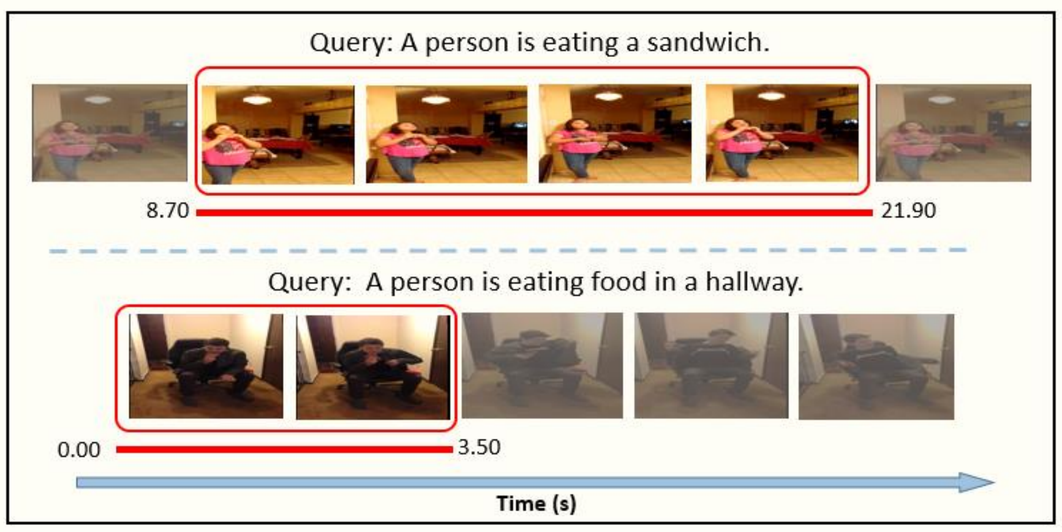
Multi-scale 2D Representation Learning for weakly-supervised moment retrieval

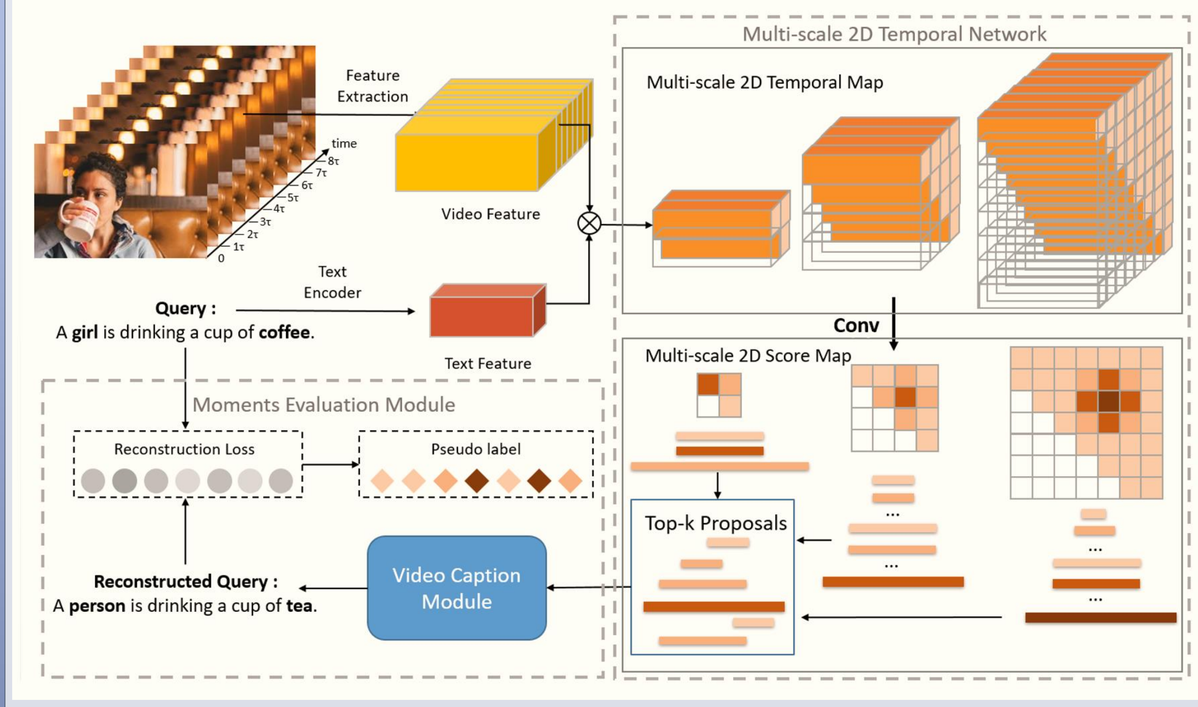
Ding Li, Rui Wu, Yongqiang Tang, Zhizhong Zhang, Wensheng Zhang
Institute of Automation, Chinese Academy of Sciences
Horizon Robotics, Beijing

Introduction



Motivation: First, most existing weakly-supervised methods resort to projecting the text features and video features of moment candidates into some learned unified space, and then calculate the alignment score of candidates with text query, the larger score indicates the higher probability to be the result. However, these methods process each moment candidate individually, thus the relations between video moments are inevitably neglected. Second, the present weakly-supervised moment retrieval methods generally overlook the fact that the variance of temporal scale of video moments is also an important influence factor for moments localization.

Multi-scale 2D Representation Learning



Framework of Multi-scale Representation Learning for weakly-supervised moment retrieval

Multi-scale 2D Temporal Network: The Multi-scale 2D Temporal Network takes an the basic video and text representation as input, and outputs $N_s \times K$ segment proposals and corresponding alignment scores. N_s is the number of scales, and K is the number of selected segment proposals in each scale.

Multi-scale 2D temporal feature map: Multi-scale temporal sampling on the untrimmed video $V = \{v_i\}_{i=1}^{N_v}$. Specifically, we first segment it into small video clips. Each video clip consists of T frames. Then, we repeatedly sample the video clips with N_s intervals. After j-th sampling, we get N_j video clips, and the original video would be converted to $V = \{S_i^j\}_{i=1,j=1}^{N_j,N_s}$, where S_i^j is the sampled clips. Further more, we extract the deep 3D CNN feature of each clip as mentioned before, denoted as $\{f^S\}_{i=1,j=1}^{N_j,Ns}$. To get a more compact representation, we pass the extracted feature through a fully-connected layer with d_v output channels. For the i-th video segment sampled in j-th scale, the final 3D feature is $f^S \in R^{d^v}$, where d_v is the feature dimension. The video feature of moment candidate is:

$$F_{x,y}^j = \begin{cases} \max\text{pool}\left(f_x^S, f_{x+1}^S, \cdots, f_y^S\right), & if\ 0 < x \le y < N_j \\ 0^S, & else \end{cases}$$

The multi-scale 2D video feature could be defined as $F^M = \{F^j\}_{j=1}^{N_s}$.
By duplicating text embedding, multi-scale text feature is $F^T = \{f^T, \cdots, f^T\}_{N_s}$.
The final cross-modal multi-scale feature are fused by the Hadamard product, formulated as:

$$F_{cro} = \left\| \left(W^T \cdot F^T \cdot 1^T\right) \odot \left(W^M \cdot F^M\right) \right\|_F$$

Problem Definition & Basic Representations

Problem Definition: Given a video denoted as $V = \{v_i\}_{i=1}^{N_v}$ and a sentence $T = \{t_i\}_{i=1}^{L}$ as text query, we aim to automatically retrieve the most relevant video segment according to the query. N_v is the number of the frames of the video, and L is the length of the sentence. Specifically, we can get the best-matched moment $M = \{\tau_s, \tau_e\}$, where τ_s, τ_e are the indices of start and end frame respectively. Note that there is no need to have access to the temporal boundary annotations of video moments in the training time.

Basic Video and Text Representation: We first split the whole video into several video clips, then each video clip would be used as the input of a pretrained 3D CNN model. The text encoder includes the word embedding and LSTM network. We use GloVe word2vec model to extract the word embedding of each word in the input sentence.

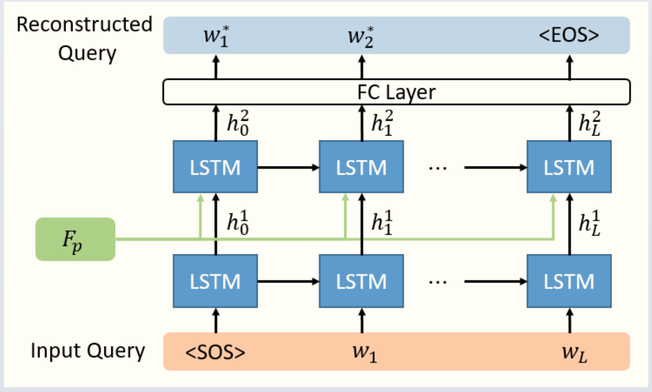
We utilize the multiple fixed intervals to sample frames from the original video, which facilitates the construction of the multiscale 2D map. The spatio-temporal feature are extracted by the pre-trained 3D CNN model, and then passed through a fully-connected layer with d_v output channels.

The final text feature are extracted as $f^T \in R^{L \times d^T}$, which encodes the input text query.

Moments Evaluation Module

Moments Caption:

Observing that the only annotation of this task is the text query, we design the caption module and make the whole model trainable.



The Caption Module for reconstruction of the text query

Moments Evaluation:

Owing to the lack of temporal boundary annotations, we are not able to compute the tIoU between moment candidates and truly-matched moments. Thus, we generate the pseudo labels for further training of the evaluation module.

Pseudo label generation:

$$l_k^j = \frac{\sum_{l=1}^{L} \log\left(w_l^* \left| F_{cro}^k, h_{l-1}^2, w_1, w_2, \cdots, w_{l-1}\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L} \log\left(w_l^* \left| F_{cro}^k, h_{l-1}^2, w_1, w_2, \cdots, w_{l-1}\right.\right)} \qquad y_k^j = \begin{cases} 0, & l_k^j \ge l_{\max} \\ 1 - l_k^j, & l_{\min} \le l_k^j < l_{\max} \\ 1, & l_k^j < l_{\min} \end{cases}$$

Loss functions:

Reconstruction Loss: $L_{rec} = \frac{-1}{N_s K L} \sum_{k=1}^{N_s K} \sum_{l=1}^{L} \log P\left(w_l^* \left| F_{cro}^k, h_{l-1}^2, w_1, \cdots, w_{l-1}\right.\right)$

Reconstruction-guided binary cross-entropy loss (RG-BCE loss):

$$L_{rg-bce} = \frac{1}{N_s K} \sum_{j=1}^{N_s} \sum_{k=1}^{K} y_k^j \log p_k^j + \left(1 - y_k^j\right) \log\left(1 - p_k^j\right)$$

Conclusion

In this work, we focus on the task of video moment retrieval without manually labelling the start and end time points of moments in training. We address the motivation of considering the various temporal scale of moment candidates as well as the temporal relations between, and propose a multi-scale 2D representation learning method, including multi-scale 2D temporal network and weakly-supervised moments evaluation module with RG-BCE loss. The multi-scale 2D temporal map could generate more precise moment candidates with various temporal scales, and moment-to-text reconstruction facilitate the weakly-supervised training in moments evaluation. Experiment results on the Charades-STA and ActivityNet Captions datasets demonstrated the effectiveness and superiority of our proposed approach.