Paper:

A Multi-modal Fusion Algorithm for Cross-modal Video Moment Retrieval

Mohan Jia*, Zhongjian Dai, Zhiyang Jia, Yaping Dai

School of Automation, Beijing Institute of Technology, Beijing, 100081, China
*E-mail: mohana620@163.com

In order to localize the most relevant moment in an untrimmed video according to the given sentence query, a multi-modal fusion algorithm for cross-modal video moment retrieval is proposed. The key idea of the multi-modal fusion algorithm is to separate the consistent features from the information across modalities and then concatenate them. The concatenated consistent features are taken as the input of the coordinates prediction network, and the regression temporal coordinates are obtained. The correlations between the consistent component pairs can enhance the expressiveness and interaction of the video features and sentence features, thus improves the accuracy. The effect of the proposed algorithm is verified on a public benchmark dataset: TACoS. The results show the effectiveness of the proposed algorithm as compared with CTRL and ACRN under TACoS dataset.

Keywords: Video Moment Retrieval, Multi-modal Fusion, Deep Learning, Attention Mechanism

1. INTRODUCTION

To watch the specific video clip which is relevant to the query, the entire video needs to be browsed to localize the relevant part in a video, which could take hours. While cross-modal video moment retrieval task can improve the retrieval efficiency by finding relevant video clips with accurate temporal boundaries (i.e., the start time and the end time) for a given query. The high practical value of the task in multimedia retrieval has attracted more and more attention [2, 12].

Video and sentence information need to be deeply incorporated to distinguish the details of different video clips and perform accurate localization. Most existing methods for this task focus on learning the cross-modal relations between the video and the sentence. Gao et al. used vector element-wise addition, vector element-wise multiplication, and vector concatenation to combine the information from both modalities [3]; Hendricks et al. presented a Moment Context Network (MCN) that can incorporate the contextual information to enhance the video clip representations [1]; Liu et al. thought that some previous work ignored the temporal and spatial information in the temporal video clips and the query, so they utilized the temporal memory attention network to explore the attentive contextual visual features of the moments and textual features of the query respectively [6, 7]. Besides they also enhanced the multi-modal representations of clip-query pairs with the outer product of their features. Yuan et al. proposed a multi-modal co-attention mechanism to generate both video and sentence attentions [11]. Zhang et al. leveraged self-attention to capture long-range semantic dependencies in video encoding[13].

However, the interaction between the video features and sentence features cannot be well modelled by the concatenation or the attended video features in the current methods, and the modal correlation also cannot be explored to enhance the expressiveness of each modality. Therefore, a multi-modal fusion algorithm for cross-modal video moment retrieval is proposed. The information across modalities can be categorized into two parts: the consistent component, and the complementary component. To separate consistent features from the information across modalities, each pair of video feature and sentence feature are assigned with a relation score. Next, the consistent features are concatenated as the input of the coordinates prediction network. The correlations between the consistent component pairs can enhance the expressiveness and interaction of the two modalities, thus improves the accuracy.

The proposed multi-modal fusion algorithm using consistent features(MMFA-CF) mainly includes three modules: feature encoding network, multi-modal co-attention interaction network and coordinates prediction network. The widely used 3D convolutional networks(C3D) is applied to encode the video clips in the feature encoding network. And to preserve video and sentence contexts, Bi-directional LSTM net is exploited to generate video clip representations. In the multi-modal co-attention interaction network, the multi-modal co-attention mechanism is introduced to set up a symmetry interaction between the video and the sentence query. In the coordinates prediction network, the concatenation of the consistent features separated from video features and sentence features are taken as the input, and then the temporal coordinates are obtained.

The effects of MMFA-CF are verified on a public dataset TACoS [9]. Compared with CTRL, MMFA-CF achieves 8.1% improvement in R@1, IoU@0.1, 1.7% improvement in R@1, IoU@0.3, and 0.8% improvement in mIoU. And compared with ACRN, MMFA-CF achieves

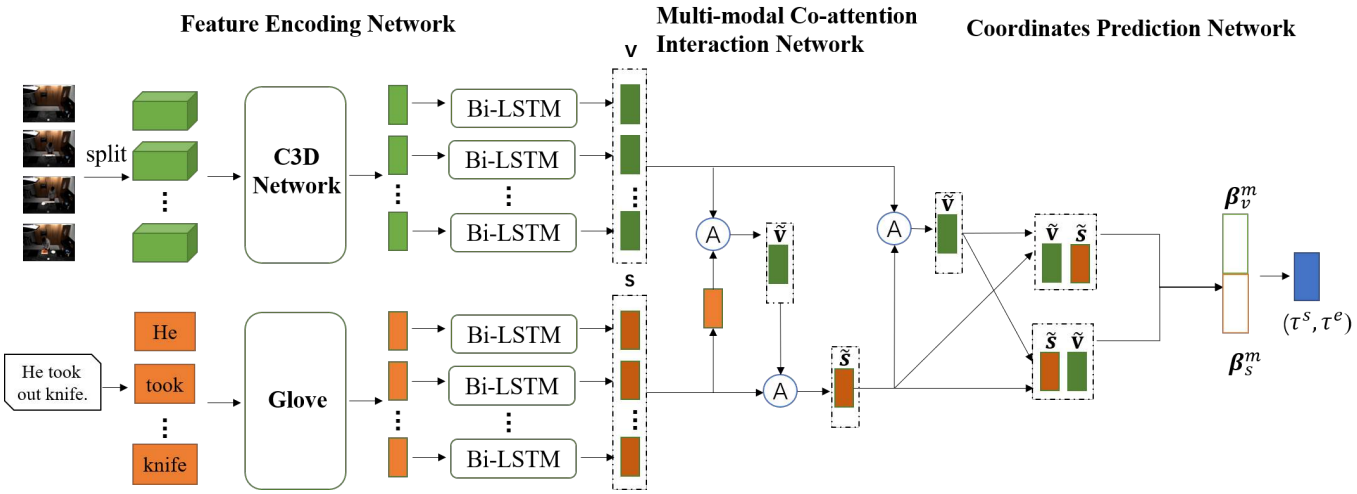

Fig. 1. The framework of MMFA-CF.

7.4%, 0.1%, 0.3% improvement in R@1, IoU@0.1, R@1, IoU@0.3, and mIoU respectively. The results show the effectiveness of MMFA-CF as compared with CTRL and ACRN under TACoS dataset.

Section 2 introduces the structure of MMFA-CF. The implementation and experimental verification of MMFA-CF are described in Section 3. Finally, the conclusions and the future work plans are discussed in Section 4.

2. METHOD

MMFA-CF includes three modules: feature encoding network, multi-modal co-attention interaction network and coordinates prediction network. In this section, the basic theory of the three modules are elaborated. The framework of MMFA-CF is shown in **Fig. 1**.

2.1. Problem Statement

Let V denotes a video and S denotes a sentence description of a video clip. The sentence description S is affiliated with a temporal annotation (τ^s,τ^e), where τ^s and τ^e are the start time and end time of the desired moment. Given the input video and sentence, our task is to predict the corresponding temporal coordinates (τ^s,τ^e).

2.2. Feature Encoding Network

Video Encoding: Firstly, each untrimmed video V is evenly split into M video clips $\{v_1,\ldots v_j,\ldots v_M\}$ in chronological order. Then, the widely used C3D network [10] is applied to encode these video clips. Finally, Bi-directional LSTM is used to generate video clip representations incorporated with the contextual information. Specific definitions are as (1)-(4):

$$\mathbf{x}_j = C3D(v_j) \tag{1}$$

$$\mathbf{h}_j^f, \mathbf{c}_j^f = LSTM^f\left(\mathbf{x}_j, \mathbf{h}_{j-1}^f, \mathbf{c}_{j-1}^f\right) \tag{2}$$

$$\mathbf{h}_j^b, \mathbf{c}_j^b = LSTM^b\left(\mathbf{x}_j, \mathbf{h}_{j+1}^b, \mathbf{c}_{j+1}^b\right) \tag{3}$$

$$\mathbf{v}_j = f\left(\mathbf{W}_v\left(\mathbf{h}_j^f || \mathbf{h}_j^b\right) + \mathbf{b}_v\right) \tag{4}$$

Here $\mathbf{x}_j$ is the fc6 layer C3D features of the video clip v_j. The Bi-directional LSTM consists of two independent streams, in which $LSTM^f$ moves from the start to the end of video and $LSTM^b$ moves end to start. The representation $\mathbf{v}_j$ of video clip v_j is obtained by transforming the concatenation of the forward and backward LSTM outputs at position j. $f()$ is the activation function, i.e., Rectified Linear Unit (ReLU).

Sentence Encoding: The Bi-directional LSTM is also employed to represent sentence. Unlike the general LSTM which encodes sentence as a whole, the Bi-directional LSTM takes a sequence of N words $\{s_1,\ldots s_j,\ldots s_N\}$ from sentence S as inputs and encodes each word s_j into a contextual incorporated feature vector $\mathbf{s}_j$. And the input features are 300-D glove [8] word features. Finally, the sentence representation is obtained, which is denoted as $\mathbf{S}=[\mathbf{s}_1,\ldots \mathbf{s}_j,\ldots \mathbf{s}_N] \in R^{h_s \times N}$.

2.3. Multi-modal Co-attention Interaction Network

Because a large part of video contents are irrelevant to the sentence and only a small clip satisfies the sentence description. Therefore, the attention mechanism is an appropriate choice to focus on the important video clips. In order to explicitly capture the varying importance of each context clip, an attentive weight is assigned to each video representation. Similarly, different words in a sentence

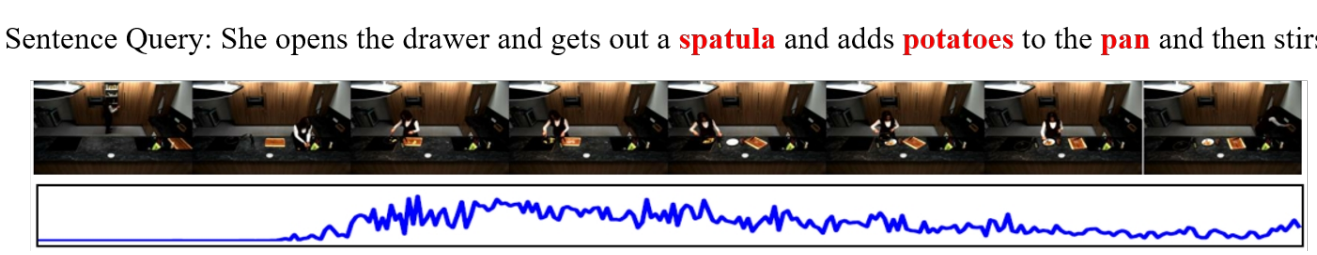

Fig. 2. Qualitative results of MMFA-CF.

have varying importance for semantic expression. Highlighting the key words or phrases by the attention mechanism can provide the localization procedure a more clear target.

In the multi-modal co-attention interaction network, the attention weights are added to the video and the sentence sequentially. The specific process is as step (1)-(3):
(1)Attend to the video based on the initial sentence feature.
(2)Attend to the sentence based on the attended video feature.
(3)Attend to the video again based on the attended sentence feature.

Specifically, the attention function $\widetilde{\mathbf{z}} = A(\mathbf{Z};\mathbf{g})$ takes the video (or sentence) feature $\mathbf{Z}$ and the attention guidance $\mathbf{g}$ derived from the sentence (or video) as inputs, and outputs the attended video (or sentence) feature as well as the attention weights. Concrete definitions are as (5)-(6):

$$\mathbf{H} = tanh\left(\mathbf{U}_z\mathbf{Z} + (\mathbf{U}_g\mathbf{g})\,1^T + \mathbf{b}_a\mathbf{1}^T\right) \tag{5}$$

$$\mathbf{a}^z = softmax\left(\mathbf{u}_a^T\mathbf{H}\right) \tag{6}$$

$$\widetilde{\mathbf{z}} = \sum a_j^z \mathbf{z}_j \tag{7}$$

Here $\mathbf{U}_g, \mathbf{U}_z \in R^{k\times h}, \mathbf{b}_a, \mathbf{u}_a \in R^k$ are parameters of the attention function, $\mathbf{1}$ is a vector with all elements to be 1, $\mathbf{a}^z$ is the attention weights of $\mathbf{Z}$, $\widetilde{\mathbf{z}}$ is the weighted sum feature.

In the first step of generating attention, $\mathbf{Z} = \mathbf{V}$ and $\mathbf{g}$ is the average representation of words in the sentence. In the second step, $\mathbf{Z} = \mathbf{S}$ and $\mathbf{g}$ is the intermediate attended video feature from the first step. In the last step, the video is again added attention based on the attended sentence feature from the second step.

2.4. Coordinates Prediction Network

In this task, the high correlation between video features and sentence features needs to be utilized to locate specific video clips. Since the information across modalities can be categorized into two parts: the consistent component and the complementary component. To adequately exploit the correlations between the consistent component pairs, the consistent component and the complementary component are split by a relation-aware attention mechanism. Specifically, the computed attention score can be used to measure the correlation between the video features and sentence features. Then, a threshold is learned for each modality to distinguish the consistent and complementary features according to the score. Finally, the consistent parts are integrated to enhance the representations. This process can be represented as (8):

$$\mathbf{s}^m = softmax\left(\mathbf{W}^m \cdot [\mathbf{v}^m, \mathbf{s}^m] + \mathbf{b}^m\right), \tag{8}$$

where $\mathbf{W}^m \in R^{D_g\times D}$ and $\mathbf{s}^m \in R^{D_h}$ denote the learnable weight matrix and relation score vector corresponding to each dimension, respectively; the D_h denotes the dimension of the video feature, and D is the dimension of the sentence feature. The $\mathbf{b}^m$ is the bias terms.

Having established the attentive relation scores, a trainable threshold ξ_o^m is used to separate out the consistent vector from the relation score vector, namely $\boldsymbol{\gamma}_o^m$. The element in the consistent vectors is defined as (9):

$$\boldsymbol{\gamma}_o^m[i] = \begin{cases} \mathbf{s}_o^m[i], & \text{if } \mathbf{s}_o^m[i] \geq \xi_o^m \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $\boldsymbol{\gamma}_o^m[i]$ is the value of i-th the dimension in the consistent weight vector $\boldsymbol{\gamma}_o^m$, indicating the degree of the consistency.

Particular, since the original functions are not continuous, a sigmoid function is introduced to make them differentiable, as (10):

$$\boldsymbol{\gamma}_o^m[i] = \frac{\mathbf{s}_o^m[i]}{1+e^{-w*(\mathbf{s}_o^m[i]-\xi_o^m)}}, \tag{10}$$

where w denotes a scalar weighting the difference between $\mathbf{s}_o^m[i]$ and ξ_o^m.

Based on these weight vectors, the consistent features are separated from the mixed information, which are the element-wise products of the original feature vector and each weight vector, as (11)-(12):

$$\boldsymbol{\beta}_v^m = \mathbf{v}^m \otimes \boldsymbol{\gamma}_v^m \tag{11}$$

$$\boldsymbol{\beta}_s^m = \mathbf{s}^m \otimes \boldsymbol{\gamma}_s^m \tag{12}$$

Based on the enhanced features, the enhanced video feature $\boldsymbol{\beta}_v^m$ and the enhanced sentence feature $\boldsymbol{\beta}_s^m$ are fused to a multi-modal representation $\mathbf{f}$, and then the temporal coordinates are regressed as (13)-(14):

$$\mathbf{f} = f\left(\mathbf{W}_f(\boldsymbol{\beta}_v^m \| \boldsymbol{\beta}_s^m) + \mathbf{b}_f\right) \tag{13}$$

$$\mathbf{t} = (t^s, t^e) = f\left(\mathbf{W}_{af}\mathbf{f} + \mathbf{b}_{af}\right) \tag{14}$$

Here $\mathbf{W}_f \in R^{h\times 2h}$ and $\mathbf{b}_f \in R^h$ are used for feature fusion. $\mathbf{W}_{af} \in R^{2\times h}$ and $\mathbf{b}_{af} \in R^2$ are parameters of the enhanced features based regression.

In addition, the parameters are updated by loss function shown as (15)-(16):

$$L_{reg} = \sum_{i=1}^{K} [R\left(\widetilde{t}_i^s - t_i^s\right) + R\left(\widetilde{t}_i^e - t_i^e\right)] \tag{15}$$

$$L_{cal} = -\sum_{i=1}^{K} \frac{\sum_{j=1}^{M} m_{i,j} log\left(a_j^{V_i}\right)}{\sum_{j=1}^{M} m_{i,j}} \tag{16}$$

Here, L_{reg} is an attention regression loss, and L_{cal} is an attention calibration loss. R($\cdot$) is the form of smooth L1 function[5].

The overall loss of MMFA-CF consists of both the attention regression and the attention calibration loss:

$$L = \alpha L_{reg} + \beta L_{cal} \tag{17}$$

α and β are hyper parameters which control the weights between the two loss terms.

With the above overall loss term, MMFA-CF can be trained end to end from the feature encoding step to the coordinates prediction step.

3. EXPERIMENTS

Our experimental environment is under ubuntu 18.04 with GeForce RTX 2080 Ti, and MMFA-CF is conducted on the Tensorflow deep learning framework.

The videos in TACoS dataset are split into 256 clips on average. Additionally, the hidden state size h of both video and sentence Bi-directional LSTM is set as 256, dropout rate as 0.5. In the multi-modal co-attention interaction network, the hidden state size k = 256, $U_g, U_z \in R^{256\times 256}$, $u_a, b_a \in R^{256}$. Besides, MMFA-CF is trained with a mini-batch of 100 and learning rate of 0.0001.

3.1. Dataset

TACoS: TACoS dataset is derived from MPII Cooking Composite Activities dataset, which contains a set of video descriptions (in natural language) and timestamp-based alignment with the videos. There are 127 videos in the dataset, which are divided into training set, verification set and test set, including 75, 27 and 25 videos respectively. And the number of clip-sentence pairs in training, verification and test sets is 10,146, 4,589 and 4,083, respectively.

Table 1. Comparison of different methods on TACoS.

Methods	R@1, IoU@0.1	R@1, IoU@0.3	R@1, IoU@0.5	mIoU
CTRL[3]	23.5	17.4	12.9	12.1
ACRN[6]	24.2	19.0	**13.1**	12.6
MMFA-CF	**31.6**	**19.1**	9.1	**12.9**

3.2. Evaluation Metrics

The standard evaluation metrics "R@1,IoU@σ" and "mIoU" [4] are used for evaluation. For each sentence query, the Intersection over Union (IoU) is calculated between the predicted and ground truth temporal coordinates. "R@1, IoU@σ" means the percentage of the sentence queries which have IoU larger than σ. Meanwhile, "mIoU" means the average IoU for all the sentence queries.

3.3. Experiment under TACoS dataset

In order to verify the effectiveness of MMFA-CF, it is reproduced under TACoS dataset. Besides, MMFA-CF is compared with the other two methods: CTRL and ACRN. The experimental results are shown in Table 1.

From the results, the following observations are obtained. Compared with CTRL, MMFA-CF achieves 8.1% improvement in R@1, IoU@0.1, 1.7% improvement in R@1, IoU@0.3, and 0.8% improvement in mIoU. Besides, compared with ACRN, MMFA-CF achieves 7.4%, 0.1%, 0.3% improvement in R@1, IoU@0.1, R@1, IoU@0.3, and mIoU respectively.

While when the IoU threshold increases to 0.5, MMFA-CF in R@1 is lower than the other two methods. It is speculated that the TACoS dataset has a single scenario, only the people and the cooking objects are changed, which makes the video clips indistinguishable. Therefore, small changes in each video clip lead to a relatively flatter wave in attention. It is can be seen in **Fig. 2**. Video attention weights are represented with the blue waves. Under this condition, MMFA-CF can effectively locate the approximate position of the sentence query, achieving better results of R@1 at lower IoU value, but can be difficult for determining the precise segment boundaries with the requirement of higher IoU threshold. As for CTRL and ACRN, the candidate video clips are split from the whole video and the sentence query is compared with each of these clips individually. Therefore, CTRL and ACRN can reduce the disturbance caused from similar scenes in TACoS videos.

In addition, since there are only 127 videos in the TACoS dataset, the small data size of the dataset limits the training accuracy and affects the performance of the algorithm.

Finally, words of high attention weights in sentence queries are highlighted by red font in **Fig. 2**. It can also be seen that the learned sentence attentions highlight some key words in sentence queries, such as some objects, actions and even words with time meaning. These highlighted words provide clear cues for localizing.

4. CONCLUSIONS

A multi-modal fusion algorithm for cross-modal video moment retrieval is proposed in this research. Firstly, Bi-directional LSTM is used to generate video clip and sentence representations incorporated with the contextual information. Then the multi-modal co-attention mechanism is introduced to set up a symmetry interaction between the video and the sentence query. The multi-modal co-attention mechanism not only learns the video attentions reflecting the global temporal structure, but also explores the crucial sentence details for localization. In the coordinates prediction network, different from previous methods which are using feature concatenation for multi-modal feature fusion, the consistent features separated from video features and sentence features are taken as the input, and then the temporal coordinates are obtained. The experimental results show the effectiveness of MMFA-CF as compared with CTRL and ACRN under TACoS dataset.

Since all the videos in TACoS share a common scene, the videos with more scenes will be considered for the experiments to extend the practical application capability of this task in future study.

References:

[1] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language", In Proceedings of the IEEE international conference on computer vision, pp. 5803–5812, 2017.

[2] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query", In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pp. 8199–8206, 2019.

[3] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query", In Proceedings of the IEEE international conference on computer vision, pp. 5267–5275, 2017.

[4] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization", In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 245–253. IEEE, 2019.

[5] R. Girshick, "Fast r-cnn", In Proceedings of the IEEE international conference on computer vision, pp. 1440–1448, 2015.

[6] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos", In The 41st international ACM SIGIR conference on research & development in information retrieval, pp. 15–24, 2018.

[7] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos", In Proceedings of the 26th ACM international conference on Multimedia, pp. 843–851, 2018.

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation", In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543, 2014.

[9] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities", In European conference on computer vision, pp. 144–157. Springer, 2012.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", In Proceedings of the IEEE international conference on computer vision, pp. 4489–4497, 2015.

[11] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression", In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pp. 9159–9166, 2019.

[12] S. Zhang, J. Su, and J. Luo, "Exploiting temporal relationships in video moment localization with natural language", In Proceedings of the 27th ACM International Conference on Multimedia, pp. 1230–1238, 2019.

[13] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos", In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 655–664, 2019.