Interaction-Integrated Network for Natural Language Moment Localization

Ke Ning, Lingxi Xie, Jianzhuang Liu, *Senior Member, IEEE*, Fei Wu, *Senior Member, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*— Natural language moment localization aims at localizing video clips according to a natural language description. The key to this challenging task lies in modeling the relationship between verbal descriptions and visual contents. Existing approaches often sample a number of clips from the video, and individually determine how each of them is related to the query sentence. However, this strategy can fail dramatically, in particular when the query sentence refers to some visual elements that appear outside of, or even are distant from, the target clip. In this paper, we address this issue by designing an Interaction-Integrated Network (I^2N), which contains a few Interaction-Integrated Cells (I^2Cs). The idea lies in the observation that the query sentence not only provides a description to the video clip, but also contains semantic cues on the structure of the entire video. Based on this, I^2Cs go one step beyond modeling short-term contexts in the time domain by encoding long-term video content into every frame feature. By stacking a few I^2Cs, the obtained network, I^2N, enjoys an improved ability of inference, brought by both (I) multi-level correspondence between vision and language and (II) more accurate cross-modal alignment. When evaluated on a challenging video moment localization dataset named DiDeMo, I^2N outperforms the state-of-the-art approach by a clear margin of 1.98%. On other two challenging datasets, Charades-STA and TACoS, I^2N also reports competitive performance.**

***Index Terms*— Temporal action localization, cross-modal learning, vision-language understanding.**

I. Introduction

VIDEO analysis [1]–[5] is a fundamental task in computer vision and artificial intelligence [6], [7], and has attracted increasing attention in recent years. It is a big challenge towards intelligent systems [8]–[10]. With the rapid development of deep learning, state-of-the-art video analysis systems are often built upon deep neural networks [11]–[13], which construct hierarchical representation in both the spatial and temporal domains.

Manuscript received March 7, 2020; revised August 27, 2020 and October 19, 2020; accepted January 4, 2021. Date of publication January 22, 2021; date of current version February 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61625107 and Grant 61702130, in part by the Key Research and Development Projects of the Ministry of Science and Technology under Grant 2020YFC0832500, in part by the Guangxi Natural Science Foundation under Grant 2020GXNSFAA159137, and in part by the Major Scientific Research Project of Zhejiang Laboratory under Grant 2018EC0ZX01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julian Fierrez. *(Corresponding author: Fei Wu.)*

Ke Ning and Fei Wu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: wufei@zju.edu.cn).

Lingxi Xie, Jianzhuang Liu, and Qi Tian are with Huawei Inc., Shenzhen 518129, China.

Digital Object Identifier 10.1109/TIP.2021.3052086

This paper considers a specific task named natural language moment localization, *a.k.a.*, temporal activity localization [14] or localizing moments in videos with natural language [15]. The goal is to find a clip (*i.e.*, a continuous set of frames) from a given video which best fits the query provided in natural language. This task requires understanding the relationship between visual and linguistic contents, for which a complex system aware of cross-modal alignment needs to be built. Most existing approaches work by first sampling a number of clips from the video. Then, two network branches are constructed, with the first one extracting semantic contents from the query sentence, and the second one doing the same work from the visual data. A network head determines the extent (a score) that these two branches are delivering the same message, and the clip with the highest score is taken to be the output.

We point out two major drawbacks of this pipeline. First, the information fusion between the two modalities, *i.e.*, vision and language, is done at a very high level, which makes it difficult to learn multi-level visual-linguistic correspondence. Second and **more importantly**, as shown in Figure 1, there are many cases in which the critical verbal description, or keywords, cannot be found in the target video clip. Instead, these visual contents can sometimes be located far away from the target clip. For example, when the query is *"tent is not in these scenes"*, the keyword is *"tent"*, yet the object does not appear in the target clip. Another difficult query is *"second time we zoom in to the stage"*, for which the algorithm should have the ability to count the number of occurrences of a given action, which can happen a long time before the target clip.

In this paper, we present a novel network structure to address the above issues. We rethink the use of natural language, and argue that the query, besides delivering verbal description, also serves as a semantic prior of the video, *i.e.*, based on it, one can roughly imagine how the video is organized even without seeing any part of it. Therefore, we suggest to offer opportunities for the two branches, visual and linguistic, to cooperate at multiple levels. We design a module named Interaction-Integrated Cell (I^2C), which integrates both cross-modal and contextual interactions between vision and language, and thus goes one step beyond the original video hierarchy built with convolutional layers. I^2C can replace the original network layers at any stage. By stacking a few I^2Cs in a chain, we obtain an Interaction-Integrated Network (I^2N), in which visual and linguistic signals are interacting at multiple stages and, consequently, the network is equipped with more accurate cross-modal alignment. Note that each I^2C has the

Fig. 1. In the task of natural language moment localization, seeing only information in the target clips (marked in the green frames) is sometimes insufficient to make the correct decision. In the first example, the algorithm needs to know where the *"tent"* is and when it disappears from the video. In the second one, the algorithm needs to count how many times the *"zoom-in"* operation has been performed. Both cases are helped by a few "context" frames (marked in the blue frames), some of which can be quite distant from the targets (*e.g.*, in the second example).

ability to build distant connections, unlike a regular convolution that only captures local contextual information. Thus, the advantage of I^{2}N becomes more significant in such difficult cases (see Figure 4 for a few examples).

We present three contributions in this paper:

1. We propose an Interaction-Integrated Cell (I^{2}C) to deal with the problem of video semantics construction, taking both long-range and short-range video contents and the given natural language description into consideration.

2. We propose a framework, Interaction-Integrated Network (I^{2}N), to tackle the natural language moment localization task. With I^{2}C as the basic building block, I^{2}N explores more detailed vision-language relations on multiple granularities, therefore achieving more accurate semantics understanding and more accurate localization results.

3. We evaluate our approach on three popular benchmarks for natural language moment localization, namely, DiDeMo [15], Charades-STA [14], and TACoS [16]. In terms of Rank@1 accuracy, I^{2}N outperforms all prior work by a large margin. On the DiDeMo dataset, I^{2}N achieves Rank@1 hit rate of 32.58% with two-stream feature, which outperforms the prior best by a margin of 4.35%. And on the Charades-STA dataset, I^{2}N achieves Rank@1 hit rates of 41.69% and 22.88% under the criteria of IoU $\geqslant$ 0.5 and IoU $\geqslant$ 0.7 with C3D feature, which outperform the prior best by margins of 2.22% and 4.01%, respectively. On the TACoS dataset, our I^{2}N achieves Rank@1 hit rate of 29.25% under the criteria of IoU $\geqslant$ 0.5 with C3D feature, which outperform the prior best by 3.93%.

II. Related Work

A. Action Recognition and Detection

Action recognition and detection [17]–[20] are fundamental problems in video analysis. Two-stream ConvNets [21], [22] and 3D ConvNets [11], [23] are popular methods to encode raw video dynamics into video representation for action recognition. In most cases, raw videos are untrimmed. Temporally localizing meaningful clips from untrimmed videos is important for video understanding. Many approaches [12], [24]–[28] were proposed to tackle this problem recently, in which temporal convolution is used as basic building blocks to model temporal hierarchy in videos, and target clips are marked by simple predefined labels.

In this paper, we address a more challenging task which directly uses natural language descriptions to localize video clips. Compared to predefined labels, natural language descriptions are much more diverse, which makes them impossible to enumerate. Besides, a natural language description contains much more complex semantic structures than a simple label, such as causal relations [29] or external knowledge [30], [31]. That is to say, a natural language description not only plays the role of a mark of the target clip, but also helps the building of video semantics. Therefore, localizing video clips by natural language descriptions requires a more accurate alignment between vision and language contents. To tackle this issue, we integrate vision-language cross-modal interaction into our I^{2}C module, enabling linguistic information not only to match visual representation, but also to provide verbal prior for visual semantics.

B. Vision-Language Cross-Modal Learning

Modeling vision-language cross-modal interaction is a major challenge of our task. Visual information and natural language information come from different modalities. How to bridge the semantic gap between them still remains less explored. Visual Question Answering (VQA) is a pioneering research field along this direction [32]–[35]. Given an image and a question in natural language, the system is required to give the correct answer by integrating both visual and linguistic contents. To solve this problem, some recent work [36]–[39] utilized natural language inputs as a module to process visual data, thus cross-modal interactions can be modeled flexibly. It was also suggested [40] to fuse visual and linguistic information from early stages to help visual processing.

In this work, we focus on video data, which is more complex than image data. A video consists of a series of images, related to each other in the temporal dimension. It is widely believed that the temporal dimension has different properties from the spatial dimensions [11], [41], and therefore, simply adopting

3D convolutions could not accurately capture visual dynamics in a complex scene. In addition, collecting video data is much more expensive than collecting image data, which results in the current situation that much fewer video data are available for model training. Therefore, improving the efficiency in exploiting training data is also a major concern for video models. In our I^{2}N, we make use of temporal convolution to encode temporal information. Along with contextual interactions, the model can flexibly construct the temporal structure for videos.

C. Natural Language Moment Localization

Natural language moment localization is attracting increasing attentions. Hendricks *et al.* [15] and Gao *et al.* [14] proposed the task of localizing video clips by natural language descriptions, as well as the baseline models named MCN and CTRL for this task. Both MCN and CTRL encoded candidate clips and description sentences independently, followed by similarity measurement between two encoded vectors. Efforts were made to perform more detailed cross-modal interactions. Reference [42] performed cross-modal fusion by an LSTM on the temporal domain and discovers frame-to-word interaction, [43] performed a multi-modal co-attention interaction for the visual information and the language information. Reference [44] gathered local features via object detection on video frames. The model then localizes the final clip together with visual feature and linguistic feature. References [45] and [46] fused visual and linguistic information at earlier stages to model vision-language relations. These works explores word-to-frame attention, but without considering these relations on different levels.

Visual contexts, in particular global visual contexts, play an important role in video moment localization. To construct visual structures at a finer level as well as to bridge the semantic gap between vision and language, it is strongly required to model the semantic dependencies over long-ranged contexts. However, few existing approaches took this issue into consideration. Reference [45] imported global information with temporal attention for clip proposal, but without exploiting contextual semantics. Reference [46] used graph convolution to explore contextual information between clips. Reference [47] modeled video contextual information by focusing on temporal words such as "before", "after", "then" and "while". Some recent [48] explored different methods for temporal modeling, and achieved significantly performance improvement.

In this work, we integrate global contextual interaction into the proposed I^{2}C module. Together with building video hierarchy via convolution-based operations, this enables a more flexible way of feature extraction. In addition, semantic dependencies from the entire video sequence become available, leading to better potential of being aligned to natural language descriptions.

III. Our Approach

A. Task Specification

We denote a video as $\mathcal{V}$ and a natural language sentence as $\mathcal{S}$. The video consists of T frames: $\mathcal{V} = \{\tilde{\mathbf{f}}_i\}_{i=1}^{T}$, and the sentence $\mathcal{S}$ consists of L words: $\mathcal{S} = \{\tilde{\mathbf{w}}_j\}_{j=1}^{L}$. $\mathcal{S}$ describes an event happening in $\mathcal{V}$. The task is to find a video clip $\{\tilde{\mathbf{f}}_i\}_{i=t_s}^{t_e}$ within $\mathcal{V}$, which has the best semantic matching with $\mathcal{S}$ among the entire $\mathcal{V}$. Each video-sentence pair corresponds to only one video interval. The system is to predict the temporal interval (t_s, t_e). For each video, there could be multiple descriptions. Different descriptions could match to different intervals.

To evaluate the system, the system is asked to generate a list of predicted intervals for each video-sentence pair. If there is a predicted interval has the IoU greater than a threshold m to the ground truth within the top k predictions, it is considered as "correct" for Recall@k, IoU $= m$.

B. Motivation

Given a frame-level visual feature sequence $\mathcal{F}^0 = \{\mathbf{f}_i^0\}_{i=1}^{T}$ and a word-level feature sequence $\mathcal{W}^0 = \{\mathbf{w}_i^0\}_{i=1}^{L}$, the most direct idea is to encode these two sequences $\mathcal{F}^0$ and $\mathcal{W}^0$ into a series of representation vectors $\{\mathbf{f}^c\}$ and a sentence representation $\mathbf{w}$ which represent the clip proposals and the given description, followed by similarity measurement between $\mathbf{w}$ and each $\mathbf{f}^c$. This pipeline is actually utilizing language information as the "final class label" for sampled video clips. In this manner, detailed video-sentence relations can not be exploited. Besides, the information available for semantics inference of each video clip is limited in the sampled range, which may lead to inaccurate video content understanding. While compared to a single class label, natural language descriptions contain much richer semantics, which themselves can be taken as cues of video understanding.

Consider the first example shown in Figure 1. To correctly understand the semantics of the target clip, information from both contextual video contents (where the tent is visible) and natural language description (*"tent is not in these scenes"*) are needed. Since the key object *"tent"* is not visible in the target clip, the model needs to first look at other frames to obtain the knowledge about the object *"tent"* shown in this video. Then, based on this knowledge and the given description, the model is able to locate the clip where the key object *"tent"* is not visible. In the second example shown in Figure 1, the model is required to count the number of occurrences and locate a specific event *"zoom in to the stage"*. The model should first observe the entire video and obtain knowledge about the event *"zoom in"*, followed by counting and localization based on the given description.

According to the above observations, we propose a joint model that exploits vision-language and contextual relations together to tackle this task. We name this model as Interaction-Integrated Network (I^{2}N). Within I^{2}N, a unified cell is utilized as the basic building block. We denote this cell as Interaction-Integrated Cell (I^{2}C). As discussed before, apart from conventional temporal video hierarchy building, I^{2}C also requires two more abilities: the ability of modeling multi-level vision-language cross-modal relation and the ability of modeling long-range context. This enables each frame of the video to receive information from all other frames, unlike in regular convolution which only allows a limited size of the temporal receptive field. As a side benefit, the language processing branch can collect cues from the vision branch and,

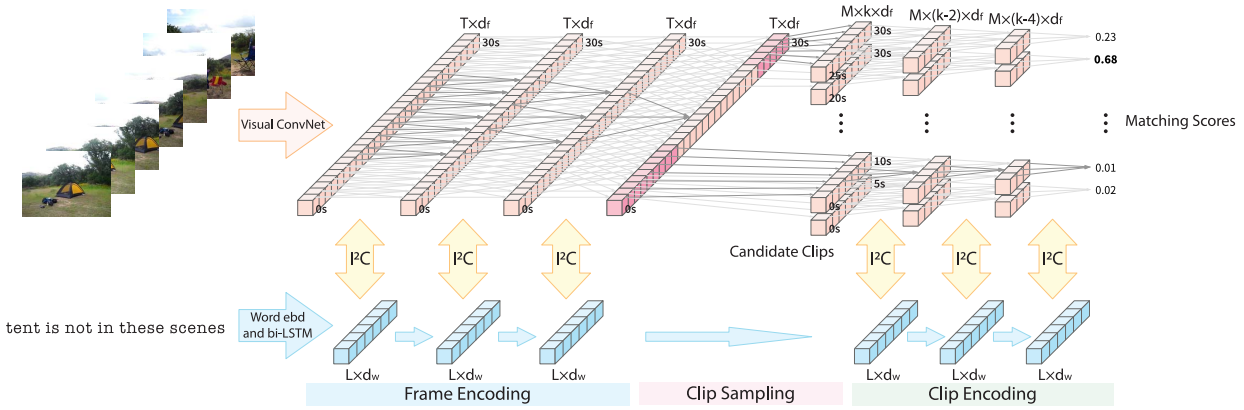

Fig. 2. Illustration of the architecture of our model, I²N (best viewed in color). It consists of three stages: frame encoding, clip sampling and clip encoding. In this figure, each encoding stage contains three interaction cells. The temporal structure made up by 1D convolutions are emphasized as gray arrows. The batch size is omitted for simplicity.

after some processing, feed them back to the vision branch. The implementation of our idea is described in the following parts.

C. Framework: Interaction-Integrated Network

To tackle this task, we first propose a basic framework for video temporal hierarchy modeling via 1D convolution, as shown in the upper half of Figure 2. Then, we introduce I²C, which is used as basic building blocks. When integrating 1D convolution operations with I²Cs, the framework becomes I²N. We take a frame-level visual feature sequence $\mathcal{F}^0 = \{\mathbf{f}_i^0\}_{i=1}^T$ and a word-level linguistic feature sequence of the description $\mathcal{W}^0 = \{\mathbf{w}_i^0\}_{i=1}^L$ as inputs, where T and L are the lengths of the video and word sequence, respectively. As outputs, it generates a series of candidate clips with temporal interval $(\hat{t_s}, \hat{t_e})$, and a confidence score $\hat{s}$ indicating whether this clip matches with the given description. In this manner, I²N is able to utilize full contents in the video as well as the language description, in order to infer whether a clip is relevant to the query, rather than determining such relevance merely based on the information contained by each clip.

A bi-LSTM first encodes the word embeddings into contextual words for further modeling. We intend to encode video clips with rich context information to model complex temporal structures. Therefore, we design the framework with three stages: *frame encoding*, *clip sampling* and *clip encoding*. In the frame encoding stage, we take the entire video sequence into consideration. We first construct the temporal structure for every frame feature $\mathbf{f}_i^0$, to transform the frames into contextual frames $\mathbf{f}_i'$. We then sample clip representations from the contextual frame sequence for every candidate clips. Lastly, we construct clip structures for every candidate clip in the clip encoding stage, and predict the final matching score for each clip.

1) Frame Encoding: In the frame encoding stage, we adopt dilated 1D convolution with kernel size 3 to construct the temporal structure. Starting from the first layer, the dilation rates increase as $\{1, 2, 4, \ldots\}$. Dilated convolution increases receptive fields in an exponential speed [49], while also preserves temporal resolution for clip sampling. This stage builds coarse context for every frame. At the end of this stage, $\mathcal{F}^0$ is transformed into contextual frame sequence $\mathcal{F}'$ with the same length T.

2) Clip Sampling: We then sample candidate clips on the contextual frame sequence $\mathcal{F}'$ following the single-shot object detection manner [50]. We sample M clips in total, with M depending on the experimental settings of different datasets. For each clip, we uniformly sample k frames from the contextual frame sequence, with linear interpolation. The interpolation enables our model to sample clips with arbitrary lengths. Each candidate clip is represented as a frame-level feature sequence $\{\mathbf{f}_1', \mathbf{f}_2', \ldots, \mathbf{f}_k'\}$ and temporal coordinates (τ_{c0}, τ_{l0}), where τ_{c0} and τ_{l0} represent the center and the length of the sampled clip, respectively.

3) Clip Encoding: In this final stage, we build temporal structures for every candidate clip by conventional 1D convolution with kernel size 3. We reduce the temporal resolution from k to 1 gradually. Each layer reduces the temporal resolution by 2. At the last cell, we predict matching score $\hat{s}$ for each clip, representing the correspondence of this clip to the given description. We also predict log deviations of temporal coordinates $\hat{\tau_c^*}$ and $\hat{\tau_l^*}$ for each clip [51], and compute:

$$\hat{\tau_c} = \hat{\tau_c^*}\tau_{l0} + \tau_{c0}, \quad \hat{\tau_l} = \exp(\hat{\tau_l^*}\tau_{t0}), \tag{1}$$

where $\hat{\tau_c}$ and $\hat{\tau_l}$ are our predicted clip center and length, respectively. And τ_{c0} and τ_{l0} represents the sampled clip center and length respectively. Finally, the predicted clip interval $(\hat{t_s}, \hat{t_e})$ can be then represented as:

$$(\hat{t_s}, \hat{t_e}) = (\hat{\tau_c} - \frac{\hat{\tau_l}}{2}, \hat{\tau_c} + \frac{\hat{\tau_l}}{2}). \tag{2}$$

D. Building Block: Interaction-Integrated Cells

The job of Interaction-Integrated Cells (I²Cs) is processing both visual and linguistic information in I²N. In the cell, we perform both cross-modal interaction and contextual interaction, along with temporal video hierarchy building. I²Cs are used as basic building blocks of I²N.

Each cell takes a frame-level visual feature sequence $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^t$ and a word-level linguistic feature sequence of the

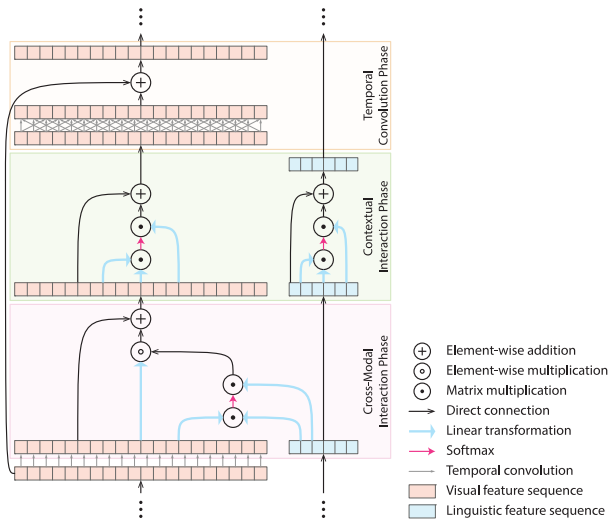

Fig. 3. Illustration of an I^{2}C, which consists of three phases: cross-modal interaction, contextual interaction and temporal convolution. We present a self-attention-based contextual interaction and a dilated temporal convolution with dilation rate 2 in this figure.

query $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^{l}$ as input, where t and l are the lengths of the input visual and linguistic sequences for this cell, respectively. $\mathbf{f}_i$ and $\mathbf{w}_i$ are the frame-level visual representation vector with dimension d_f and the word-level representation vector with dimension d_w, respectively. For each frame in $\mathbf{F}$, the cell applies transformation $\mathrm{f}(\mathbf{f}_i) = (\mathrm{f_{cv}} \circ \mathrm{f_{ct}} \circ \mathrm{f_{cm}})(\mathbf{f}_i, \mathcal{F}, \mathcal{W})$, where $\mathrm{f_{cm}}$, $\mathrm{f_{ct}}$ and $\mathrm{f_{cv}}$ represent the **cross-modal interaction function**, the **contextual interaction function** and the **temporal convolution function**, respectively. They are executed serially, and Figure 3 shows the overall structure of an I^{2}C.

1) The Cross-Modal Interaction Phase: In the cross-modal interaction phase, we explore cross-modal relations between every frame of the visual feature and every word of the linguistic feature. The reason we utilize word-to-frame relation is that we want to align vision-language information on finer granularities. Compared to using the whole sentence as the language cue, the word-level information enables more detailed interaction, therefore more accurate semantic alignment. After exploring the word-to-frame relation, we transform the word-level linguistic features into filters for the visual feature.

We first compute a frame-to-word similarity matrix, and obtain a weighted sum of word vectors for each frame $\mathbf{f}_i$:

$$\mathbf{w}_i' = \mathrm{softmax}\left(\frac{\theta(\mathbf{f}_i)^\top \phi(\mathcal{W})}{d_f}\right) \cdot \psi(\mathcal{W}), \tag{3}$$

where d_f is the dimension of the feature vectors, θ, ϕ and ψ are linear transformation functions.

We then use $\mathbf{w}_i'$ as the dynamic filter for $\mathbf{f}_i$, transforming $\mathbf{f}_i$ by element-wise multiplication [52]:

$$\mathrm{f_{cm}}(\mathbf{f}_i, \mathcal{W}) = \mathbf{f}_i + \mathbf{w}_i' \odot \omega(\mathbf{f}_i), \tag{4}$$

where $\odot$ denotes the element-wise multiplication, and ω is a linear transformation function.

2) The Contextual Interaction Phase: In the contextual interaction phase, each frame in $\mathcal{F}$ interacts with all other frames. With this phase, the frames can break the limit of convolution receptive fields, incorporating the information from all over the video. Modeling long-ranged semantic dependencies becomes possible.

In our I^{2}Cs, we implement this phase by summarizing a context vector into each frame feature. The method for context summarizing is not unique. Here we introduce three methods for it.

Average Pooling: Average pooling is a simple and popular approach to summarize information from multiple feature vectors into one feature vector. In our case, it can be formulated as:

$$\mathrm{f_{ct}}(\mathbf{f}_i, \mathcal{F}) = \mathbf{f}_i + \phi(\mathrm{avg}(\mathcal{F})). \tag{5}$$

Squeeze-and-Excitation: The Squeeze-and-Excitation (SE) module [53] first aggregates features to produce a descriptor that contains global channel distribution. The descriptor is then transformed into a gate, to adaptively re-calibrate the channel-wise visual feature. The SE module can be formulated as:

$$\mathrm{f_{ct}}(\mathbf{f}_i, \mathcal{F}) = \mathbf{f}_i + \sigma(\phi(\mathrm{avg}(\mathcal{F}))) \odot \mathbf{f}_i, \tag{6}$$

where σ is the sigmoid function.

Self Attention: Self attention explores relations for every pair of elements from the input sequence. It has been proven effective for natural language processing [54] and computer vision [55]. The self attention mechanism enables every element to interact with all other elements despite the limit of receptive fields. Compared to the previous two mechanisms, it directly summarizes global context based on content similarities between each element pair. It can be formulated as:

$$\mathrm{f_{ct}}(\mathbf{f}_i, \mathcal{F}) = \mathbf{f}_i + \mathrm{softmax}\left(\frac{\theta(\mathcal{F})^\top \phi(\mathcal{F})}{d_f}\right) \cdot \psi(\mathcal{F}). \tag{7}$$

In the contextual interaction phase, we also apply self attention to word sequence, to enable it to dynamically adjust semantic representation along with visual hierarchy.

3) Temporal Convolution Phase: After the two phases above, we apply a 1D convolution with kernel size 3 to build the temporal video hierarchy, as we proposed in the framework. At the frame encoding stage, we utilize dilated convolution as $\mathrm{f_{cv}}$ to build the temporal structure for every frame, while also preserving the temporal resolution of the video. At the clip encoding stage, we use conventional temporal convolution as $\mathrm{f_{cv}}$ to build the temporal structure for each clip. The temporal resolution gradually reduces to 1. At the end of the clip encoding, we predict the matching score and log deviations of temporal coordinates for each clip as proposed in Section III-C.

IV. Experiments

A. Datasets and Implementation Details

1) DiDeMo: DiDeMo consists of 33,005, 4,180 and 4,021 video-sentence pairs for training, validation and testing, respectively. DiDeMo videos are taken in unconstrained

open-world scenes. To simplify the process of annotation, each video has a fixed length of 30 seconds. The 30-second videos are then divided into 5-second segments. Multiple annotators are asked to select the most corresponding segment. Therefore, there are 21 candidate clips for each video. We directly sample these 21 clips at the clip sampling stage. The evaluation metrics are Rank@1, Rank@5 and mIoU. Rank@k measures the proportion of localization results which the ground truth clip is contained in the top k proposals. mIoU denotes the mean IoU between the ground truth and localization results.

For the DiDeMo dataset, we sample 256 frames of two-stream representation [22] across 30 seconds as our input visual feature. Specifically, following MCN [15] and TGN [42], we use VGG16 [56] as the RGB branch and Inception-BN [57] as the optical flow branch.

At the last cell, we only predict the matching score, since the candidate clips are predefined by the dataset. We apply a 21-way cross-entropy loss over the candidate clips to train the model in a discriminative manner. The ground truth clip is treated as positive sample, and the rest 20 clips are treated as negative samples. The loss function for this model is represented as: $L = -\log(e^{\hat{y}_g} / \sum_i e^{\hat{y}_i})$.

2) Charades-STA: Charades-STA consists of 12,408 video-sentence pairs for training and 3,720 pairs for testing. The videos of Charades-STA come from the Charades dataset [58]. These videos focus on daily activities from indoor scenes. There are no predefined candidate clips provided by the dataset. The evaluation metrics are Recall@1 and Recall@5 at IoU 0.5 and 0.7. Recall@k, IoU m measures the proportion of localization results, where within the top k localized clips, one clip and the ground truth have the IoU greater than m.

We sample 5 frames per second and totally 256 frames of C3D representation [23] as our input visual feature following Gao *et al*. [14]. Since there are no predefined clip provided, we sample clips from 6 scales, $\{\frac{1}{1}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$, of the complete model length. Neighboring two clips within the same scale have an overlap of $\frac{2}{3}$ length of the clip. At the last cell, we predict both matching scores and temporal coordinate deviations. These two targets are trained by the 2-way cross entropy loss and the smooth-L1 regression loss respectively. Any candidate clip which has IoU greater than 0.5 with the ground truth clip is treated as positive sample. The rest samples are treated as negative samples. The loss function is represented as: $L = -\log(e^{\hat{y}_g} / \sum_i e^{\hat{y}_i}) + \sum_{i\in\{c,l\}} \text{smooth}_{\text{L1}}(\hat{\tau_i^*} - \tau_i)$. At inference, we apply Non-Maximum Suppression (NMS) with threshold 0.8 to suppress highly overlapped localization results.

To make comparisons with most recent methods, we also utilize I3D feature [11] pretrained on the Kinetics dataset with our I^{2}N. Surrounding 16 RGB frames are used for I3D feature extraction.

3) TACoS: TACoS [16] consists of 9,030 video-sentence pairs for training and 3,607 pairs for testing. There are no predefined candidate clips provided by the dataset. The evaluation metrics are Recall@1 and Recall@5 at IoU 0.3 and 0.5. For TACoS, we use C3D feature following most previous methods. The clip sampling strategy and training strategy are same with Charades-STA.

TABLE I

ABLATION STUDIES. THE FIRST SECTION SHOWS THE ABLATION STUDY ON COMPONENTS OF I^{2}C, FOLLOWING WHICH ARE ABLATION STUDIES ON CROSS-MODAL INTERACTION PHASE AND CONTEXTUAL INTERACTION PHASE

Methods	DiDeMo			Charades-STA			
	Rk@1	Rk@5	mIoU	Rec@1 IoU0.5	Rec@1 IoU0.7	Rec@5 IoU0.5	Rec@5 IoU0.7
Baseline	27.99	73.28	43.97	36.45	18.09	69.52	40.38
Baseline 2	30.89	75.67	47.57	36.96	18.76	71.42	41.83
Full model	33.06	78.80	49.45	41.69	22.88	75.73	46.85
Full model 2	32.89	78.52	49.12	41.02	22.12	73.44	43.90
Addition	32.15	78.37	48.27	40.51	22.15	73.71	44.52
Concatenation	31.05	77.30	47.72	38.41	21.26	71.21	43.20
Sentence	31.48	77.92	47.90	40.13	22.12	72.12	44.92
Avg Pooling	32.58	78.06	49.38	38.27	21.83	72.61	43.20
Sqz&Exc	32.15	78.21	48.57	40.56	22.82	75.05	43.87
w/o word self-att	32.13	78.52	48.94	40.40	22.50	75.54	46.69
w/o dilation	32.70	78.21	49.29	38.63	20.94	70.11	40.62

For all datasets, we utilize GloVe embedding [59] as our input word feature and Adam [60] with a learning rate 0.0005 as our optimizer. In our experiments, we use four and three cells to construct the frame encoding stage and the clip encoding stage, respectively. We choose self attention as our contextual interaction function by default, and implement our model in TensorFlow [61].

B. Ablation Studies

In ablation studies, we evaluate the importance of various modules in I^{2}N. First, we study the impact of each component of I^{2}C. We then perform ablation studies for each phase of I^{2}C. The experiments are performed on the validation subset of DiDeMo and the testing set of Charades-STA. We use concatenated two-stream feature as input for DiDeMo, and C3D feature as input for Charades-STA.

1) Ablation Study on the Components of I^{2}C: In this experiment, we study the influence of the components of I^{2}C.

The **baseline** model is a plain framework. It contains only basic temporal convolution phases, plus cross-modal interaction phases only in the clip encoding stage. In this model, vision-language relations are built after clip sampling, which is a strategy similar to [45].

Based on the **baseline** model, we include the cross-modal interaction phases into the frame encoding stage. It allows the model to exploit vision-language relations in earlier stages. We denote this model as **baseline 2**.

We then introduce the contextual interaction phases into the **baseline 2** model, allowing I^{2}N to model long-range relations between frames, which becomes our **full model**. The results are shown in the first part of Table I. Both the cross-modal interaction and the contextual interaction improve the model's performances significantly.

The order of components in I^{2}C is an interesting point to study. The temporal convolution phase aggregates visual information on the temporal domain, and generates visual features on higher levels. Performing it early makes the language information hard to interact with low-level and more detailed

visual information. We change the order of the cross-modal interaction phase and the contextual interaction phase in this part, denoted as **full model 2**. On both datasets, **full model 2** significantly outperformed by **full model**. The reason is performing cross-model interaction early enables the model to gather information from inputs from both modalities to help visual information processing, which is also suggested by [40].

2) Ablation Study on the Cross-Modal Interaction Phase: In this ablation study, we analyze the influence of different cross-modal interaction methods.

In the cross-modal interaction phase of I^{2}C, the interaction between word feature $\mathbf{w}_i'$ and frame feature $\omega(\mathbf{f}_i)$ is performed by element-wise multiplication. Language data can be seen as convolutional filters for visual data. In this part, we replace the element-wise multiplication with other two popular fusion methods: **addition** and **concatenation**.

We also try to replace language information from the word-level feature sequence $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^{L}$ to the sentence-level feature vector $\mathbf{s}$. $\mathbf{s}$ is the concatenation of the last hidden states from the bi-LSTM. We denote this model as **sentence**. The results are shown in the second part of Table I. By replacing the element-wise multiplication with addition and concatenation, both the models achieve worse performances. Also, the sentence-level description representation does not discover the relation between frames to words, leading to less accurate alignment.

3) Ablation Study on the Contextual Interaction Phase: We analyze the influence of different context modeling methods in this ablation study.

We choose average pooling, squeeze-and-excitation and self attention. Our full model makes use of self attention as the context summarization mechanism. The third part of Table I shows the experimental results of the **average pooling** model and the **squeeze-and-excitation** model. On both the datasets, the self attention model achieves the best results on all metrics. The self attention mechanism encodes context information for each frame based on its own content. Compared to the other two methods, it is more flexible.

Compared to the **baseline 2** model, all models with the contextual interaction phase achieve significant better performances. Even with a simple average pooling, the models can be beneficial a lot from the coarse contexts. This also demonstrates the importance of global context awareness.

In the contextual phase of I^{2}C, we apply self attention for word-level feature sequence to build language hierarchy. We evaluate the influence of this module for language data. The last part of Table I shows the experimental results. Removing linguistic attention causes a performance drop on all metrics slightly. Compared to the visual counterpart, linguistic attention plays a less important role in our model.

4) Ablation Study on the Temporal Convolution Phase: The dilated convolution is an important part to build temporal video structure. We analyze the influence of dilated convolution in this ablation study. We replace the dilated convolution with conventional convolution in the frame encoding stage, denoted as **w/o dilation** in Table I. By removing dilation, the performances significantly dropped. This ablation study demonstrates even though the global context is available for each frame, the local receptive field and local temporal structure are still important.

TABLE II
COMPARISON TO STATE-OF-THE-ARTS ON DIDEMO

Methods	Rank@1	Rank@5	mIoU
TMN [62]	22.92	76.08	35.17
MCN [15]	28.10	78.21	41.08
TGN [42]	28.23	**79.26**	42.97
I^{2}N-TS	32.45	77.10	**48.55**
I^{2}N-TSLF	**32.58**	78.36	48.16
MAN [46]	27.02	**81.70**	41.16
I^{2}N-I3D	**29.00**	73.09	**44.32**

C. Comparison With the State-of-the-Arts

1) DiDeMo: Table II shows our experimental results compared to state-of-the-arts on the DiDeMo dataset. **I^{2}N-TS** indicates the model with the concatenated two-stream feature as visual input. Following MCN [15] and TGN [42], we also train two models with RGB and optical flow features as visual inputs independently, followed by late fusion on the results, denoted as **I^{2}N-TSLF**. Our two models can outperform state-of-the-arts on Rank@1 and mIoU by a large margin. **I^{2}N-TS** achieves 4.22% and 5.58% performance improvement. And **I^{2}N-TSLF** achieves 4.34% and 5.19% performance improvement. It achieves 78.36% on Rank@5, which is a 0.90% performance inferior to TGN. MAN [46] uses I3D feature as visual input. I3D captures video dynamics on RGB frames without the help of optical flow. By replacing the two-stream feature with I3D, our I^{2}N achieves significantly worse performances. Compared to MAN, **I^{2}N-I3D** achieves 1.98% performance improvement on Rank@1.

2) Charades-STA: Table III shows our experimental results compared to state-of-the-arts on the Charades-STA dataset. **I^{2}N-C3D** outperforms state-of-the-arts by 2.22% on R@1, IoU $=$ 0.5 and 4.01% on R@1, IoU $=$ 0.7. On R@5s, [45] outperforms our I^{2}N by 3.67% for IoU $=$ 0.5. But for IoU $=$ 0.7, our I^{2}N achieves a better accuracy of 1.45% improvement. By replacing the C3D feature with the I3D feature, **I^{2}N-I3D** achieves significantly performance improvement due to more accurate visual representation. In particular, it outperforms MAN by 5.75% and 8.60% in terms of R@1 accuracy at IoUs of 0.5 and 0.7, respectively. SCDM [70] outperforms our I^{2}N on R@1, IoU $=$ 0.5, R@1, IoU $=$ 0.7 and R@5, IoU $=$ 0.7. But our I^{2}N outperformes SCDM on R@5, IoU $=$ 0.5. The video temporal modeling in SCDM is implement by a lot of convolution layers. On Charades-STA, it also perceives wide range of temporal context. With I3D feature finetuned on the Charades dataset,[1] I^{2}N gains further improvement of 4.33% and 2.82% on R@1, IoU $=$ 0.5 and R@1, IoU $=$ 0.7. We denote this model as **I^{2}N-I3D*** in Table III. Our I^{2}N outperforms ExCL [72] by a clear margin. LGI [73] achieves the best performances among all methods. It benefits from its carefully designed language encoder.

[1] https://github.com/piergiaj/pytorch-i3d

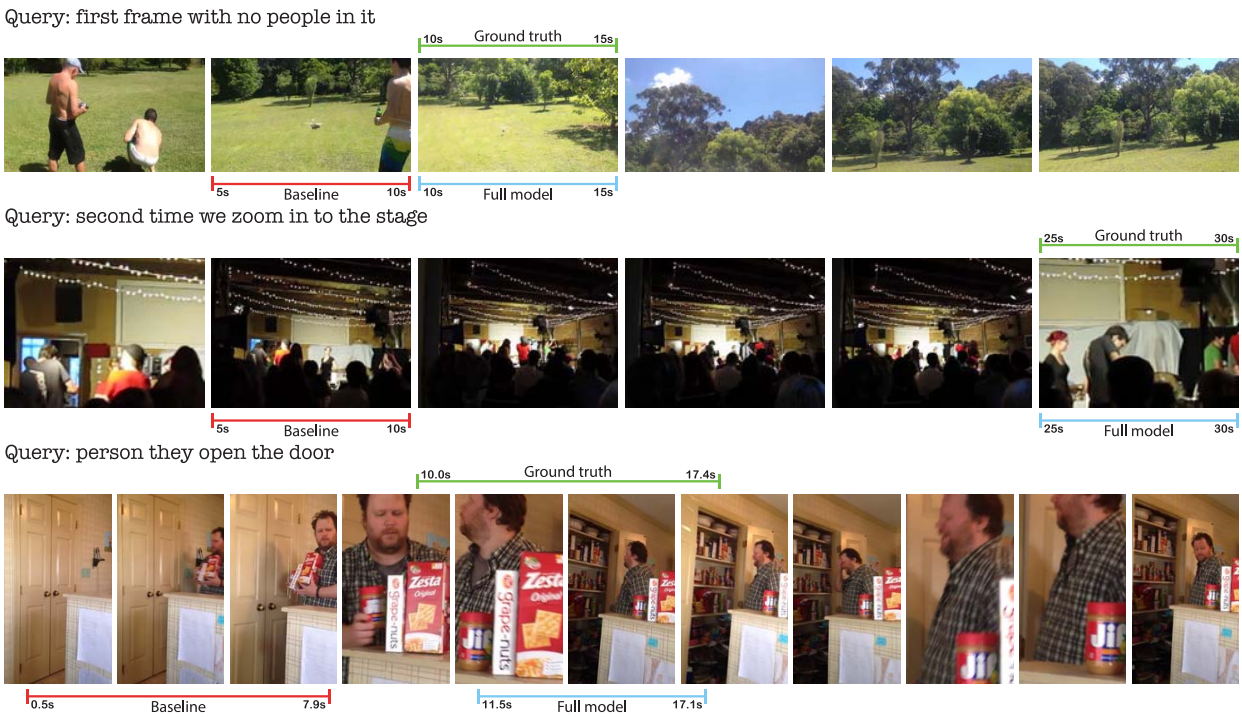

Fig. 4. Qualitative analysis of our model with Rank@1 predictions. The first two cases are from DiDeMo. The last case comes from Charades-STA. Ground truth, wrong prediction and correct prediction are marked as green, red and blue. In these three cases, our model can successfully localize the clips by queries with complex context description.

TABLE III
COMPARISON TO STATE-OF-THE-ARTS ON CHARADES-STA

Methods	Rec@1 IoU0.5	Rec@1 IoU0.7	Rec@5 IoU0.5	Rec@5 IoU0.7
CTRL [14]	23.63	8.89	58.92	29.52
ROLE [63]	12.12	-	40.59	-
ACL [64]	30.48	12.20	64.84	35.13
SAP [65]	27.42	13.36	66.37	38.15
LSTM [45]	35.6	15.8	**79.4**	45.4
TSP-PRL [66]	37.39	17.69	-	-
DEBUG [67]	37.39	17.69	-	-
GDP [68]	39.47	18.49	-	-
CBP [69]	36.80	18.87	70.94	**50.19**
I^{2}N-C3D	**41.69**	**22.88**	75.73	46.85
MAN [46]	46.53	22.72	**86.23**	53.72
SCDM [70]	**54.44**	33.43	74.43	**58.08**
[71]	52.02	**33.74**	-	-
I^{2}N-I3D	52.28	31.32	80.65	54.17
ExCL [72]	44.1	22.4	-	-
LGI [73]	**59.36**	**35.48**	-	-
I^{2}N-I3D*	56.61	34.14	**81.48**	**55.19**

TABLE IV
COMPARISON TO STATE-OF-THE-ARTS ON TACOS

Methods	Rec@1 IoU0.3	Rec@1 IoU0.5	Rec@5 IoU0.3	Rec@5 IoU0.5
CTRL [14]	18.32	13.30	36.69	25.42
ACRN [74]	19.52	14.62	34.97	24.88
ABLR [43]	19.7	9.4	-	-
TGN [42]	21.77	18.90	39.06	31.02
SCDM [70]	26.11	21.17	40.16	32.18
GDP [68]	24.14	-	-	-
DEBUG [67]	23.45	11.72	-	-
2D-TAN [48]	**37.29**	25.32	**57.81**	45.04
CBP [69]	27.31	24.79	43.64	37.40
I^{2}N-C3D	31.47	**29.25**	52.65	**46.08**

3) TACoS: Table IV shows our experimental results compared to state-of-the-arts on the TACoS dataset. **I^{2}N-C3D** outperforms other methods by a clear margin on R@1, IoU = 0.5. On other metrics, our method also achieves competitive results. Though SCDM achieves better performances on Charades-STA than our I^{2}N, our model outperforms SCDM on the TACoS dataset. The reason is videos in TACoS is much longer than videos in Charades-STA. The receptive field of multi-layer convolution is still a limit for long videos. 2D-TAN clearly outperforms our method on R@1, IoU = 0.3 and R@5, IoU = 0.3. While under the critirion IoU = 0.5, I^{2}N outperforms 2D-TAN on both R@1 and R@5. The 2D temporal operation in 2D-TAN enables it to explore temporal relation in a much more smart way. But our I^{2}N also achieves competitive performances with rather simple temporal modeling strategy. It demonstrates the effectiveness of our arrangement of modules.

The above experimental results show the effectiveness of our I^{2}N. Compared to the state-of-the-arts, our I^{2}N improves a lot, especially in terms of top-1 localization accuracy. We notice that our method achieves relatively weak performances on the Rank@5 metrics, especially at low IoU criterions. We believe the main reason is the annotation bias of the datasets. The length of the most ground truth clips are 5s for DiDeMo, and there are only 6 possible clips with length 5s. As for

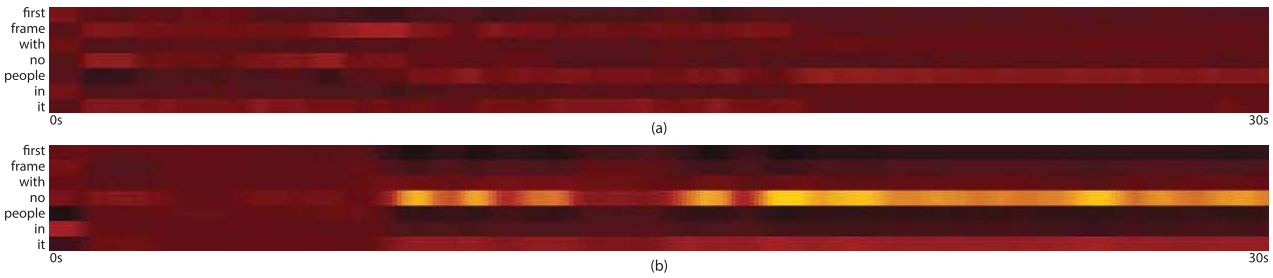

Fig. 5. Visualization of attention weights of case "First frame with no people in it". (a) shows the frame-to-word attention matrix of the first I^{2}C, in which each frame pays nearly equal attention on every words. (b) shows the frame-to-word attention matrix of a later I^{2}C, in which significant attentions are assigned on the word "no". Brighter cells indicate higher attention values.

Charades-STA, the most ground truth clips are about 10s, and the videos are about 30s. Randomly selecting serveral clips is highly possible to hit on Rank@5, especially at low IoU criterions, therefore achieving high Rank@5 accuracies. So we believe more strict metrics such as Rank@1 and higher IoU criterions are more reliable and convincing.

D. Qualitative Analysis and Discussions

In Figure 4, we show qualitative analysis of our method. The first two cases are from DiDeMo and the last one is from Charades-STA. We visualize the top 1 localization results of our full model and the baseline model. In the first case, the query looks for a clip that contains *"first frame with no people in it"*. The baseline model localizes a clip where people can be seen. It indicates that the baseline model failed to utilize the linguistic information correctly. Phrase *"no people"* is the target event, and it relies on the knowledge *"people"*. To correctly localize the clip with this query, the understanding of content from the entire video, knowing *"people can be seen"* at the start of the video is mandatory. Then the first frame with *"no people"* can be successfully localized.

In the second case, the baseline model localizes a clip that near the first *"zoom in to the stage"*, and fails to focus on *"second time"*. The major challenge here is to count the occurence of the target event. Our I^{2}N encodes the positional information implicitly into visual features through 1D convolutions with large receptive fields. With the Context-Interaction Phase of the I^{2}C, I^{2}N can easily notice similar events disregard of the temporal distance, therefore achieving the ability of counting.

In the third case, the person approaches the door in 0–8s, without opening the door. He first put down the snacks, and opens the door from 10s. In the training data, most *"open door"* queries are strongly related to the *"approach to door"* scene. Without being aware of contexts, the baseline model mistakenly connect these two concepts, leading to incorrect predictions. I^{2}N, by utilizing contextual restriction and semantic dependencies, manages to recover from confusion and thus produces correct localization results.

Figure 5 shows the frame-to-word attention matrices from I^{2}N for the case *"first frame with no people in it"*. A major challenge in this case is to correctly understand the phrase *"no people"*. In the frame-to-word attention matrix from the first I^{2}C, as shown in Figure 5 (a), each frame pays nearly equal attentions on every words. It implies the early stage of I^{2}N is doing raw event understanding. Figure 5 (b) shows the frame-to-word attention matrix from a later-stage I^{2}C. Significant attentions are assigned to the word *"no"*, indicating that I^{2}N is trying find the frames without people. The visualization of frame-to-word attention matrices verifies the attention changes in I^{2}Cs throughout the network, and suggests that I^{2}N first obtains rough understanding of the entire video, based on which complicated events are detected. Multi-level visual-linguistic interaction is the key of this process.

V. Conclusion

This paper studies a challenging task in video understanding, known as natural language moment localization. Conventional pipelines suffer from an individual manner of checking the relevance of each clip, which limits the model from seeing visual cues across a long time range and taking them into consideration. To deal with this drawback, we present a novel module named I^{2}C, which integrates both cross-modal and contextual interactions and is thus able to make use of contents from both inside and outside of each candidate clip, and facilitate constructing the hierarchy of visual contents with linguistic cues. Stacking a few I^{2}Cs obtains I^{2}N which achieves the state-of-the-art results on two popular benchmarks, namely, DiDeMo and Charades-STA.

Our research reveals that the interaction between vision and language is far from well studied. Also, it remains unclear if there is a better way of supervising these two sources of data to cooperate. These are left for future research.

References

[1] S. Abu-El-Haija *et al.*, "YouTube-8M: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016.
[2] R. Goyal *et al.*, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5842–5850.
[3] C. Gu *et al.*, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6047–6056.
[4] Y. Yang *et al.*, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.
[5] J. Li, S. Zhang, and T. Huang, "Multi-scale temporal cues learning for video person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 4461–4473, Feb. 2020.
[6] Y. Zhuang, M. Cai, X. Li, X. Luo, Q. Yang, and F. Wu, "The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI," *Engineering*, vol. 6, no. 3, p. 245, 2020.
[7] Y. Zhu *et al.*, "Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, Mar. 2020.

[8] Y.-G. Lv, "Artificial intelligence: Enabling technology to empower society," *Engineering*, vol. 6, pp. 205–206, Jan. 2020, doi: 10.1016/j.eng.2020.01.005.

[9] Y. Pan, "Multiple knowledge representation of artificial intelligence," *Engineering*, vol. 6, no. 3, pp. 216–217, Mar. 2020.

[10] Y.-T. Zhuang, F. Wu, C. Chen, and Y.-H. Pan, "Challenges and opportunities: From big data to knowledge in AI 2.0," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 1, pp. 3–14, 2017.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[12] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.

[13] N. Lei *et al.*, "A geometric understanding of deep learning," *Engineering*, vol. 6, no. 3, pp. 361–374, Mar. 2020.

[14] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5267–5275.

[15] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5803–5812.

[16] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

[17] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[18] Y.-G. Jiang *et al.*, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. ECCV Int. Workshop Competition Action Recognit. Large Number Classes*, 2014.

[19] W. Kay *et al.*, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[20] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8668–8678.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 568–576.

[22] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[24] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 344–353.

[25] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5783–5792.

[26] M. Xu, M. Gao, Y.-T. Chen, L. Davis, and D. Crandall, "Temporal recurrent networks for online action detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5532–5541.

[27] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2923.

[28] Y. Xu, Y. Han, R. Hong, and Q. Tian, "Sequential video VLAD: Training the aggregation locally and temporally," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4933–4944, Oct. 2018.

[29] K. Kuang *et al.*, "Causal inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.

[30] L.-K. Zhou, S.-L. Tang, J. Xiao, F. Wu, and Y.-T. Zhuang, "Disambiguating named entities with deep supervised learning via crowd labels," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 1, pp. 97–106, Jan. 2017.

[31] X.-Y. Duan *et al.*, "Temporality-enhanced knowledgememory network for factoid question answering," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 104–115, Jan. 2018.

[32] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[33] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2901–2910.

[34] H. Xue, Z. Zhao, and D. Cai, "Unifying the video and question attentions for open-ended video question answering," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5656–5666, Dec. 2017.

[35] Z. Zhao, S. Xiao, Z. Song, C. Lu, J. Xiao, and Y. Zhuang, "Open-ended video question answering via multi-modal conditional adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 3859–3870, 2020.

[36] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1989–1998.

[37] P. Gao *et al.*, "Question-guided hybrid convolution for visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 469–485.

[38] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. M. Snoek, "Actor and action video segmentation from a sentence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5958–5966.

[39] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1–10.

[40] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6594–6604.

[41] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[42] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 162–171.

[43] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 9159–9166.

[44] B. Jiang, X. Huang, C. Yang, and J. Yuan, "Cross-modal video moment retrieval with spatial and language-temporal attention," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, Jun. 2019, pp. 217–225.

[45] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 9062–9069.

[46] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1247–1257.

[47] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with temporal language," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 1380–1390.

[48] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks formoment localization with natural language," in *Proc. AAAI*, 2020, pp. 12870–12877.

[49] A. V. D. Oord *et al.*, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[50] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[52] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *CoRR*, vol. abs/1609.09106, 2016.

[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[54] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[58] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 510–526.

[59] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[60] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[61] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[62] B. Liu, S. Yeung, E. Chou, D.-A. Huang, L. Fei-Fei, and J. Carlos Niebles, "Temporal modular networks for retrieving complex compositional activities in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 552–568.
[63] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proc. 26th ACM Int. Conf. Multimedia (ACMMM)*, Oct. 2018, pp. 843–851.
[64] R. Ge, J. Gao, K. Chen, and R. Nevatia, "MAC: Mining activity concepts for language-based temporal localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 245–253.
[65] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8199–8206.
[66] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 1–9.
[67] C. Lu, L. Chen, C. Tan, X. Li, and J. Xiao, "DEBUG: A dense bottom-up grounding approach for natural language video localization," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5147–5156.
[68] L. Chen *et al.*, "Rethinking the bottom-up framework for query-based video localization," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 10551–10558.
[69] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12168–12175.
[70] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 536–546.
[71] C. Rodriguez-Opazo, E. Marrese-Taylor, F. S. Saleh, H. Li, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2464–2473.
[72] S. Ghosh, A. Agarwal, Z. Parekh, and A. G. Hauptmann, "Excl: Extractive clip localization using natural language descriptions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 1984–1990.
[73] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10810–10819.
[74] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Jun. 2018, pp. 15–24.

Ke Ning received the B.Sc. degree in computer science from Zhejiang University, in 2015, where he is currently pursuing the Ph.D. degree with the Digital Media Computing and Design (DCD) Laboratory, under the supervision of Prof. F. Wu. His main research interests include video analysis and multimedia analysis.

Lingxi Xie received the B.E. and Ph.D. degrees in engineering from Tsinghua University, in 2010 and 2015, respectively. He served as a Postdoctoral Researcher for the CCVL Laboratory from 2015 to 2019, having moved from the University of California, Los Angeles, CA, USA, to the Johns Hopkins University. He is currently a Senior Researcher at Huawei Inc. His research interest includes computer vision, in particular, the application of deep learning models. His research covers image classification, object detection, semantic segmentation, and other vision tasks. He is also facilitating the application of automated machine learning to the above research fields. He has published more than 50 articles in top-tier international conferences and journals. In 2015, he received an outstanding Ph.D. Thesis Award from Tsinghua University. He is also the Winner of the Best Paper Award at ICMR 2015.

Jianzhuang Liu (Senior Member, IEEE) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a Postdoctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong. In 2011, he joined the Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a Professor. He is currently a Principal Researcher with Huawei Technologies Company Limited, Shenzhen, China. He has authored more than 150 articles. His research interests include computer vision, image processing, deep learning, and graphics.

Fei Wu (Senior Member, IEEE) received the B.Sc. degree from Lanzhou University, in 1996, the M.Sc. degree from the University of Macau, in 1999, and the Ph.D. degree from Zhejiang University, in 2002, all in computer science. From October 2009 to August 2010, he was a Visiting Scholar with the Prof. Bin Yu's Group, University of California, Berkeley, CA, USA. He is currently a Qiushi Distinguished Professor with the College of Computer Science, Zhejiang University. He is also the Vice-Dean of the College of Computer Science, and the Director of the Institute of Artificial Intelligence, Zhejiang University. His main research interests include artificial intelligence, multimedia analysis and retrieval, and machine learning. He has been the Chairman of the IEEE CAS Hangzhou-Chapter, since October 2018. He is also an Associate Editor of *Multimedia System*, the editorial members of *Frontiers of Information Technology and Electronic Engineering*. He has won various honors, such as the Award of National Science Fund for Distinguished Young Scholars of China in 2016.

Qi Tian (Fellow, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, the M.S. degree in ECE from Drexel University, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign (UIUC). He was a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), from 2002 to 2019. From 2008 to 2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA). From 2018 to 2020, he was the Chief Scientist in Computer Vision with the Huawei Noah's Ark Laboratory. He is currently a Chief Scientist in Artificial Intelligence with Cloud BU, Huawei. His research interests include computer vision, multimedia information retrieval, and machine learning, and published more than 610 refereed journal and conference papers. His Google citation is more than 25100 with H-index 77. He was a coauthor of best articles, including IEEE ICME 2019, ACM CIKM 2018, ACM ICMR 2015, PCM 2013, MMM 2013, ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Student Contest Paper in ICASSP 2006, and coauthor of a Best Paper/Student Paper Candidate in ACM Multimedia 2019, ICME 2015, and PCM 2007. He received the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement awards from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), ACM TOMM, MMSJ, and in the Editorial Board of *Journal of Multimedia* (JMM) and *Journal of MVA*. He is the Guest Editor of IEEE textscTransactions on Multimedia (TMM) and *Journal of CVIU*.