Attentive Moment Retrieval in Videos

Meng Liu*
Shandong University
School of Computer Science and Technology
mengliu.sdu@gmail.com

Xiang Wang
National University of Singapore
School of Computing
xiangwang@u.nus.edu

Liqiang Nie
Shandong University
School of Computer Science and Technology
nieliqiang@gmail.com

Xiangnan He
National University of Singapore
School of Computing
xiangnanhe@gmail.com

Baoquan Chen
Shandong University
School of Computer Science and Technology
baoquan.chen@gmail.com

Tat-Seng Chua
National University of Singapore
School of Computing
dcscts@nus.edu.sg

ABSTRACT

In the past few years, language-based video retrieval has attracted a lot of attention. However, as a natural extension, localizing the specific video moments within a video given a description query is seldom explored. Although these two tasks look similar, the latter is more challenging due to two main reasons: 1) The former task only needs to judge whether the query occurs in a video and returns an entire video, but the latter is expected to judge which moment within a video matches the query and accurately returns the start and end points of the moment. Due to the fact that different moments in a video have varying durations and diverse spatial-temporal characteristics, uncovering the underlying moments is highly challenging. 2) As for the key component of relevance estimation, the former usually embeds a video and the query into a common space to compute the relevance score. However, the later task concerns moment localization where not only the features of a specific moment matter, but the context information of the moment also contributes a lot. For example, the query may contain temporal constraint words, such as “first”, therefore temporal context is required to properly comprehend them.

To address these issues, we develop an Attentive Cross-Modal Retrieval Network. In particular, we design a memory attention mechanism to emphasize the visual features mentioned in the query and simultaneously incorporate their context. In the light of this, we obtain an augmented moment representation. Meanwhile, a cross-modal fusion sub-network learns both the intra-modality and inter-modality dynamics, which can enhance the learning of moment-query representation. We evaluate our method on two datasets: DiDeMo and TACoS. Extensive experiments show the effectiveness of our model as compared to state-of-the-art methods.

CCS CONCEPTS

• **Information systems → Video search**; **Novelty in information retrieval**;

KEYWORDS

Temporal Memory Attention, Cross-modal Retrieval, Moment Localization, Tensor Fusion

ACM Reference Format:
Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *SIGIR ’18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210003

1 INTRODUCTION

Searching videos of interests from large collections has long been an open problem in the field of multimedia information retrieval [36]. Since this task needs to answer queries by relevant videos only, most prior efforts cast it as a matching problem [33] by estimating the relevance score between a video and the given query. Such direct video-query matching works well for judging whether the description query occurs in an entire video that depicts simple scenes solely. However, in some real-world scenarios (e.g., robotic navigation, autonomous driving, and surveillance), the untrimmed videos usually contain complex scenes and involve a large number of objects, attributes, actions, and interactions, whereby only some parts of the complex scene convey the desired cues or match the description. For a prepared surveillance video lasting for several minutes, as Figure 1 shows, one may only have interest in the moment, “a girl in orange first walks by the camera”, where the start and end points are at the 24s and the 30s, respectively. Therefore, localizing temporal moments of interest within a video is more useful yet challenging, as compared to simply retrieving an entire video.

In this paper, we focus on the task of moment retrieval, aiming to identify the specific start and end points within a video to precisely respond to the given query. In our work, a desired moment refers to a query-aware temporal segment whose content is in accordance with the given query[1]. In general, automatic moment

[*] Meng Liu is a visiting student of the National University of Singapore, supervised by Dr. Xiang Wang, Dr. Xiangnan He, and Prof. Tat-Seng Chua.

[1] Note that an entire video may contain multiple moments related to the given query.

SIGIR ’18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210003

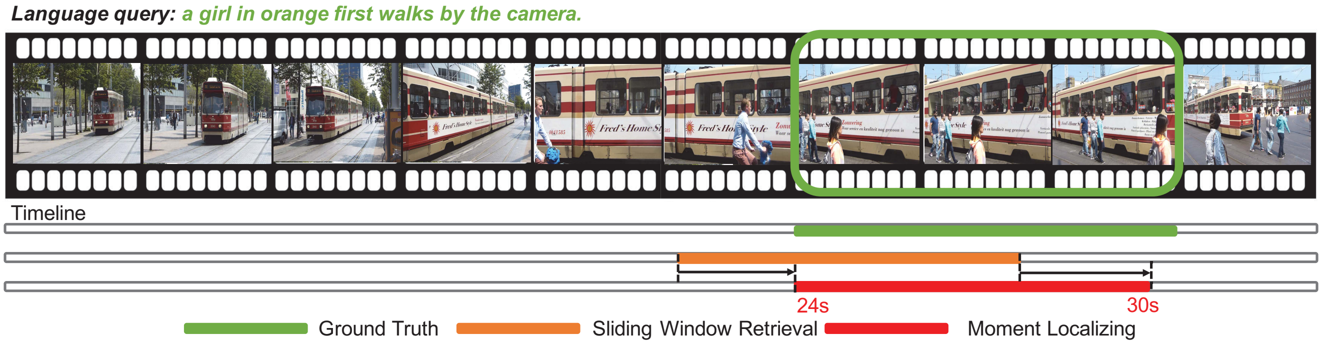


Figure 1: Temporal video moment localization is designed to localize a moment (the red bar) with a start point (24th s) and an end point (30th s) in the video according to the given language query. Here the green bar denotes the ground truth, the orange bar stands for the result of sliding window moment retrieval, and the red bar refers to the localizing result.

retrieval from a video requires two components, namely, fine-grained moment candidates localization and relevance estimation. The key challenges are, first, different moments in a video have varying durations and diverse spatial-temporal characteristics; thereby uncovering the underlying moments is already highly challenging, not to mention the estimation of moment-query relevance. To generate the moment candidates, a direct way is to densely sample sliding windows at different scales. However, such moment derivation methods are limited, not only for the expensive computational costs, but also the exponential search space. Second, the relevance estimation is a typical cross-modal retrieval problem. A viable solution as employed in [2] is to first project the visual features of the moment candidates and textual features of the query into a common latent space and then calculate the relevance based on their similarity. Nevertheless, such workflow overlooks the spatial-temporal information inside the moment and the query. Taking the query of "a girl in orange first walks by the camera" as an example, the term "first" is relative and requires temporal context for proper comprehension.

To address the aforementioned problems, we develop an Attentive Cross-Modal Retrieval Network, dubbed as ACRN, for the task of moment retrieval. For moment derivation, we propose a temporal memory attention network to explore the attentive contextual visual features of the moments. For each pre-segmented moment, its surrounding context, consisting of pre- and post-moments, encodes consistent signal to imply the continuous scenes [10]. Inspired by this, we utilize a memory network to memorize the contextual information for each moment, and treat the natural language query as the input to an attention network to adaptively assign weights to the memory representation. In the light of this, we obtain the augmented moment representation. Thereafter, we introduce a cross-modal fusion network to enhance the moment-query representation. It is built on the inter- and intra-modal embedding interactions. The former aims to explicitly model the interactions between the visual and textual embeddings, and the latter targets at exploring the embedding interactions within each individual modality. Finally, we feed the moment-query representation into a boundary regression model to predict the relevance scores and moment offsets.

The key contributions of this work are three-fold:

- We present a novel Attentive Cross-Modal Retrieval Network, which jointly characterizes the attentive contextual visual feature and the cross-modal feature representation. To the best of our knowledge, the existing studies either consider only one of the above models or not integrate them within a unified model.
- For the purpose of accurately localizing moments in a video with natural language, we are the first to introduce a temporal memory attention network to memorize the contextual information for each moment, and treat the natural language query as the input of an attention network to adaptively assign weights to the memory representation.
- We perform extensive experiments on two benchmark datasets to demonstrate the performance improvement. As a side contribution, we released the data and codes[2].

The rest of the paper is organized as follows. The related work is briefly introduced in Section 2. Section 3 details the proposed approach. We present experiment results in Section 4. Finally, Section 5 concludes the work and points out the future directions.

2 RELATED WORK

Localizing specific moments within a video responding to a textual query is related to many vision tasks including video retrieval, temporal action localization, as well as video description and question answering.

2.1 Video Retrieval

Given a set of video candidates and a language query, video retrieval algorithms aim to retrieve the videos that match the query. Technically, the retrieval problem is usually tackled as a ranking task [5–7], returning moments based on their matching scores. Similar to image-language embedding models [8, 31], current methods [22, 44] are designed to incorporate deep video-language embeddings. Lin et al. [18] proposed a retrieval model to match the visual concepts in the videos with the semantic graphs generated by parsing the sentence descriptions. Bojanowski et al. [1] introduced a strategy to tackle the problem of video-text alignment by assigning a temporal interval to each sentence given a video and a set of

[2] https://sigir2018.wixsite.com/acrn.

sentences with the temporal ordering. Different from the discussed algorithms, the input of our model is only one sentence query and the temporal ordering is not used.

There are also some efforts dedicated to retrieving temporal segments within a video in constrained settings. Tellex et al. [34] considered retrieving video segments from a home surveillance camera via text queries with a fixed set of spatial prepositions. Later, Lin et al. [19] developed a model to retrieve temporal segments in 21 videos from a dashboard car camera. Recently, Hendricks et al.[2] proposed a joint video-language model to retrieve moments within a video based on texture queries. However, these models can only verify the segments containing the corresponding moment. Namely, there are many background noises in the returned results. Although they could densely sample video moments at different scales and utilize these models to retrieve the corresponding video moment, it is not only computationally expensive but also makes the matching task more challenging with the search space increasing. As we know that adjusting the temporal boundaries of proposals by learning regression parameters has succeeded in the object localization, as in [26]. In this paper, we adopted a similar strategy to predict the start and end time points of the desired video moment.

2.2 Temporal Action Localization

Gaidon et al. [9] introduced the problem of temporally localizing actions in the untrimmed videos, focusing on limited actions such as “drinking and smoking” and “open the door and sit down”. Later, researchers worked on building large-scale datasets consisting of complex action categories, and proposed different models for localizing activities in videos. Shou et al. [28] proposed an end-to-end segment-based 3D Convolutional Neural Network (CNN) framework, which outperforms other Recurrent Neural Network (RNN)-based methods by capturing spatio-temporal information simultaneously. And Singh et al. [30] presented a multi-stream bi-directional RNN network for fine-grained action detection. Gao et al. [11] proposed a novel temporal unit regression network model, which can jointly predict action proposals and refine the temporal boundaries by temporal coordinate regression. Due to the fact that these methods are restricted to a pre-defined list of actions, Gao et al. [10] proposed to use natural language queries to localize activities. They leveraged all the context moments surrounding the current input, without explicitly considering the semantic information of the input query. It thus considers the video moments unrelated to the input query, which is unnecessary or even misleading.

2.3 Video Description and Question Answering

More recently, attention mechanism [43] is a standard part of the deep learning toolkit, contributing to the impressive results in neural machine translation [21], video captioning [23, 41] and video question answering [45]. Visual attention models for video captioning leverage the video frames at every time step, without explicitly considering the semantic attributes of the predicted words. It is unnecessary or even misleading. To tackle this issue, Song et al. [32] proposed a hierarchical Long Short-term Memory (LSTM) network [20] with an adjusted temporal attention model for video captioning. Later, Hori et al.[14] expanded the attention model to selectively attend not just to specific times or spatial regions, but to specific modalities of the inputs such as image features, motion features, and audio features. Their new modality-dependent attention mechanism provides a natural way to fuse multimodal information for video description. Recently, Xu et al. [42] proposed a multimodal attention LSTM network, which fully exploits both multimodal streams and temporal attention to selectively focus on specific elements during the sentence generation.

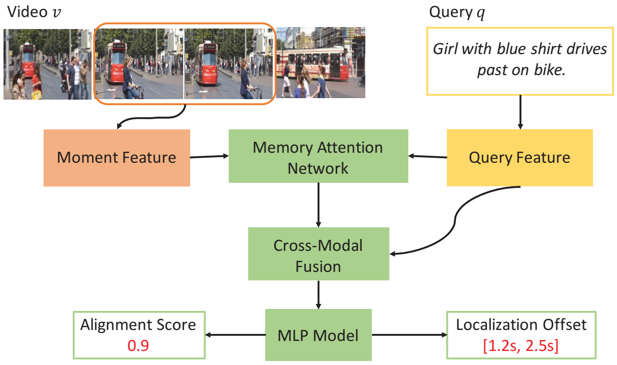


Figure 2: An illustration of our proposed ACRN model.

Video question answering is a relatively new task, where a video and a natural language question are provided and a model is designed to reply the question. Zhao et al. [48] developed a hierarchical dual-level attention networks to learn the question-aware video representations with word-level and question-level attention mechanisms. And they also proposed a hierarchical spatio-temporal attention network [49] to learn the joint representation of the dynamic video contents according to the given question. Unlike the aforementioned studies, Ye et al. [46] studied the problem of video question answering by modeling its temporal dynamics with frame-level attention mechanism. And Xu et al. [40] proposed to refine the attention by gradually using both coarse-grained question feature and fine-grained word feature. Motivated by these attention mechanisms, we presented a temporal memory attention model to dynamically select context moments consistent with the input query and simultaneously memorize the context moment information.

3 OUR PROPOSED ACRN MODEL

As Figure 2 illustrates, our proposed ACRN model comprises of the following components: 1) the memory attention network leverages the weighting contexts to enhance the visual embedding of each moment; 2) the cross-modal fusion network explores the intra-modal and the inter-modal feature interactions to generate the moment-query representations; and 3) the regression network estimates the relevance scores and predicts the location offsets of the golden moments.

3.1 Problem Formulation

Let v and q denote a video and a query, respectively. We present a video as a sequence of frames $v = \{f_t\}$, where f represents a frame and $t \in \{0, \cdots, \tau\}$ indexes the time point. The query is affiliated with a temporal annotation (t_s, t_e), where t_s and t_e is the start and end point[3], respectively. Our task is to identify a golden moment $c = \{f_t\}_{t=\tau_s}^{\tau_t}$ corresponding to the description of the given query q, whereby $(\tau_s, \tau_e) = (t_s, t_e)$. Towards this end, we pre-segment the video v into a set of moment candidates $C = \{c_i\}_{i=1}^{M}$ via multi-scale temporal sliding windows, where M is number of the moments[4]. For the given query q and a moment candidate c overlapping with the golden moment, we align the moment-query pair as a positive training sample. Moreover, due to the segmentation strategy, the positive candidates overlap with the golden moment on different scales, we hence pair each positive moment-query pair (c, q) with a time location offset (i.e., $(t_s - \tau_s, t_e - \tau_e)$). We will detail the process of data construction in Section 4.1. As such, the moment retrieval problem can be formally defined as:

Input: A set of moment candidates C and the given query q.

Output: A ranking model mapping each moment-query pair (c, q) to a relevance score and estimating their location offsets of the golden moment.

3.2 Memory Attention Network

To estimate the matching score between each moment candidate and the sophisticated query, a direct way is to project the visual embeddings of the moment candidates and the textual embedding of the query into a latent common space, and then feed them into a well-designed similarity function to calculate their relevance. Finally, it returns the moment with the highest score as the retrieval result. Formally, we summarize the above process as follows,

$$\begin{cases} \widehat{\mathbf{x}}_c = f_\Theta(\mathbf{x}_c), \\ \widehat{\mathbf{q}} = f_\Theta(\mathbf{q}), \\ c^* = \arg\max_{c \in C} g(\widehat{\mathbf{x}}_c, \widehat{\mathbf{q}}), \end{cases} \tag{1}$$

where $\mathbf{x}_c \in \mathbb{R}^{D_1}$ and $\mathbf{q} \in \mathbb{R}^{D_2}$ denote the embeddings of the moment c and the input query q, $f_\Theta(\cdot)$ is the mapping function [47] to project $\mathbf{x}_c$ and $\mathbf{q}$ to $\widehat{\mathbf{x}}_c$ and $\widehat{\mathbf{q}}$ in a common space, and g is the similarity function.

Although feasible, solely considering the current moment candidate overlooks the spatial-temporal information within its surrounding context, leading to information loss and suboptimal performance. For example, the term "first" in the query of "a girl in orange first walked by the camera" is a temporal constraint word and requires temporal context for a proper understanding. Recent work [10] has observed that the pre-context and post-context moments can be regarded as the context of the current moment and provide its relative temporal position in a video. Inspired by the observations, we consider to leverage the context information to complement the current moment.

Suppose the context moments of each video moment $c \in C$ are $\mathcal{N}_c = \{c_j\}$, where $j \in [-n_c, n_c]$ and n_c denote the shift boundary[5]. We utilize $j > 0$, $j = 0$, and $j < 0$ to index the post-, current, and pre-context moments, respectively. The embedding of the central moment c is denoted as $\mathbf{x}_c$, and its context embeddings are denoted as $\mathbf{x}_{c_j}$. Given these contextual embeddings, how to integrate them is crucial to strengthen the representation discrimination of the current moment. A simple strategy adopted in [11] is to employ the average pooling on the context embeddings to capture the interactions between the current and context moments. And the output of the pooling operator is used as the enhanced representation of the current moment, formulated as,

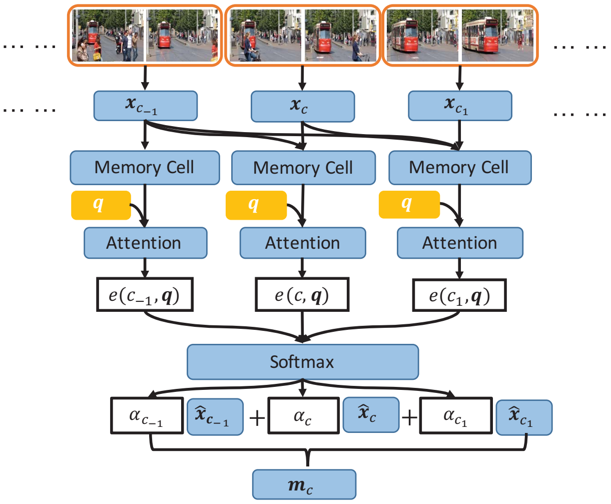


Figure 3: An illustration of our proposed memory attention model.

$$\widehat{\mathbf{x}}_c = \frac{1}{|\mathcal{N}_c|} \sum_{c_j \in \mathcal{N}_c} \mathbf{x}_{c_j}. \tag{2}$$

Although such average pooling is capable of fusing all contextual embeddings into a single one, we argue that it is insufficient to capture the consistent information and complex interactions among the moment contexts. Particularly, the average pooling assumes that the moments are linearly independent and equally contribute to the final relevance estimation. Thereby, it fails to identify the importance of each moment, and is unable to eliminate the useless even noisy features.

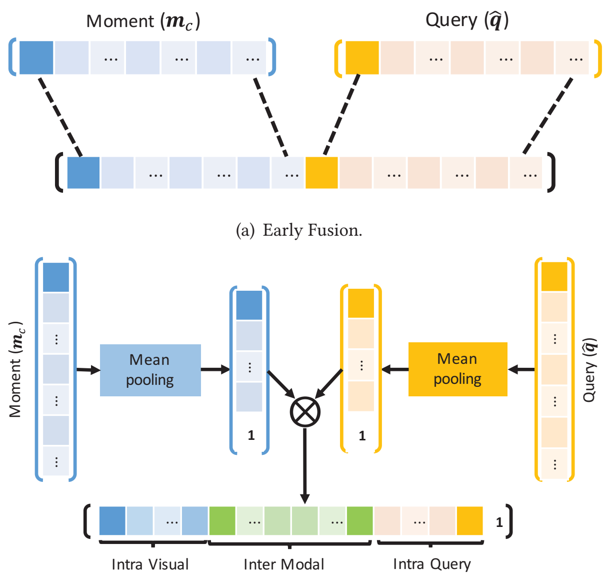
To tackle the aforementioned problem, we consider to explicitly capture the varying importance of each context moments by assigning an attentive weight to the embedding of each moment [37]. The detail of our memory attention is illustrated in Figure 3. Here we design a memory attention network by considering two components contributing to the attentive weights. Given the representation vector of the basic context moment $\mathbf{x}_{c_j} \in \mathbb{R}^{D_1}$ and the one of the given query $\mathbf{q} \in \mathbb{R}^{D_2}$, we use a one-layer network to estimate the attention score α_{c_j}, which explicitly reflects the consistency between the moment and the query. Moreover, for each moment in the contexts, we add the representations of the prior moments to memorize the temporal information and model the importance weights better. Formally, we present the memory attention network as follows,

$$\begin{cases} e(c_j, \mathbf{q}) = \sigma(\sum_{i=-n_c}^{j} \mathbf{W}_c \mathbf{x}_{c_i} + \mathbf{b}_c)^T \cdot \sigma(\mathbf{W}_q \mathbf{q} + \mathbf{b}_q), \\ \alpha_{c_j} = \frac{e(c_j, \mathbf{q})}{\sum_{k=-n_c}^{n_c} e(c_k, \mathbf{q})}, \; j \in [-n_c, n_c], \end{cases} \tag{3}$$

[3] As mentioned before, a description query may correspond to multiple moments in a given video. To simplify the notation, we only formulate one relevant moment.

[4] The generation of the moment candidates, the visual embedding of each moment, and the textual embedding of the given query will be described in Section 4.1.

[5] For example, n_c=1 denotes that the context moment number is 1. Namely, there is one pre-context and one post-context moment. The generation of the context moments is illustrated in Section 4.1.


Moment ($\boldsymbol{m}_c$)
Query ($\widehat{\boldsymbol{q}}$)
...
Moment ($\boldsymbol{m}_c$)
Mean pooling
Mean pooling
Query ($\widehat{\boldsymbol{q}}$)
1
1
1
Intra Visual
Inter Modal
Intra Query

(a) Early Fusion.

(b) Cross-modal Fusion.

Figure 4: An illustration of the commonly used early fusion and our proposed cross-modal feature fusion model. Top: Early fusion (multimodal concatenation). Bottom: Our proposed cross-modal feature fusion model with intra-modal and inter-modal intersections.

where $\mathbf{W}_c \in \mathbb{R}^{D\times D_1}$ and $\mathbf{W}_q \in \mathbb{R}^{D\times D_2}$ are used to transform the query and video embeddings to the same underlying embedding space; $\mathbf{b}_c \in \mathbb{R}^D$ and $\mathbf{b}_q \in \mathbb{R}^D$ denote the trainable bias vector for the moment and the query, respectively; and $\sigma(\cdot)$ is the tanh activation function to restrict the attention weight to be in (0,1).

With the attention weight α_{c_j}, the fused feature $\mathbf{m}_c$ is computed as follows,

$$\begin{cases}\widehat{\mathbf{x}}_{c_j} = \mathbf{W}_c\mathbf{x}_{c_j} + \mathbf{b}_c,\\ \mathbf{m}_c = \sum_{j\in[-n_c,n_c]}\alpha_{c_j}\widehat{\mathbf{x}}_{c_j},\end{cases} \tag{4}$$

where $\mathbf{m}_c \in \mathbb{R}^D$ is the representation of the current moment c attended by the current input query, and $\mathbf{W}_c$ and $\mathbf{b}_c$ are the common space embedding matrix and bias vector in Eqn.(3), respectively. As such, our memory attention network can leverage the context weights of the various importance of each moment to enhance the moment representations. And we can obtain the embedding query feature,

$$\widehat{\mathbf{q}} = \mathbf{W}_q\mathbf{q} + \mathbf{b}_q, \tag{5}$$

where $\mathbf{W}_q$ and $\mathbf{b}_q$ respectively denote the query embedding matrix and bias vector in Eqn.(3).

3.3 Cross-Modal Fusion Network

Previous multimodal studies do not leverage both intra-modality and inter-modality dynamics directly. Instead, they apply the commonly-used feature concatenation as an approach for multimodal feature fusion (as shown in Figure 4(a)). This fusion approach, nevertheless, does not efficiently model the inter-modality dynamics.

In this paper, we aim to build a fusion sub-network that disentangles unimodal and bimodal dynamics by modeling each of them explicitly. Having established the attentive embedding, we then obtain an enhanced moment representation. To estimate the relevance between the moment and the query, we design a cross-modal fusion network to explore the intra-modal and inter-modal embedding interactions. The former is implemented by the tensor fusion operation to explicitly model the interactions between the visual and textual embeddings. Meanwhile, the latter, implemented by the concatenation operation, targets at retaining the information within each individual modality. Thereafter, we concatenate these intra-modal and inter-modal embeddings into a fused moment-query representation.

As shown in the Figure 4(b), the cross-modal fusion network consists of two parts: the mean pooling and the tensor fusion. Due to the fact that high dimensional vectors will lead to expensive time complexity when computing tensor fusion, we introduce a mean pooling layer before conducting the tensor fusion. In particular, assuming that we obtain a D-dimension moment embedding $\mathbf{m}_c$ and a D-dimension query embedding $\widehat{\mathbf{q}}$ from the preceding memory attention network. We aim to learn a dimension reduction and high-level representation based upon mean pooling. Representation learning based on mean pooling is equivalent to applying a linear filter with the size n to each input embedding, and each entry in the output is the mean of the corresponding size kernel window in value. We employ the mean pooling layer on $\mathbf{m}_c$ and $\widehat{\mathbf{q}}$ to obtain the dimension reduction and high-level representation features $\widetilde{\mathbf{m}}_c$ and $\widetilde{\mathbf{q}}$ for the moment and the query, respectively. Hereafter, we input these two embeddings into the tensor fusion model. The tensor fusion, technically speaking, can be viewed as a differentiable outer product between the visual representation $\widetilde{\mathbf{m}}_c$ and the query representation $\widetilde{\mathbf{q}}$,

$$\mathbf{f}_{cq} = \begin{bmatrix}\widetilde{\mathbf{m}}_c\\ 1\end{bmatrix} \otimes \begin{bmatrix}\widetilde{\mathbf{q}}\\ 1\end{bmatrix} = [\widetilde{\mathbf{m}}_c, \widetilde{\mathbf{m}}_c \otimes \widetilde{\mathbf{q}}, \widetilde{\mathbf{q}}, 1], \tag{6}$$

where $\otimes$ indicates the outer product between vectors, and $\mathbf{f}_{cq}$ is all the possible combinations of the unimodal embeddings with three semantically distinct subregions. The two subregions $\widetilde{\mathbf{m}}_c$ and $\widetilde{\mathbf{q}}$ form unimodal interactions in tensor fusion, and subregions $\widetilde{\mathbf{m}}_c \otimes \widetilde{\mathbf{q}}$ capture bimodal interactions in tensor fusion.

3.4 Learning

Above the tensor fusion sub-network $\mathbf{f}_{cq}$, we place a multi-layer perceptrons (MLP) [12, 13, 38] to get the matching score of the moment-query pair (c, q) as well as the localization offset between the moment candidate and the golden moment $(t_s - \tau_s, t_e - \tau_e)$. Formally, the hidden layers are defined as follows[6],

$$\begin{cases}\mathbf{e}_1 = \theta_1(\mathbf{W}_1\mathbf{f}_{cq} + \mathbf{b}_1),\\ \mathbf{e}_2 = \theta_2(\mathbf{W}_2\mathbf{e}_1 + \mathbf{b}_2),\\ \cdots\\ \mathbf{e}_L = \theta_L(\mathbf{W}_L\mathbf{e}_{L-1} + \mathbf{b}_L),\end{cases} \tag{7}$$

where $\mathbf{W}_l$, $\mathbf{b}_l$, θ_l and $\mathbf{e}_l$ denote the weight matrix, bias vector, activation function, and output vector of the l-th hidden layers, respectively. As for the activation function in each hidden layer,

[6] In our experiments, the number of layers in MLP is set as two.

we opt for the ReLU unit. Particularly, the out vector $\mathbf{e}_L = [s_{cq}, \delta_s, \delta_e] \in \mathbb{R}^3$ comprises of the matching score s_{cq} and the localization offsets of $\delta_s = t_s - \tau_s$ and $\delta_e = t_e - \tau_e$.

Therefore, the loss function of our proposed model consists of two parts: one is utilized to compute the loss of the alignment scores, and the other is on the localization offsets. In the following subsections, we will detail them one by one.

3.4.1 Alignment Loss. Similar to the spirit in [10], we adopt the alignment loss to encourage the aligned moment-query pairs to have positive scores and misaligned pairs to have negative scores. Formally, we restate it as,

$$\begin{aligned} L_{align} = \alpha_1 & \sum_{(c,q)\in\mathcal{P}} \log(1 + \exp(-s_{cq})) \\ +\alpha_2 & \sum_{(c,q)\in\mathcal{N}} \log(1 + \exp(s_{cq})), \end{aligned} \tag{8}$$

where $\mathcal{P}$ is the set of positive moment-query pairs, namely aligned moment-query pairs; $\mathcal{N}$ is the set of negative moment-query pairs, namely misaligned moment-query pairs; and α_1 and α_2 are the hyper parameters controlling the weights between the positive and the negative moment-query pairs.

3.4.2 Localization Regression Loss. As the multi-scale temporal sliding window is adopted to segment videos, different moment candidates have different durations. Hence for each moment-query pair, we need to not only judge whether the moment is relevant to the query, but also decide the localization offsets compared to the golden moment. Here we adopt the moment boundary adjustment strategy presented in [11]. Formally, we denote the offset values for the start and end points as follows,

$$\begin{cases} \delta_s^* = t_s - \tau_s, \\ \delta_e^* = t_e - \tau_e, \end{cases} \tag{9}$$

where (t_s, t_e) is the start and end points of the given query, and (τ_s, τ_e) is the start and end points of a candidate moment in $\mathcal{P}$. Meanwhile, we use $\delta^* = [\delta_s^*, \delta_e^*]$ to denote the ground truth localization offsets.

Based on the ground truth offsets, we can adaptively adjust the alignment points of the current moments to match the exact temporal duration. Towards this end, we design a location offset regression modal as,

$$L_{loc} = \sum_{(c,q)\in\mathcal{P}} [R(\delta_s^* - \delta_s) + R(\delta_e^* - \delta_e)], \tag{10}$$

where $\mathcal{P}$ is the set of positive moment-query pairs and R is the L_1 norm function.

We devise the optimization framework consisting of the alignment loss and the localization regression loss processes, as,

$$L = L_{align} + \lambda L_{loc}, \tag{11}$$

where λ is a hyper-parameter to balance the two losses.

4 EXPERIMENT

4.1 Data Description

4.1.1 TACoS. The first dataset is constructed by [25]. It is built on the top of MPII-Compositive dataset [27] and contains 127 videos. Each video is associated with two type of annotations. One is the fine-grained activity label with temporal annotation (i.e., the start and end points). The other is natural language descriptions for the temporal annotations. The dataset[7] is used in [10] for temporal activity localization, dubbed as TACoS.

Table 1: The summary of the TACoS and DiDeMo datasets.

Dataset	# Videos	# Queries	# Moments	Domain	Video Source
TACoS	100	14,229	2,326	Cooking	Lab Kitchen
DiDeMo	10,464	40,543	26,892	Open	Flickr

We briefly describe the dataset construction process. In paper [10], each training video is sampled by multi-scale temporal sliding windows with size of [64, 128, 256, 512] frames and 80% overlap. As for the testing samples, they are coarsely sampled using sliding windows with size of [128, 256] frames. For a sliding window moment c from C with temporal annotation (τ_s, τ_e) and a query description q with temporal annotation (t_s, t_e), they are aligned as a pair of training sample if they satisfy the following conditions: 1) the Intersection over Union (IoU) is larger than 0.5; 2) the non Intersection over Length (nIoL) is smaller than 0.15; and 3) one sliding window moment can be aligned with only one query description. In the dataset, there are 75 training videos, 25 testing videos, and 26,963 training moment-query pairs satisfying the above conditions. Besides, they utilized 3D ConvNets (C3D) [35] as the moment-level visual encoder and Skip-thoughts [17] as the query description embedding extractor. Therefore, the dimension of the visual embedding and the query description embedding are 4,096 and 4,800, respectively.

4.1.2 DiDeMo. The second dataset is constructed by [2] for language-based moment retrieval, named the Distinct Describable Moments (DiDeMo) dataset[8]. It includes 10,464 personal videos with duration of 25-30 seconds, 26,892 video moments, and 40,543 localized descriptions. Descriptions in DiDeMo refer to expressions, describing the specific moments in a video. What is more, the construction of the DiDeMo dataset contains a verification step to ensure that the descriptions align with a single moment within a video. In the dataset, each video is broken into six five-second moments and represented by a 6 × 4096 feature matrix, where each column represents a 4,096-d VGG [29] feature of one moment. For language features, they adopted 300 dimensional dense word embeddings Glove [24] to represent each word.

The statistics of the datasets are summarized in Table 1. The reported experimental results in this paper are based on datasets mentioned above[9]. Besides, we carried out experiments with the help of Tensorflow, selecting function AdamOptimizer as our optimizer. We trained it over a server equipped with 16 Tesla K80s.

4.2 Experimental Settings

4.2.1 Evaluation Protocols. To thoroughly measure our model and the baselines, we adopt "R@*n*, IoU=*m*" proposed by [15] as the evaluation metric. To be more specific, given a query, it is the percentage of top-*n* results having IoU larger than *m*. In the following, we use $R(n, m)$ to denote "R@*n*, IoU=*m*". This metric

[7] https://github.com/jiyanggao/TALL.

[8] https://github.com/LisaAnne/LocalizingMoments.

[9] In the following experiments, we set the context moment number n_c as 1. And the length of context window is set as 128 frames on the TACoS dataset and 5 seconds on the DiDeMo dataset.

Table 2: Performance comparison between our proposed model and the state-of-the-art baselines on TACoS. (p-value*: p-value over $R(1, 0.5)$)

Method	R@1 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.1	R@5 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.1	p-value*
MCN	1.25%	1.64%	3.11%	1.25%	2.03%	3.11%	3.62E-10
VSA-STV	8.84%	13.59%	17.58%	16.41%	26.40%	35.86%	2.16E-06
VSA-RNN	9.96%	16.16%	20.92%	18.32%	29.19%	40.66%	1.82E-05
TALL	12.46%	16.85%	21.69%	24.44%	33.38%	45.38%	5.71E-05
ACRN	**14.62%**	**19.52%**	**24.22%**	**24.88%**	**34.97%**	**47.42%**	-

itself is on the query level, so the overall performance is the average among all the queries,

$$R(n,m) = \frac{1}{N_q} \sum_{i=1}^{N_q} r(n,m,q_i), \tag{12}$$

where $r(n, m, q_i)$ is the recall [3] for a query q_i, N_q is the total number of queries, and $R(n, m)$ is the averaged overall performance.

4.2.2 Baselines. We compared our proposed ACRN with the following several state-of-the-art baselines to justify the effectiveness of our proposal:

- **TALL** [10]: This is a cross-modal temporal regression localizer that jointly captures the interaction between the query description and video moments, as well as outputs alignment scores and action boundary regression results for the moment candidates.
- **MCN** [2]: This method is designed for the moment-query retrieval task. It emphasizes the local and global moment features, aiming to strengthen the expressiveness ability.
- **VSA-RNN** [10]: This method is the variant of the Deep Visual-Semantic Alignment (DVSA) model [16]. It transforms the local visual feature and the texture feature encoded by the LSTM model into a common space, and then estimates the matching score of each moment candidate and the query (as formulated in Eqn.(1)).
- **VSA-STV** [10]: Instead of using RNN to extract the query description embedding, this work uses an off-the-shelf Skip-thoughts [17] sentence embedding extractor. A skip-thought vector is in the 4,800-dimensional space, and we linearly transformed it to 1,000 dimension. Visual encoder is the same with that of the VSA-RNN.

Note that VSA-RNN and VSA-STV are two baseline models in [10], but the source codes and the involved parameters are not released by the authors. We implemented these two models by our own, and tried our best to tune their parameters to achieve the optimal performance. In our paper, we represented each word with 500 dimensional dense word embeddings (specifically Glove [24]) when training the VSA-RNN. The size of the hidden state of LSTM is 1,024 and the output size is 1,000. Video moments are processed by a visual encoder and linearly transformed to 1,000 dimensional, which are used as the moment-level embeddings. Besides, cosine similarity is used to calculate the confidence score between the moment candidates and the given query. And hinge loss is used to train the two models, which is defined as follows,

$$L = \sum_k [\sum_l \max(0, s_{kl} - s_{kk} + 1) + \sum_l \max(0, s_{lk} - s_{kk} + 1)], \tag{13}$$

Table 3: Performance comparison between our proposed model and the state-of-the-art baselines on DiDeMo. (p-value*: p-value over $R(1, 0.5)$)

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@1 IoU=0.9	R@5 IoU=0.5	R@5 IoU=0.7	R@5 IoU=0.9	p-value*
MCN	23.33%	15.37%	15.32%	41.03%	20.37%	19.77%	6.14E-09
VSA-STV	25.38%	14.49%	14.39%	68.56%	26.92%	24.24%	1.98E-03
VSA-RNN	24.94%	14.52%	14.44%	68.39%	26.10%	23.95%	3.31E-06
TALL	26.45%	15.36%	15.31%	68.78%	28.43%	26.15%	2.32E-02
ACRN	**27.44%**	**16.65%**	**16.53%**	**69.43%**	**29.45%**	**26.82%**	-

where k is the index of moment candidates, l is the index of query descriptions, s_{kk} denotes the cosine score of the aligned moment-query pair, and s_{kl} or s_{lk} denotes the cosine score of the misaligned moment-query pair.

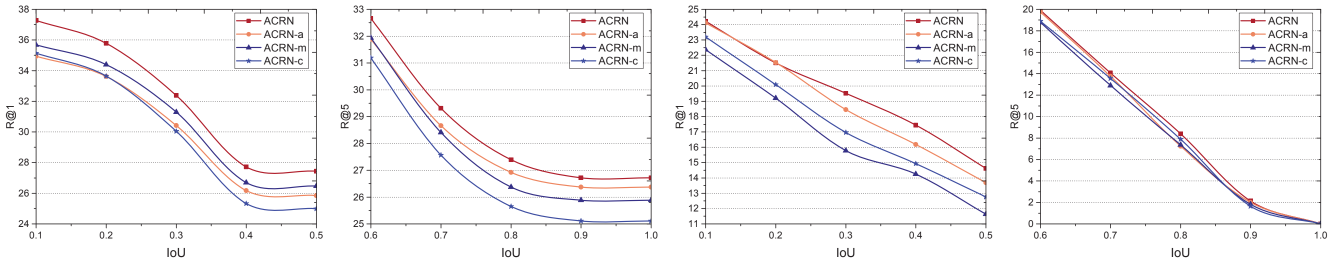
4.3 Performance Comparison

Table 2 displays the performance comparison w.r.t. $R(n, m)$ on TACoS. We have the following observations:

- VSA-STV and VSA-RNN achieve poor performance since they overlook the context information of moment candidates. They hence fail to exploit the spatial-temporal cues to guide the retrieval process, highlighting the necessity of modeling the context in moment retrieval.
- While MCN considers the features from the surrounding moments, it treats the average pooling of all the context representations as the context of each current candidate, ignoring the adaptive importance of the context moments. Assigning equal importance with each context moments may lead to introduce noisy features and lead to negative transfer. That is why MCN achieves the unstable performance on two datasets. It hence verifies the feasibility of revising the attention weight of each context moment.
- When performing our moment retrieval task, TALL outperforms MCN, VSA-STV and VSA-RNN. The observed results make sense since TALL is capable of exploiting the interactions across the visual and textual modalities and strengthens the expressiveness of the moment-query pairs.
- ACRN achieves the best performance, substantially surpassing all the baselines. Particularly, ACRN shows consistent improvements over TALL and MCN, verifying the importance of memorizing the context information and employing the attention mechanism on identifying the adaptive importance attention of each context moment.

We also evaluated our proposed ACRN model and the baseline methods on DiDeMo, and reported the results regarding IoU∈{0.5, 0.7,0.9} and R@{1, 5}. Note that since the positive moment-query pairs in this dataset are well aligned, namely there are no location offsets between them, we only used the alignment loss to train the ACRN and TALL model for localizing the corresponding moment. The results are shown in Table 3. It can be seen that the results are consistent with those on TACoS. ACRN shows a significant improvement over non-attention models (TALL and MCN) and non-context models (VSA-RNN and VSA-STV).

In addition, we also conducted the significance test between our model and each of the baselines. We can see that all the p-values are substantially smaller than 0.05, indicating that the advantage of our model is statistically significant.



(a) R@1 vs IoU on DiDeMo (b) R@5 vs IoU on DiDeMo (c) R@1 vs IoU on TACoS (d) R@5 vs IoU on TACoS

Figure 5: Performance comparison among our model variants on the TACoS dataset and the DiDeMo dataset. From left to right: (a) is the R@1 vs IoU $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ **on the DiDeMo dataset; (b) is the R@5 vs IoU** $\in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ **on the DiDeMo dataset; (c) is the R@1 vs IoU** $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ **on the TACoS dataset; (d) is the R@5 vs IoU** $\in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ **on the TACoS dataset.**

4.4 Study of ACRN

In the following section, we first explore how our proposed memory attention network and cross-modal fusion network affect the moment retrieval results. We then visualize the alignment and offset regression process of our proposed ACRN.

4.4.1 Component-Wise Comparison. We experimented with variants of our model to verify the effectiveness of the memory attention and cross-modal fusion networks:

- **ACRN-a**: We utilized the average pooling in Eqn.(2) to replace our proposed memory attention network for memorizing the context embeddings.
- **ACRN-m**: We eliminated the memory part of our memory attention model in Eqn.(3). That is, each context attention value is only related to itself and the query without considering the context information.
- **ACRN-c**: Instead of using cross-modal fusion model in Eqn.(6), we adopted the early fusion strategy, i.e., concatenating the multimodal feature.

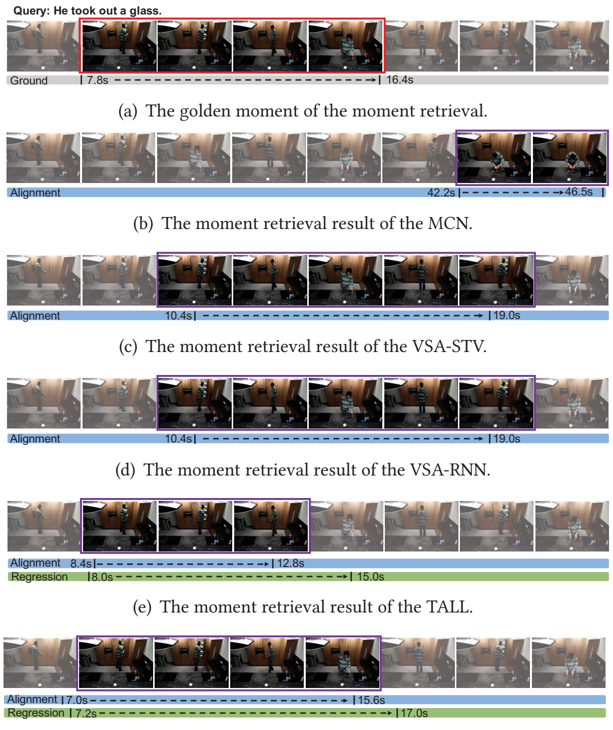
We tested these model variants on the TACoS and DiDeMo dataset, respectively. And the component-wise comparison results are shown in the Figure 5.

By jointly analyzing Figure 5, we have the following findings:

- As Figures 5(a) and 5(b) demonstrate, ACRN outperforms ACRN-a by a large margin on the DiDeMo dataset, and achieves considerable improvement on the TACoS dataset as shown in the Figures 5(c) and 5(d). It reveals that simply operating average pooling is insufficient to capture the consistent information and underlying interactions among the moment contexts. As average pooling assumes that the context moments are linearly independent and equally contributing to the final relevance estimation. It hence fails to identify the adaptive importance of each moment and hardly eliminates the irrelevant even noisy features. Therefore, the improvement achieved by ACRN verifies the effectiveness of the attention mechanism.
- The performance of ACRN-m indicates that removing the memory attention network hurts the expressiveness of the moment representation and further degrades the retrieval performance. Particularly, ACRN-m assumes that the representation of one moment candidate is independent with its surrounding context moments, which cannot exploit the spatial-temporal information encoded in the contexts. Taking the advantage of the memory attention network, ACRN is capable of enriching the moment representation.
- ACRN shows consistent improvement over ACRN-c on two datasets, verifying the crucial influence of modality interaction. Concatenation of the moment and query representations models the intra-modal interactions solely and limits the expressiveness of the moment-query pairs' representations. Our proposed cross-modal fusion network can exploit the intra-modal and inter-modal feature interactions and further enhance the moment-query representations.

4.4.2 Qualitative Results. To gain the deep insights into our proposed ACRN model, we show an example of moment retrieval. The video illustrated in Figure 6 describes a complex cooking scene, in which a man firstly took out a glass from the cupboard and placed it on the countertop, and then he went back to the cupboard and took out a second glass. Later, he cracked an egg from the fridge and drained the egg white by holding the halves of the shell together over the glass. We choose the description "He took out a glass" from the dataset as the given query, and utilize the aforementioned models to retrieve the relevant moments. From the results shown in the Figure 6, we observe that:

- As Figure 6(b) illustrates, MCN returns a moment that "The man drained the egg white" from the moment candidates, not the moment that "He took out a glass". Although it considers the local moment feature and the global feature, MCN forces all the background moment contexts as the global feature to enhance the representation of the visual embedding. As most of the moment candidates within this video are related to the scene "cracked egg and drained egg white", the global visual embedding fails to represent the desired scene.
- Both VSA-STV and VSA-RNN return a moment contain two sub-scenes which are "He took out a glass" and " He took a second glass", as shown in Figure 6(c) and 6(d), respectively. Because these two models only consider the current moment information instead of the temporal context information, they cannot identify the relative order of the moments. Hence, they only return all frames contain the action "took" and object "glass" as the output. The poor performance



(a) The golden moment of the moment retrieval.

(b) The moment retrieval result of the MCN.

(c) The moment retrieval result of the VSA-STV.

(d) The moment retrieval result of the VSA-RNN.

(e) The moment retrieval result of the TALL.

(f) The moment retrieval result of the ACRN.

Figure 6: Moment retrieval results on the TACoS dataset. All of the above figures are the R@1 results. The gray, blue and green bars denote the time line of the ground truth, alignment result and regression fine tune result, respectively.

admits the importance of the spatial-temporal information within the surrounding context.

- TALL generates more accurate alignment result than MCN, VSA-STV, and VSA-RNN, as Figure 6(e) displays. This certifies the importance of cross-modal fusion, which enriches the moment-query representations by modeling the intra-modal and inter-modal feature interactions.
- Our alignment retrieval performs better than all the state-of-the-art baselines. As shown in Figure 6(f), our alignment result has larger IoU with the golden moment. Moreover, ACRN generates better result than TALL. Although TALL model utilizes context information, it respectively pools the pre- and post-contexts into one vector and then concat them with the current moment to enhance the visual embedding. It ignores the complex interactions among contexts and fails to identify the importance of each moment, therefore it misses some important cues. This indicates the effectiveness of our proposed attentive moment retrieval network.
- Even in the case that the alignment retrieval results have small IoU with the golden moment, ACRN and TALL can correct the alignment time points via their regression part and further provide a more accurate result. This highlights the effectiveness of the location offset regression. In addition, our proposed ACRN further performs better offsets than TALL, this demonstrates that the ACRN can generate better moment-query representation.

5 CONCLUSION AND FUTURE WORK

In this paper, we develop an attentive cross-modal retrieval scheme to retrieve specific moments from a long video responding to a given query. To well align the moment candidates and the given query, we design a memory attention model to dynamically compute the visual attention over the query and its context information. Meanwhile, we adopt a cross-modal fusion sub-network to incorporate cross-modal information into the moment-query alignment. To evaluate our model, we perform extensive experiments on two public datasets. And the results show that our model can achieve better performance compared to the state-of-the-art baselines.

In future, we will extend our work in three directions. First, we plan to design an end-to-end model, which observes the moments and decides both where to look at next and when to make a prediction. It will not need to pre-segment videos with multi-scale sliding windows, and can quickly narrow down the searching space. Second, we shall study different attention networks on frame-level and incorporate them into our model, because different parts of a frame have varying influences on the scene and query understanding. Third, we will consider our framework in personalized moment recommendation, where the retrieval result is relevant to personal interests of users. In particular, when given a video, the personal query history is treated as the user-item interactions similar in [4, 13, 39] to better capture a user's preference towards moments.

6 ACKNOWLEDGMENTS

We would like to thank all reviewers for their comments. This work is supported by Joint NSFC-ISF Research Program (No.61561146397) funded by the National Natural Science Foundation of China and the Israel Science Foundation, the National Basic Research Program of China (973) (No.2015CB352501, No.2015CB352502), National Natural Science Foundation of China (No.61772310, No.61702300, and No.61702302) and the One Thousand Talents Plan of China (No.11150087963001). This work is part of NExT research, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised Learning from Narrated Instruction Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4575–4583.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing Moments in Video with Natural Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5803–5812.

[3] Da Cao, Xiangnan He, Liqiang Nie, Xiaochi Wei, Xia Hu, Shunxiang Wu, and Tat-Seng Chua. 2017. Cross-platform App Recommendation by Jointly Modeling Ratings and Texts. *ACM Transactions on Information Systems* 35, 4 (2017), 37.

[4] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding Factorization Models for Jointly Recommending Items and User Generated Lists. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 585–594.

[5] Zhiyong Cheng, Xuanchong Li, Jialie Shen, and Alexander G Hauptmann. 2016. Which Information Sources are More Effective and Reliable in Video

Search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1069–1072.

[6] Fuli Feng, Xiangnan He, Yiqun Liu, Liqiang Nie, and Tat-Seng Chua. 2018. Learning on Partial-order Hypergraphs. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1523–1532.

[7] Fuli Feng, Liqiang Nie, Xiang Wang, Richang Hong, and Tat-Seng Chua. 2017. Computational Social Indicators: A Case Study of Chinese University Ranking. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 455–464.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A Deep Visual-semantic Embedding Model. In *Proceedings of the Advances in Neural Information Processing Systems*. NIPS, 2121–2129.

[9] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. 2011. Actom Sequence Models for Efficient Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3201–3208.

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5267–5275.

[11] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. 2017. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3628–3636.

[12] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 355–364.

[13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 173–182.

[14] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based Multimodal Fusion for Video Description. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4203–4212.

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural Language Object Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4555–4564.

[16] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3128–3137.

[17] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought Vectors. In *Proceedings of the Advances in Neural Information Processing Systems*. NIPS, 3294–3302.

[18] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2657–2664.

[19] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2657–2664.

[20] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards Micro-video Understanding by Joint Sequential-Sparse Modeling. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 970–978.

[21] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1412–1421.

[22] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning Joint Representations of Videos and Sentences with Web Image Search. In *Proceedings of the European Conference on Computer Vision*. Springer, 651–667.

[23] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. Seeing Bot. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1341–1344.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1532–1543.

[25] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Proceedings of the Advances in Neural Information Processing Systems*. NIPS, 91–99.

[27] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script Data for Attribute-based Recognition of Composite Activities. In *Proceedings of the European Conference on Computer Vision*. Springer, 144–157.

[28] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1049–1058.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[30] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-grained Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1961–1970.

[31] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics* 2 (2014), 207–218.

[32] Jingkuan Song, Zhao Guo, Lianli Gao, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. 2017. Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning. *arXiv preprint arXiv:1706.01231* (2017).

[33] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural Compatibility Modeling for Clothing Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 753–761.

[34] Stefanie Tellex and Deb Roy. 2009. Towards Surveillance Video Search by Natural Language Query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 38.

[35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4489–4497.

[36] David Vallet, Frank Hopfgartner, Joemon M Jose, and Pablo Castells. 2011. Effects of Usage-based Feedback on Video Retrieval: a Simulation-based Study. *ACM Transactions on Information Systems* 29, 2 (2011), 11.

[37] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. TEM: Tree-enhanced Embedding Model for Explainable Recommendation. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1543–1552.

[38] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item Silk Road: Recommending Items from Information Domains to Social Users. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 185–194.

[39] Xiang Wang, Liqiang Nie, Xuemeng Song, Dongxiang Zhang, and Tat-Seng Chua. 2017. Unifying Virtual and Physical Worlds: Learning Toward Local and Global Consistency. *ACM Transactions on Information Systems* 36, 1 (2017), 4.

[40] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of the ACM Conference on Multimedia*. ACM.

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5288–5296.

[42] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. 2017. Learning Multimodal Attention LSTM Networks for Video Captioning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 537–545.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*. ACM, 2048–2057.

[44] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework.. In *Proceedings of the American Association for Artificial Intelligence*, Vol. 5. AAAI, 6.

[45] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked Attention Networks for Image Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 21–29.

[46] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video Question Answering via Attribute-Augmented Attention Network Learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 829–832.

[47] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5532–5540.

[48] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Dual-Level Attention Network Learning. In *Proceedings of the ACM Conference on Multimedia*. ACM, 1050–1058.

[49] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 3518–3524.