}^v + H_1^s \mathbf{q}^s\right). \tag{29}$$

The experimental results are shown in Table VI. Please note that, we follow the XML [7] (ECCV 2020, the pioneer work for TV show Retrieval) and use the same setting (without NMS, an extra trick to remove the highly overlapped but with lower score predictions) to report the results on both validation and test-public sets. In HERO [9], NMS is used in the inference phase of validation set to report the results. From Table VI, we can find that, compare to the state-of-the-art work XML, our method outperforms it on both splits on all metrics. Especially, I^{2}Transformer yields improvements of 16.0% and 6.3% on R@1, as well as 13.5% and 4.2% on R@5. Compared to the HERO, when training is implemented without additional multi-modal data for pre-training, our method surpasses it on most metrics. When using additional 7.6M multi-modal samples for pre-training, HERO achieves the performance superior to ours. However, as reported in [9], this requires expensive 7.6M multi-modal data for pre-training and takes about 3 weeks with 16 Nvidia V100 GPUs. Hence, when both methods are trained without multi-modal samples, the I^{2}Transformer is more competitive. Experimental results validate that learning both intra- and inter-relations is also important in this task, because each modality not only can be sufficiently exploited, but also can embed cross-modal information as supplement during multi-modal semantic interactions.

G. Performance comparison on VMT

We also apply the IAE and IEE to the video-guided machine translation (VMT) task on VATEX dataset, where this task aims to translate an English sentence under the guidance of the video into a Chinese sentence. Hence, similar to TVC, the translation model also requires understanding not only the contents of video and source sentence, but also the relations in between. Since VMT is also a language generation task, the IAE and IEE can be easily applied by replacing the subtitle input in Fig. 2 with English captions. Besides, since the previous method in VMT [42] used the LSTM as the encoder and decoder, for fair comparison, we also couple the IAE and IEE with the LSTM to conduct experiments, and the hidden size of LSTM is set to 512. Furthermore, when using the LSTM, similar to VMT, we also equip the decoder with (w/) or without (w/o) temporal attention.

TABLE VII
EVALUATION OF THE IAE AND IEE ON VMT, WHERE M DENOTES THE MOTION MODALITY REPRESENTED BY I3D AND T INDICATES TEXT MODALITY OF SOURCE ENGLISH SENTENCE.

Method	BLEU-4
VMT-LSTM w/o attention (M+T) (2019)[42]	27.43
VMT-LSTM w/ attention (M+T) (2019) [42]	29.12
I^{2}LSTM w/o attention (M+T)	32.11
I^{2}LSTM w/ attention (M+T)	32.46
I^{2}Transformer (M+T)	**32.66**

The experimental results are shown in Table VII. There are the following observations, that is, 1) the performance achieves the best when coupling both IAE and IEE with the transformer; 2) with or without temporal attention, our methods all obtain the significant improvements over the VMT; 3) the performances are close in our LSTM-based methods. These indicate that 1) compared to the LSTM, the transformer is more powerful for the language generation task [38], [27]; 2) both IAE and IEE are still able to capture the intra- and inter-relations in multiple modalities in this task and suitable for different encoder-decoder frameworks; 3) when capturing the intra-relation in a video, the IAE can mine the underlying temporal relation in a video besides the semantic relation. Therefore, the IAE can help reduce the dependency for the temporal attention mechanism and decrease the computation cost. Besides, the experimental results on TVC, TVR and VMT show that the proposed IAE and IEE have a good generalization ability for addressing the tasks of video with text inputs.

H. Qualitative Analysis

Fig. 6 shows four examples of captions generated by humans, the MMT [7] and the I^{2}Transformer. For the first example, we can intuitively observe that the caption generated by the MMT is logically correct but not in accordance with the video contents. Our conjecture is that the MMT cannot capture the intra-relation in the subtitle and thus neglects the clue "dropping" provided by the subtitle. For the second example, we can observe that the generated captions by the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3159472, IEEE Transactions on Image Processing

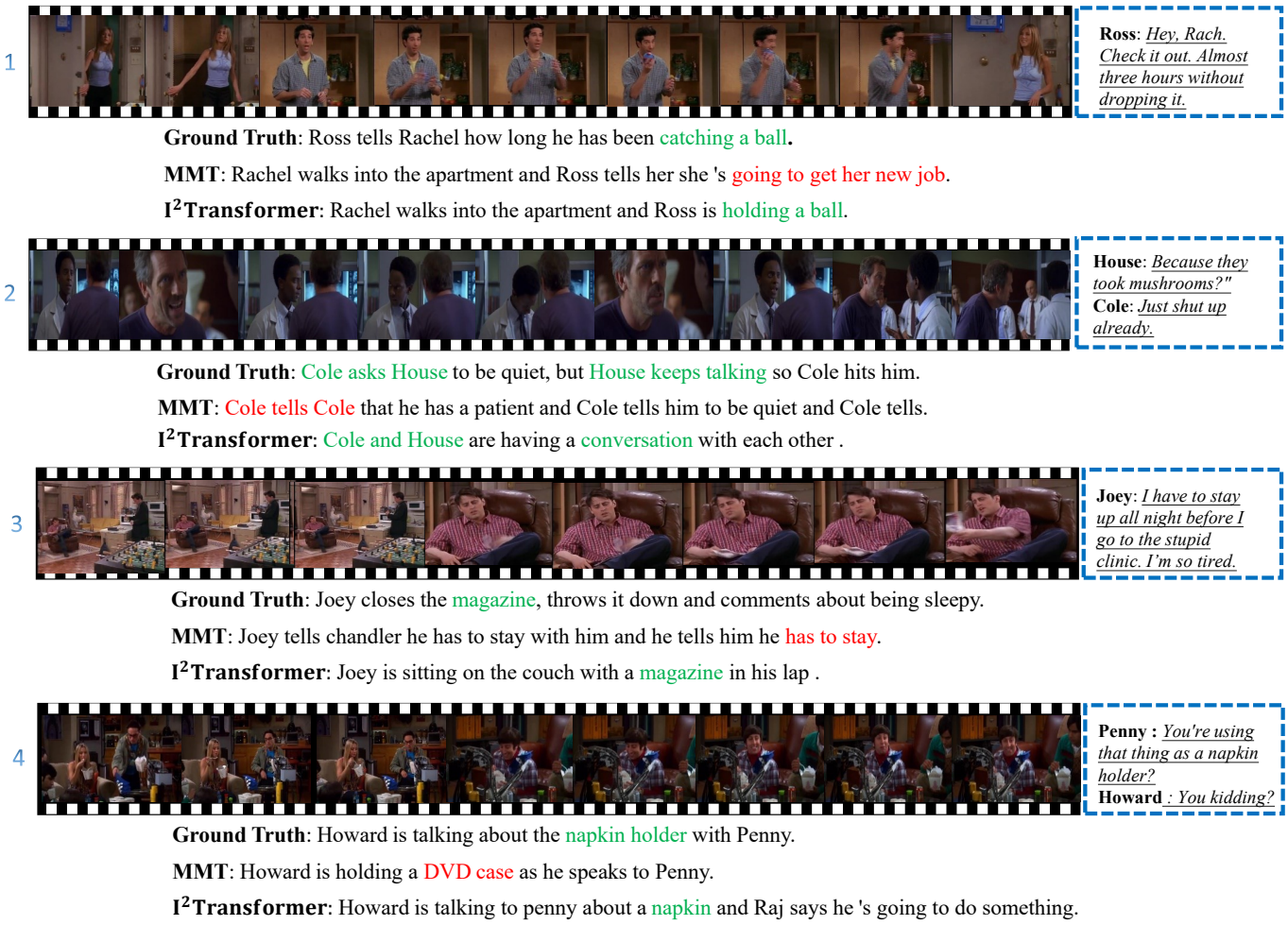


Fig. 6. Four examples from the validation set of TVC, which involve captions generated by humans, the state-of-the-art MMT and our I^{2}Transformer. The words in green and red color respectively denote correct and incorrect words with respect to the ground truth words.

MMT has described the main contents of both video and subtitle, but it is logically incorrect. Our conjecture is that the MMT cannot fully exploit the inter-relations in cross modalities and thus generates this confused caption. For the third example, we can observe that the paired subtitle is very abstract and cannot directly correspond to the video content. In this case, the captioning model should use more video information. However, the MMT cannot dynamically select useful information from each modality, so it generates a totally wrong sentences. For the last example, the detail "napkin" is difficult to represent by either appearance or motion modality, but it appears in the actors' dialogue. However, it is surrounded by a set of scrappy information. If the captioning model is unable to reorganize the subtitle information, this key clue will be ignored, as shown in generated sentence by the MMT. By contrast, the captions generated by the I^{2}Transformer are not only logically correct, but also can capture the main information conveyed by the video and subtitle. This benefits from the intra- and inter-relations are embedded into the omni-representation of video and subtitle. Hence, this representation can convey the explicit information (e.g. actions) in the video and implicit information (e.g. intentions) in the subtitle.

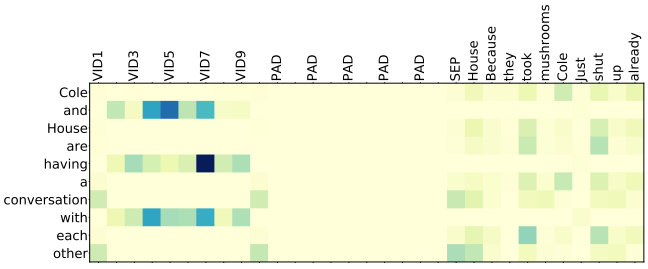


Fig. 7. The visualization of attention distributions for caption words with respect to video frames and subtitle words. "VID" denotes video frames.

In Fig. 7, we select the second example of Fig. 6 to visualize its attention distributions for generated words with regard to video frames and subtitle words. The caption is generated by the I^{2}Transformer. From Fig. 7, we can observe that, on one hand, when generating the words reflecting the actions or relations of actors ("having", "and" "with"), the captioning model will pay more attention to video frames. On the other hand, when generating other words such as the abstract noun ("conversation") and names of person ("House",

"Cole"), the captioning model focuses more on subtitle words. The visualization result shows that our method not only accurately understands the contents of video and subtitle, but also adaptively exploits each of them to generate target words. This mainly benefits from that our method can generate a powerful omni-representation for both video and subtitle, and thus clearly represent entitles, actions and intentions conveyed by them.

V. Conclusion

In this paper, we propose an intra- and inter-relation embedding Transformer (I^2Transformer) for TV show caption generation. To derive an omni-representation for both video and subtitle, the model first utilizes the intra-relation embedding block (IAE) to model the relation in appearance, motion, and text modalities, respectively. Then, it leverages the inter-realtion embedding block (IEE) to learn the mutual relations in cross modalities. Through this network, each modality not only can be fully exploited, but also can be augmented with cross-modal information during the multi-modal semantic interactions. Finally, the learned omni-representation is fed into the Transformer for caption generation. Extensive experiments with different input modalities on TVC dataset show that the I^2Transformer achieves the state-of-the-art performance. Besides, in order to validate the generalization ability of the proposed IAE and IEE, we additionally conduct experiments on TV show retrieval and video-guided machine translation which are two tasks of video with text inputs. The encouraging performances are also achieved on both datasets, which validate that our proposed IAE and IEE have a good generalization ability.

In the future, we will attempt to exploit pre-training strategy to further boost the performance of the proposed method, and most importantly, to seek a trade-off between the hardware requirement and good performance.

References

[1] M. Yang, J. Liu, Y. Shen, Z. Zhao, X. Chen, Q. Wu, and C. Li, "An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network," *IEEE Transactions on Image Processing*, vol. 29, pp. 9627–9640, 2020.

[2] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1180–1192, 2020.

[3] H. Liu, S. Zhang, K. Lin, J. Wen, J. Li, and X. Hu, "Vocabulary-wide credit assignment for training image captioning models," *IEEE Transactions on Image Processing*, vol. 30, pp. 2450–2460, 2021.

[4] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H. T. Shen, and Y. Ji, "Video captioning by adversarial lstm," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5600–5611, 2018.

[5] B. Zhao, X. Li, and X. Lu, "Cam-rnn: Co-attention model based rnn for video captioning," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5552–5565, 2019.

[6] J. Zhang and Y. Peng, "Video captioning with object-aware spatio-temporal correlation and aggregation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6209–6222, 2020.

[7] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *ECCV*, 2020, pp. 447–463.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[9] L. Li, Y. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: hierarchical encoder for video+language omni-representation pre-training," in *EMNLP*, 2020, pp. 2046–2065.

[10] S. Yao and X. Wan, "Multimodal transformer for multimodal machine translation," in *ACL*, 2020, pp. 4346–4350.

[11] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Task-driven dynamic fusion: Reducing ambiguity in video description," in *CVPR*, 2017, pp. 3713–3721.

[12] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with pos sequence guidance based on gated fusion network," in *ICCV*, 2019, pp. 2641–2650.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.

[14] W. Wu, D. Tao, H. Li, Z. Yang, and J. Cheng, "Deep features for person re-identification on metric learning," *Pattern Recognition*, vol. 110, p. 107424, 2021.

[15] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 1, pp. 163–174, 2018.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *ICCV*, 2013, pp. 433–440.

[18] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *AAAI*, 2015, pp. 2346–2352.

[19] Y. Tu, X. Zhang, B. Liu, and C. Yan, "Video description with spatial-temporal attention," in *ACM MM*, 2017, pp. 1014–1022.

[20] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *CVPR*, 2020, pp. 13 096–13 105.

[21] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia (2019)*, 2019.

[22] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating part of speech guidance for image captioning," *IEEE Transactions on Multimedia*, vol. 23, pp. 92–104, 2021.

[23] J. Perez-Martin, B. Bustos, and J. Perez, "Improving video captioning with temporal composition of a visual-syntactic embedding," in *IW-CACV*, 2021, pp. 3039–3049.

[24] Y. Tu, C. Zhou, J. Guo, S. Gao, and Z. Yu, "Enhancing the alignment between target words and corresponding frames for video captioning," *Pattern Recognition*, vol. 111, p. 107702, 2021.

[25] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *In Neural Computation,*, pp. 1735–1780, 1997.

[27] J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal, "Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning," in *ACL*, 2020, pp. 2603–2614.

[28] M. Suin and A. Rajagopalan, "An efficient framework for dense video captioning," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 039–12 046.

[29] J. S. Park, T. Darrell, and A. Rohrbach, "Identity-aware multi-sentence video description," in *ECCV*. Springer, 2020, pp. 360–378.

[30] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, "Learning modality interaction for temporal sentence localization and event captioning in videos," in *ECCV*. Springer, 2020, pp. 333–351.

[31] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *ICCV*, 2017, pp. 4193–4202.

[32] W. Hao, Z. Zhang, and H. Guan, "Integrating both visual and audio cues for enhanced video caption," in *AAAI*, 2018, pp. 6894–6901.

[33] X. Wang, Y.-F. Wang, and W. Y. Wang, "Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning," in *NAACL-HLT*, 2018, pp. 795–801.

[34] T. Rahman, B. Xu, and L. Sigal, "Watch, listen and tell: Multi-modal weakly supervised dense event captioning," in *ICCV*, 2019, pp. 8908–8917.

[35] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognition*, vol. 98, p. 107075, 2020.

[36] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018, pp. 711–727.

[37] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, "Object relational graph with teacher-recommended learning for video captioning," in *CVPR*, 2020, pp. 13 278–13 288.

[38] B. Pan, H. Cai, D.-A. Huang, K.-H. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *CVPR*, 2020, pp. 10 867–10 876.

[39] Q. Huang, J. Wei, Y. Cai, C. Zheng, J. Chen, H.-f. Leung, and Q. Li, "Aligned dual channel graph convolutional network for visual question answering," in *ACL*, 2020, pp. 7166–7176.

[40] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[42] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, 2019, pp. 4580–4590.

[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[44] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL*, 2005, pp. 65–72.

[45] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[46] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[47] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *In IJCV*, pp. 211–252, 2015.

[49] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[50] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[51] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

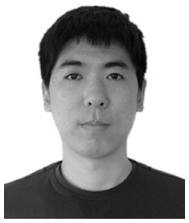
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8024–8035.

[54] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019, pp. 5103–5114.

[55] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," in *ACL*, 2020, pp. 3025–3035.

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



Liang Li received his B.S. degree from Xi'an Jiaotong University in 2008, and Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China in 2013. From 2013 to 2015, he held a post-doc position with the Department of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China. Currently he is serving as the associate professor at Institute of Computing Technology, Chinese Academy of Sciences. He has also served on a number of committees of international journals and conferences. Dr. Li has published over 60 refereed journal/conference papers. His research interests include multimedia content analysis, computer vision, and pattern recognition.

Li Su received the Ph.D. degree in computer science from the Graduate University of Chinese Academy of Sciences, Beijing, in 2009. She is currently an Associate Professor of the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China. Her research interests include image processing and media computing.

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

Chenggang Yan received the B.S. degree in control science and engineering from Shandong University, Shandong, China, in 2008 and the Ph.D. degree in computer science from Chinese Academy of Sciences University, Beijing, China, in 2013. He is currently a professor in the Department of Automation, Hangzhou Dianzi University. His research interests include computational photography and pattern recognition and intelligent system.

Zhengjun Zha received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. He is currently a Full Professor with the School of Information Science and Technology, University of Science and Technology of China, the Vice Director of National Engineering Laboratory for BrainInspired Intelligence Technology and Application. He has authored or coauthored more than 100 papers in these areas with a series of publications on top journals and conferences. His research interests include multimedia analysis, retrieval and applications, as well as computer vision etc. Prof. Zha was the recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia, etc.

Zhengtao Yu received his Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in 2005. He is currently a professor in the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language processing, information retrieval and machine learning.

Qingming Huang received the B.S. degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Chair Professor and the Deputy Dean with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He has coauthored over 400 academic papers in international journals, such as IEEE TPAMI, TIP, TKDE, TMM and TCSVT, and top level international conferences, including NeurIPS, ACM Multimedia, ICCV, CVPR, ECCV, VLDB, AAAI and IJCAI. He is a Fellow of IEEE. His current research interests include multimedia computing, image/video processing, pattern recognition, and computer vision.