Multimodal Interaction Fusion Network Based on Transformer for Video Captioning

Hui Xu, Pengpeng Zeng(✉), and Abdullah Aman Khan

Sichuan Artificial Intelligence Research Institute, Yibin 644000, China
pengpengzeng@gmail.com

Abstract. Learning to generate the description for a video is essentially a challenging task as it involves an understanding of vision and language. Existing methods are mainly based on Recurrent Neural Networks (RNN). Nevertheless, there are some limitations, such as feeble representation power and sequential nature. The transformer-based architecture was proposed to address such issues, and it is widely used in the domain of image captioning. Although it has achieved success in existing methods, the applicability to video captioning is still largely under-explored. To fully explore its significance in video captioning, this paper proposes a novel network by utilizing the transformer for video captioning named Multimodal Interaction Fusion Network (MIFN). To effectively learn the relationship between multiple features, a cross-attention module is introduced within the encoder, which provides a better representation. Moreover, in the decoder, we use a gated mechanism for filtering the essential information to produce the next word. Moreover, we evaluate the proposed approach by using the benchmark MSR-VTT and MSVD video captioning datasets to illustrate its quantitative and qualitative effectiveness and employ extensive ablation experiments to fully understand the significance of each component of MIFN. The extensive experimental results demonstrate that MIFN obtains performance comparable to the state-of-the-art methods.

Keywords: Multimodal · Transformer · Video captioning · Attention

1 Introduction

Recently, the task of vision and language has attracted widespread attention due to one of the key initiatives to achieve artificial intelligence (AI). Meanwhile, the development of deep learning has promoted many multimodal learning tasks such as image-text matching [1], visual question answering (VQA) [2–4] and captioning [5–9]. Captioning task is defined as generating the natural language descriptions of an image or a video. Captioning requires a deep understanding of visual concepts, linguistic semantics, and the alignment of the two. The model needs to recognize the visual information provided to generate accurate and descriptive sentences. However, compared with image captioning, video captioning presents many challenges. First, reasoning over the sequence of images rather than the static image can be more difficult. To describe the video, the model

S. Yang and H. Lu (Eds.): ISAIR 2022, CCIS 1700, pp. 21–36, 2022.
https://doi.org/10.1007/978-981-19-7946-0_3

needs to identify the content of each frame in the video. Secondly, the model needs to consider long-range temporal structures without missing the relationship between the frames in the video. Thirdly, it is a common problem for both image captioning and video captioning to build the relationships between the visual and language to generate descriptions consistent with visual semantics. Thus, video captioning models require reasoning ability on spatial and long-range temporal structures of both video and text to generate an accurate sentence.

The well-known encoder-decoder framework with a Recurrent Neural Network (RNN) is usually used as the basic structure for video captioning tasks. First, a convolutional neural network (CNN) has been adept in our framework to extract frame-based features from an input video. And then, the encoder captures the temporal information to obtain video-based features. Finally, the decoder generates the caption words about the video by taking the output of the encoder. The encoder-decoder framework was optimized end-to-end through a word-level cross-entropy loss. Based on this framework, some researchers have proposed plenty of improved algorithms to upgrade the model for video captioning tasks. For instance, to build the relationship between the caption words and their related frames in the video, an attention mechanism has also been adopted for video captioning, including spatial and temporal attention. To provide a better visual representation of the video, multiple features (appearance and motion) can be extracted from different networks to represent the diverse information.

Although it has succeeded in existing methods, the applicability to video captioning is still largely under-explored. Although it has succeeded in existing methods in video captioning, we observed the following limitations: First, when encoding the multiple features, the encoder neglects intra-modal interactions (e.g., appearance to appearance or motion to motion) and inter-modal interactions (e.g., appearance to motion or motion to appearance). Secondly, the decoder also ignores the self-attention between the output caption words (*i.e.,* word-to-word). Third, the fusion strategy in the decoder is to simply fuse different features, such as concatenation and element-wise adding.

To deal with the above-mentioned issues, we design a novel Multimodal Interaction Fusion network with a transformer (MIFN) to upgrade the effectiveness of the video captioning task. Especially, the property of transformer [10] can capture some intra-modal interactions (e.g., appearance-to-appearance, motion-to-motion, and word-to-word) and inter-modal interactions (e.g., word-to-motion and word-to-appearance) which can address partial limitation and the second limitations. To resolve the first limitation, we inject the cross-attention module to learn the relationship between the input visual features (*i.e.,* appearance-to-motion and motion-to-appearance). Furthermore, we apply a gated mechanism to select the key information from the multiple features.

Our MIFN model achieves comparable performance with the competing methods over MSR-VTT and MSVD datasets, two large-scale video captioning benchmarks. In summary, the major contributions of the proposed method include:

- This paper proposes a novel approach, Multimodal Interaction Fusion Network (MIFN), based on a transformer for video captioning by jointly modeling the interactions with self-attention and co-attention. The MIFN model uses a modular attention block to model a better representation by extracting five types of relationships, e.g.,

word-to-word, appearance-to-motion, motion-to-appearance, word-to-appearance, and word-to-motion.
- We introduce a Gated mechanism in the decoder part of the MIFN model to obtain key information across different modalities.
- The extensive ablation experiments illustrate that MIFN obtains performance improvement over two publicly available MSR-VTT and MSVD datasets for video captioning tasks. In addition, our experimental results obtained comparable results to the state-of-the-art methods.

2 Related Work

Here, we review the last few years studies about video captioning and transformer.

2.1 Video Captioning

Learning to generate video descriptions is a very challenging task that involves understanding both vision and language. Early approaches [11–14] on video captioning is often based on template methods, which apply the word and language rules to design a sentence template. According to the predefined template, the model can align with video content and languages. For example, the work in [13] adopted a Conditional Random Field (CRF) to generate the semantic features for description by modeling the relationships of each two different components from visual inputs. In [14], they proposed a unified framework with two joint models to model two types of features, compositional semantics language and video, in video-text space. However, fixed templates also limit the capabilities of the model for language generation.

The well-known encoder-decoder framework is broadly used in neural machine translation and captioning, which is more flexible than the aforementioned method. The encoder-decoder framework includes encoder and decoder components. In captioning tasks, the former is used to encode video information, and the latter is used to produce human-specific sentences. The work in [7] first averaged each frame feature and then used an LSTM to decode it into captions. The temporal information of video is not adequately used. The attention mechanism is introduced into video captioning to focus on a specific frame in the video relevant to the generated word [15]. The work in [16] proposes a general video caption method (S2VT) without an explicit attention model which learns the temporal structure of the video (optical flow) and applies an LSTM in both encoder and decoder. In [6], they adapt a parallel two-stream 3D-CNN to gain better visual features from the video. Since then, numerous works have adopted multi-modal visual features to upgrade the effectiveness of video captioning. The work in [17] proposes an adaptive attention (hLSTMat) hierarchical LSTM framework to select the key information between the visual content and the language context content and adapt the multi-modal features too.

However, the works mentioned above directly input the various visual outputs from the encoder into the decoder while ignoring the relationship between those features. To learn such a relationship, MIFN injects a cross-attention into the encoder to learn the relationship between the multiple features and a gated mechanism into the decoder to generate the video's description by selecting the key information.

2.2 Video Captioning

The first work on the transformer network [10] introduced a new encoder-decoder architecture. This architecture was applied to machine translation and it achieved better results than the previous works [18–20]. The transformer includes self-attention and cross-attention. The former can characterize the intra-modal interaction within each modality (e.g., appearance-to-appearance), and the latter can characterize inter-modal interaction across different modalities (*i.e.,* Chinese to English). Some work also attempted to use the transformer to solve the captioning task since the captioning involves multimodal interactions. The work in [21] proposed the Object Relation transformer (ORT), explicitly modeling the spatial relationships. The work in [9] introduced EnTangled Attention (ETA) for image captioning that enables the transformer to bridge the relationship between semantic and visual information simultaneously. The work in [8] proposed $\hat{\text{M2}}$, a meshed transformer with memory that model the prior knowledge between the low and high two-level information.

Nevertheless, there are only a few studies that use transformers to deal with video captioning tasks. In our work, we adopt the transformer architecture that introduces the different encoder and decoder designs to deal with video captioning tasks. The encoder learns a better representation from different multimodal features. The decoder can select more important information with the gate mechanism.

3 Preliminaries

Next, this section introduces the primary formulation of the transformer, which is the core structure for our model. The transformer has two basic blocks: the multi-head attention module and the feed-forward network (FFN).

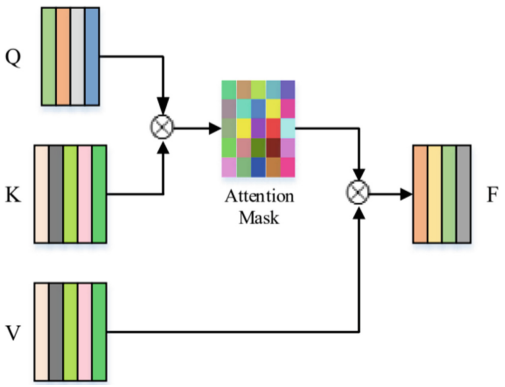

Fig. 1. The Scaled Dot-Product Attention network.

Multi-Head Attention module layer (MHA): The core component of the multi-head attention module is h paralleled scaled independent dot-product attention layers, namely

'Head', as shown in Fig. 1. Specifically, the input of 'Head' contains $q \in R^d$, $k^t \in R^d$, and $v^t \in R^d$, where q is the query, k is the key and belongs to set K, v is the value and belongs to set V, $t \in \{1, 2, ..., n\}$ is the number of key-value pairs, and d is the dimension number. Then, we can achieve the attention weights for the values by calculating the dot products of the query with all keys through a softmax function. Eventually, through this dot-product attention mechanism, we obtain queries-based aware attended feature defined as F. For matrices $Q = [q_1, ..., q_m] \in R^{m\times d}$, $K = [k_1, ..., k_n] \in R^{n\times d} and V = [v_1, ..., v_n] \in R^{n\times d}$, we formulate the above process as:

$$F = \text{Attention}(Q, K, V) = \sigma\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where the dimension of F is $m \times d$, δ is the softmax function.

MHA allows the model not only conduct with a single attention layer but also to explore subspaces. Thus, the final attended feature F will be obtained as follows:

$$\begin{aligned} M - \text{head}(Q, K, V) &= \text{Concat}(H_1, \ldots, H_h)W^O, \\ H_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \end{aligned} \tag{2}$$

where M-head is the short form of multihead, $W_i^K, W_i^V \in R^{d\times d_h}$ are linear function parameters of the $i - th$ 'Head' which respectively project $Q, KandV$ into the latent space. $W_i^V \in R^{d\times d_h}$ is the output projection matrix. We denote the sequence number of the 'Head' as h and set $d_h = d/h$.

Feed-Forward Network layer (FFN): Another basic block of the transformer is FFN. FFN takes the output from the MHA as input and it is implemented by two linear projections to obtain the high-level representation. This process is formulated as:

$$FFN(x) = FC(\text{Re}LU(FC(x))), \tag{3}$$

where both input and output dimensions are d, ReLU is the ReLU activation function. And the dimensionality of inner-layer is usually $d_{ff} = 2048$.

Both encoder and decoder of transformer consist of multiple above building blocks where each building block is composed of the MHA and FFN modules. According to whether the input features are the same, the MHA module can be divided into self-attention and cross-attention two types. The original encoder of the transformer only has self-attention, which can characterize the intra-modal interaction within each modality. The difference in the decoder is that it has these two attention mechanisms, and cross-attention can characterize Inter-modal interaction across different modalities. Besides, each layer follows layer normalization (LayerNorm), and the residual connection is used for all building blocks.

4 Methodology

For a video sequence$V = [v_1, ..., v_m]$, the goal of video captioning is to produce a natural sentence $S = [s_1, ..., s_n]$ that expresses the semantic meaning of the video, where m is the length of frames, n is the length of a sentence.

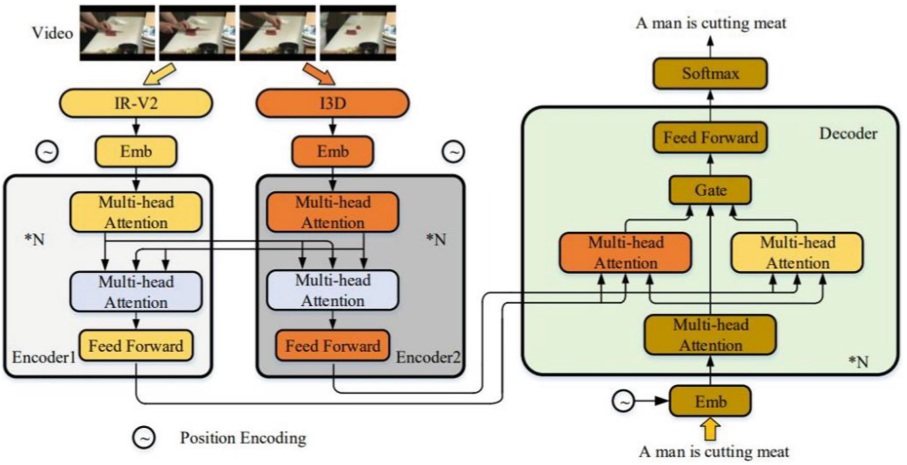

Fig. 2. The framework of the proposed MIFN. We first inject the cross-attention module into the encoder of the transformer to fully exploit the differential but related semantic information. Then, we select the key information to generate textual captions by representing various features through the encoder.

Next, we detail the proposed framework, *i.e.*, Multimodal Interaction Fusion Network (MIFN) based on the transformer for video captioning, which can upgrade the quality of the produced sentence descriptions. In our method, we first adopt the basic structure of the transformer and then modify its internal structure to meet our specific needs. Specifically, a cross-attention module is injected into the encoder of the transformer to fully utilize the different but relevant semantic information. Furthermore, since we get representations of various features through the encoder, we need to produce textual captions by choosing the key information. Thus, we propose a gated mechanism to deal with such issue. The overall framework of the MFIN is illustrated in Fig. 2. The following sections provide details about each component.

4.1 Encoder with Multimodal Features

The process of the encoder includes two stages. The first stage aims to extract the different visual representations, namely appearance and motion features, defined as a multimodal feature extraction module. Moreover, the second stage aims to learn the relationship between the multimodal visual representation to obtain the final video features.

4.2 Multimodal Feature Extraction Module

For appearance representation, we utilize Inception-Resnet-v2 (IR-v2) network [22] pre-trained on ILSVRC-2012-CLS dataset [23] to extract it. The input of each frame v_i in the video to the IR-v2 network is resized to 299×299. The output feature of the last pooling layer is used to encode the frame-based visual feature $a_i \in R^{1536}$, where 1536 is the dimension. Therefore, the total appearance features A of video V is defined as

follows:

$$\begin{aligned} A = [a_1, \ldots, a_m], a_i \in R^{1536} \\ a_i = IR - v2(v_i), \end{aligned} \tag{4}$$

where v_i represents the RGB image of the frame in the video.

For motion representation, we adopt the Inflated 3D ConvNet (I3D) [24] pre-trained on the Kinetics dataset [12] to obtain clip-based visual features. First, we use both horizontal and vertical directions between each frame and adjacent frames, using TVL1-flow networks [25], to calculate the optical flow information. Then, each continuous 64 optical flow flips are input to the I3D network to obtain 2048-dimensional motion features $m_i \in R^{2048}$. We formulate this process as below:

$$\begin{aligned} M = [m_1, \ldots, m_m], a_i \in R^{1536} \\ m_i = I3D(c_i) \end{aligned}, \tag{5}$$

where c_i represents a clip, the clip duration is 64 frames.

In order to represent sequence order for the transformer model, the model uses position encoding (PE) to encode the position information of the frames.

We formulate the process of PE as:

$$\begin{aligned} PE_{(\text{ position },2i)} = \sin\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \\ PE_{(\text{ position },2i+1)} = \cos\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \end{aligned}, \tag{6}$$

where i is the dimension. The PE chose the sinusoidal version.

4.3 Cross-Attention Module with Multi-modal Features

The cross-attention module is composed of two identical encoders, and each encoder encodes a different video feature. Each encoder encodes its features without interaction, and there is no relationship between the learned features. To deal with this issue, this paper introduces a cross-attention strategy to the encoder that modifies the transformer's original encoder. Each encoder stack L cross-attention block, and each cross-attention block has three sub-networks, a self-attention network, a cross-attention network, and a FFN. Take the outputs of the $l - th(0 \leq l < N)$ block of two encoders $V_A^i \in R^{m \times d}$ and $V_M^i \in R^{m \times d}$ as an example. Specifically, the input from the $(i+1) - th$ self-attention network is the output for the $i - th$ block, which learns Intra-modal interaction within each modality.

$$D_{.,t}^{l+1} = MHA_{\text{ self}}\left(H_{.,t}^{l}, H_{\leq t}^{l}, H_{\leq t}^{l}\right), \tag{7}$$

And then the $V_A^{i^{`}} and\ V_M^{i^{`}}$ are input to the $(i+1) - th$ cross-attention network to learn inter-modal interaction across different modalities. This process is formulated as follows:

$$\begin{aligned} g^a = MHA_{\text{ cross}}\left(d_t, V_A^L, V_A^L\right) \\ g^m = MHA_{\text{ cross}}\left(d_t, V_M^L, V_M^L\right) \end{aligned}, \tag{8}$$

Before feeding into this module, the appearance feature A and the motion feature M are mapped into $V_A^0 \in R^{m\times d}$ and $V_M^0 \in R^{m\times d}$, and the residual connection and layer normalization are used.

4.4 Decoder with Gated Mechanism

Given previously generated words and video features, the decoder generates the next tokens of the output caption. Note that the video features extracted from multiple encoders focus on aspects of the video that have different importance to the generated words, making it challenging to select key information from different features. To address such an issue, our decoder block inserts a gated mechanism module between the MHA sub-layer and FFN sub-layer, which can empower the ability of the decoder block to perform attention over the difference encoder outputs simultaneously.

Given the previously generated words$S_{<\mathrm{t}} = [s_0, ..., s_{t-1}]$, the decoder is to generate the next $t-th$ word. Each word $w_t \in R^d$ is represented by a vector for the word at position t in the sentence. Moreover, w_t is obtained by word embedding and positional encoding. Notably, w_0 represents the start of a sentence.

Similar to the encoder, the decoder has N identical blocks, and each block consists of five sub-layers, namely oneMHA_{self}, twoMHA_{cross}, one a gated mechanism and one FFN. For the $(l+1)-th$ block, the output of the $l-th$ block $H^l \in R^{t\times d} = [h_1^l, ..., h_t^l]$, are fed into a $(l+1)-th$ MHA_{self} sub-layer in the $(l+1)-th$ block, notice that h_0^l corresponds to w_{t-1} :

$$\begin{aligned} c\_a_t &= sigmoid(W_a\left[d_t, g^a\right] + b_a \\ c\_m_t &= sigmoid(W_m\left[d_t, g^m\right] + b_t \\ c_t &= Relu(\left[(c\_a_t \odot f(g^a)) \oplus (c\_m_t \odot f(g^m)\right]) \end{aligned} \quad , \tag{9}$$

where $H_{.,\mathrm{t}}^l \in R^{1\times d}$, $E_{.,\mathrm{t}}^{\mathrm{l}=1} \in R^{1\times d}$, and $h_0^l = s_{t-1}$. Subsequently, the MHA_{self} output $D_{.,\mathrm{t}}^{l+1}$ is passed into the two MHA_{cross} to provide the proper guidance for the attention in different target modalities. The process can be formulated as follows:

$$L_\theta = -\sum_{t=1}^{N} \log P(w_t|w_{<t}, V, \theta), \tag{10}$$

where V_A^L and V_M^L are the output of the two encoders.

The gating mechanism is to select key information between the different modalities and flow to the subsequent layers. Such a gate is good at dealing with gradient explosion and vanishing, enabling information to spread unimpeded through a long time step or a deeper layer. The context gates c_{a_t} and c_{m_t} are determined by the current self-attention output d_t, the appearance guidance g^a and the motion guidance g^m.

$$F = Attention(Q, K, V) = \mathrm{MultiH}(\mathrm{QUOTE\ ta_l}h_{V_l^{enc}}^u, \mathrm{W_{k_1}}^{TA_l} h_{TA_l}^u, \mathrm{W_{v_1}}^{TA_l} h_{TA_l}^u$$

where W_a and W_m are parameters that needed to be learned, and b_a and b_t are bias terms. Note that [,], $\odot$, and $\oplus$ is the concatenation, element-wise multiplication and

element-wise addition operation separately.$f(.)$ can be an activation function. We feed the c_t into the FFN sub-layer to obtain the final feature $H_{\leq t}^{t+1} \in R^{t \times d}$.

$$F = Attention(Q, K, V) = \text{MultiH}(\text{QUOTE ta}_{\text{l}} h_{V_l^{enc}}^{u}, \text{W}_{\text{k}_1}^{TA_l} h_{TA_l}^{u}, \text{W}_{\text{v}_1}^{TA_l} h_{TA_l}^{u}$$

Suppose the decoder is generating the $n-th$ word of the target sentence. Similar to the encoder, the decoder also consists of N uniform multi-head layers. Each layer consists of three MHA modules, more preciously, two FFN modules and one gated mechanism module. Moreover, the first two MHA modules learn motion-guided attention from the caption words separately (appearance to word and motion to word) and the appearance-guided attention from the caption words. The last MHA module models the self-attentions on the caption words (word to word).

4.5 Training Processing

In the training process, we follow the standard protocol of video captioning to implement the training process. Specifically, we pre-train the model by using a word-level cross-entropy objective:

$$F = Attention(Q, K, V) = \text{MultiH}(\text{QUOTE ta}_{\text{l}} h_{V_l^{enc}}^{u}, \text{W}_{\text{k}_1}^{TA_l} h_{TA_l}^{u}, \text{W}_{\text{v}_1}^{TA_l} h_{TA_l}^{u}$$

where N denotes the length number of the sentence, and θ denotes the model parameter.

5 Experiments

In this section, we demonstrate the effectiveness through extensive experiments. we first introduce two public datasets used for the video captioning task. Then, we compare MIFN with other competing baselines. Finally, qualitative results of ablation experiments are conducted to verify the effectiveness of each component.

5.1 Datasets

To evaluate the effectiveness of the proposed video captioning models, this part reports the results over MSR-VTT and MSVD, two large-scale video captioning datasets.

MSR-Video to Text (MSR-VTT) dataset: MSR-VTT dataset collected by [5] is one of the largest video captioning datasets for generating the description of the video, which contains 10,000 video clips. The duration of each clip is between 10 and 30 s, and each video clip is annotated with approximately 20 different captions by Amazon Mechanical Turk (AMT) workers. The total of video-sentence pairs is 200K. Following the existing work, we use the splits provided by [5], *i.e.,* 6,513 for training, 497 for validation, and 2,990 for testing respectively.

Microsoft Video Description (MSVD) dataset: MSVD dataset consists of 1,970 video clips, each of them also annotated by AMT workers. Each video clip has roughly 40 descriptions and the dataset has approximately 80,000 video-sentence pairs. Similar to the existing works [16, 26], there are 1,200 training video clips, 100 validation video clips, and 670 test video clips.

Table 1. The experiment results over MSR-VTT dataset.

Model	B@1 B@2 B@3 B@4 M R C
MP-LSTM [7]	- - - 30.4 23.7 52.0 35.0
Soft-Attention [15]	- - -28.5 25.0 53.3 37.1
Res-Attention [31]	77.1 62.1 48.7 37.0 26.9 – 40.7
S2VT [16]	- - - 28.5 25.0 52.0 37.1
v2t_navigater [32]	- - - 40.8 28.2 60.9 44.8
Aalto [33]	- - - 39.8 26.9 59.8 45.7
VideoLAB [34]	- - - 39.1 27.7 60.6 44.1
hLSTMat [17]	76.2 62.9 50.6 39.7 27.0 – 42.1
RecNet [35]	- - - 39.1 26.6 59.3 42.7
PickNet [36]	- - - 41.3 27.7 59.8 44.1
LSGN + LNA [37]	- - - 39.5 27.4 60.9 46.5
baseline (our)	- - - 38.5 26.8 58.9 44.8
MIFN (our)	78.6 65.1 51.9 41.0 27.6 60.1 46.6

5.2 Implementation Details

Preprocessing. To preprocess the text information, we first convert all words to lowercase letters, remove the punctuation and add three additional tokens, namely unknown $< UNK >$, the begin-of-sentence $< BOS >$, and the end-of-sentence $< EOS >$. Thus, it yields a vocabulary of 23,665 and 15,906 in size for the MSR-VTT and MSVD datasets, respectively.

Evaluation Settings: Following the previous works, we employ the same captioning metrics: BLEU [27], METEOR [28], ROUGE [29], and CIDEr [30].

Model Settings & Training: All MHA layers and FFN layers have 512-dimensional, and the input feature's dimension of the co-encoder is mapped into 512 with a linear projection. We represent words with word embeddings, whose size is also set as 512. During training, the model is optimized using Adams' optimizer. The batch size is set as 50, and a beam size is equal to 5. Besides, we introduce gradient clipping and dropout technology during training.

5.3 Comparison with Competing Baselines

Here, we present the results of our evaluation followed by a comparison of the proposed method with several competing baselines over MSR-VTT and MSVD datasets.

5.3.1 The Experiment Results Over MSR-VTT Dataset

Comparing Methods. We first consider the MSR-VTT dataset and compare the performance of the proposed approach to other competing methods. Specifically, we compare MIFN with Mean-Pooling [7], Soft-Attention [15], Res-Attention [31] S2VT [16], v2t_navigater [32], Aalto [33], VideoLAB [34], hLSTMat [17], RecNet [35], PickNet [36], and LSGN + LNA [37].

Results: The experimental results are demonstrated in Table 1. In this experiment, the proposed method (MIFN) achieves better performance than other baseline models. The results show that the Cider of our MIFN can achieve 47.6, which exceeds the highest performance reported over MSR-VTT dataset. For the rest of the evaluation metrics, our model obtains comparable scores with 41.0 B@4 (vs. 41.3 B@4), 27.6 Meteor (vs. 28.2 Meteor), and 60.1 Rough (vs. 60.0 Rough). It proves the effectiveness of our approach.

5.3.2 The Experiment Results over MSVD Dataset

Comparing Methods: Next, we assess our model over MSVD dataset. For the experiments, we compare the proposed MIFN with previous works, *i.e.,* S2VT [16], Res-Attention [31], hLSTMat [17], HRNE [38], MA [39], SCN [40], SCN [40], TSA [41], RecNet [35], PickNet [36], and ASGN + LNA [37].

Results: We demonstrate the experimental results in Table 2. We can see that the proposed MIFN model outperforms other competing baselines. The proposed method achieves improvements of 1.3 B@4, 0.6 Meteor, 1.1 Rough, and 2.3 Cider respectively, compared to the LSGN + LNA model [37].

Table 2. The Experiment Results over MSVD Dataset.

Model	B@4 M R C
S2VT [16]	- 29.2 -
Res-Attention [31]	53.4 34.3 72.9
hLSTMat [17]	33.5 – 72.8
HRNE [38]	43.8 33.1 - -
MA [39]	50.4 31.8 – 69.9
SCN [40]	51.1 33.5 – 77.7
TSA [41]	51.7 34.0 – 74.9
RecNet [35]	52.334.169.880.3
PickNet [36]	52.3 33.3 69.6 76.5
ASGN + LNA [37]	52.1 33.3 70.3 80.3
MIFN (Our)	53.4 33.9 71.4 82.6

5.4 Ablation Study

We employ extensive ablation studies to illustrate the effectiveness of each component over the MSR-VTT dataset.

5.4.1 The Effect of Transformer

Here, we demonstrate the effectiveness of the transformer. The results are shown in Table 3 which has three blocks, and each block takes a different feature as an input. Especially, A and M denote appearance features and motion features, respectively.

Moreover, $A + M$ denotes the fusing of the appearance feature and motion feature by simple concatenating. Furthermore, each block is divided into three categories: 1) LSTM-ATT model, which uses a two-layer LSTM with attention mechanism, 2) transformer with three layers, and 3) transformer with six layers.

As shown in Table 4, we first see that the performance of $A + M$ is better than A and M. This result shows the beneficial effects of multiple features. Then, the result illustrates that the original transformer for video captioning in effectiveness. The original configuration of the transformer is six layers and self/cross attention. Comparing LSTM (Row 1) with transformer (Row 3) in the first block of Table 4, it has obtained the improvement by the original transformer, by 0.4 in B@4, 0.3 in Meteor, 0.3 in Rough, and 1.5 in Cider, respectively. Furthermore, by changing the number of layers of the transformer (three layers vs. six layers), we notice that the performance has also slightly increased. This validates the impact of the number of transformer layers. The possible reason is that training data quantity is reduced, and the complexity of the sentence is lower for video description compared with machine translation. In the following experiments, we utilize multiple features ($A + M$) to train our model and adopt the transformer with three layers.

5.4.2 Effects of Proposed Individual Components

Here, an ablation study is designed to verify the effect of our MIFN individual component in MIFN. The results are shown in Table 3. The baseline model is the basic-transformer encoder with three layers, which takes input by simply concatenating multiple features into one visual feature. Comparing Row 0 and Row 1, we observe the efficiency of the total model MIFN. The improvement over the baseline is significant, by 2.5 in B@4, 0.8 in Meteor, 1.2 in Rough, and 2.0 in Cider, respectively. Then we start with the MIFN model and successively remove Cross-attention from the encoder and the Gated mechanism from the decoder to demonstrate their importance.

Effects of the encoder with cross-attention: By removing the Cross-attention in the encoder (Row 2), we notice that the model has fallen on the MSR-VTT dataset and have fallen by 0.2, 0.7, 0.7, and 0.4 on B@4, Meteor, Rough, and Cider, respectively, almost fallen on all evaluation metrics. These results show the effectiveness of such cross-attention from the encoder to generate a better visual feature from the video.

Effects of decoding with gated mechanism: Further, we verify the strength of decoding with a gated mechanism by comparing it (Row 1) with the other model (Row 5) that replaces the gated mechanism with the concatenation operation in the decoder. The result of Row 4 (the RCM model with only supervised learning) validates the superiority of the gated mechanism in the decoder. The gated mechanism in the decoder improves the performance of video caption, particularly for B@4 by 0.6, Meteor by 0.5, Rough by 0.4, and cider by 0.9, respectively.

Table 3. The effectiveness of the transformer in the MSR-VTT dataset. The table has three blocks, and each block uses a different feature to train the model. A and M denotes the appearance feature and motion feature, respectively. A + M denotes the fusing between the appearance feature and motion feature by simple concatenation. L represents the number of the transformer's layers.

Model		B@4 M R C
A	LSTM	37.8 26.0 58.3 41.4
	transformer (L = 3)	38.0 26.5 58.7 43.5
	transformer (L = 6)	38.2 26.3 58.6 42.9
M	LSTM	37.2 25.3 57.9 39.9
	transformer (L = 3)	37.8 26.2 58.3 42.5
	transformer (L = 6)	37.5 25.8 58.2 41.8
A + M	LSTM	38.0 26.2 58.5 42.3
	transformer (L = 3)	38.5 26.8 58.9 44.8
	transformer (L = 6)	38.3 26.4 58.7 44.3

Table 4. Ablation study on MSR-VTT. The result demonstrates the performance of the basic transformer as the baseline, which takes features by simply concatenating multiple features as input. Row 1–3 shows the influence of individual components by removing them from the final model (Row 1).

#	Model	B@4	M	R	C
0	Baseline (transformer with three layers)	38.5	26.8	58.9	44.8
1	MIFN	41.0	27.6	60.1	46.6
2	—Cross-attention in the encoder	39.8	26.9	59.4	45.5
3	—Gated mechanism in the decoder	39.4	27.1	59.7	45.9

6 Conclusion

In this paper, we propose a novel framework, named Multimodal Interaction Fusion Network (MIFN) which is based on the transformer, for the video captioning task. Firstly, we inject the cross-attention module into the original transformers' encoder to learn the relationship between the input visual feature which provides a better visual representation of video. Besides, we apply a gated mechanism to replace the simple fusion strategy, which can select the key information from the multiple features in the decoder. Experimental results on MSR-VTT and MSVD datasets illustrate that MIFN achieves performance comparable with the competing methods. Moreover, extensive ablation studies also indicate the effectiveness of the proposed model for video captioning.

Acknowledgment. This work is supported by the Sichuan Science and Technology Program (No. 2022119).

References

1. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4654–4662 (2019)
2. Antol, S., et al.: VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, pp. 2425–2433 (2015)
3. Teney, D., Anderson, P., He, X., Van Den Hengel A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4223–4232 (2018)
4. Gao, L., Zeng, P., Song, J., Liu, X., Shen, H.T.: Examine before you answer: multi-task learning with adaptive-attentions for multiple-choice vqa. In: Proceedings of the 26th ACM international conference on Multimedia, pp. 1742–1750 (2018)
5. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 5288–5296 (2016)
6. Zhang, C., Tian, Y.: Automatic video description generation via lstm with joint two-stream encoding. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2924–2929, IEEE (2016)
7. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1494–1504 (2015)
8. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: M ˆ2: Meshed-memory transformer for image captioning. arXiv preprint arXiv:1912.08226 (2019)
9. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8928–8937 (2019)
10. Vaswani, A.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
11. Guadarrama, S.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2712–2719, (2013)
12. Kay, W., et al., The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
13. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 433–440 (2013)
14. Xu, R., Xiong, C., Chen, W., Corso, J.J.: Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
15. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
16. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
17. Gao, L., Li, X., Song, J., Shen, H.T.: Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Trans. Pattern Anal. Mach. Intell. p. 1 (2019)

18. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1243–1252 (2017)
19. Shazeer, N., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
20. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
21. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: transforming objects into words. In: Advances in Neural Information Processing Systems, pp. 11135–11145 (2019)
22. Szegedy, C., Loffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)
23. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
25. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: TV-L1 optical flow estimation. Image Processing On Line 2013, pp. 137–150 (2013)
26. Pan,Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4594–4602 (2016)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318 (2002)
28. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
29. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, (Barcelona, Spain) (2004)
30. Vedantam, R., Zitnick, C. L., Parikh, D.: CIDER: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015)
31. Li, X., Zhou, Z., Chen, L., Gao, L.: Residual attention-based LSTM for video captioning. World Wide Web **22**(2), 621–636 (2019)
32. Jin, Q., Chen, J., Chen, S., Xiong, Y., Hauptmann, A.: Describing videos using multi-modal fusion. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1087–1091 (2016)
33. Shetty, R., Laaksonen, J.: Frame-and segment-level features and candidate pool evaluation for video caption generation. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1073– 1076 (2016)
34. Ramanishka, V., et al.: Multimodal video description. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1092–1096 (2016)
35. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7622–7631 (2018)
36. Chen, Y., Wang, S., Zhang, W., Huang, Q.: Less Is More: Picking Informative Frames for Video Captioning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 367–384. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_22

37. Xiao, X., Wang, L., Fan, B., Xiang, S., Pan, C.: Guiding the flowing of semantics: Interpretable video captioning via POS tag. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2068–2077 (2019)
38. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1029–1038 (2016)
39. Hori, C., et al.: Attention-based multimodal fusion for video description," in Proceedings of the IEEE International Conference on Computer Vision, pp. 4193–4202 (2017)
40. Gan, Z., et al.: Semantic compositional networks for visual captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5630–5639 (2017)
41. Wu, X., Li, G., Cao, Q., Ji, Q., Lin, L.: Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6829–6837 (2018)
42. Lu, H., Li, Y., Chen, M., et al.: Brain Intelligence: go beyond artificial intelli-gence. Mobile Netw. Appl. **23**, 368–375 (2018)
43. Lu, H., Zhang, Y., Li, Y., et al.: User-oriented virtual mobile network resource management for vehicle communications. IEEE Trans. Intell. Transp. Syst. **22**(6), 3521–3532 (2021)