Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding

Daizong Liu[1,2], Xiaoye Qu[3], Pan Zhou[1*]

[1]The Hubei Engineering Research Center on Big Data Security, School of
Cyber Science and Engineering, Huazhong University of Science and Technology
[2]School of Electronic Information and Communication,
Huazhong University of Science and Technology
[3]Huawei Cloud

`dzliu@hust.edu.cn, quxiaoye@huawei.com, panzhou@hust.edu.cn`

Abstract

A key solution to temporal sentence grounding (TSG) exists in how to learn effective alignment between vision and language features extracted from an untrimmed video and a sentence description. Existing methods mainly leverage vanilla soft attention to perform the alignment in a single-step process. However, such single-step attention is insufficient in practice, since complicated relations between inter- and intra-modality are usually obtained through multi-step reasoning. In this paper, we propose an Iterative Alignment Network (IA-Net) for TSG task, which iteratively interacts inter- and intra-modal features within multiple steps for more accurate grounding. Specifically, during the iterative reasoning process, we pad multi-modal features with learnable parameters to alleviate the nowhere-to-attend problem of non-matched frame-word pairs, and enhance the basic co-attention mechanism in a parallel manner. To further calibrate the misaligned attention caused by each reasoning step, we also devise a calibration module following each attention module to refine the alignment knowledge. With such iterative alignment scheme, our IA-Net can robustly capture the fine-grained relations between vision and language domains step-by-step for progressively reasoning the temporal boundaries. Extensive experiments conducted on three challenging benchmarks demonstrate that our proposed model performs better than the state-of-the-arts.

1 Introduction

Temporal localization is an important topic of visual understanding in computer vision. There are several related tasks proposed for different scenarios involving language, such as video summarization (Song et al., 2015; Chu et al., 2015), temporal action localization (Shou et al., 2016; Zhao et al.,
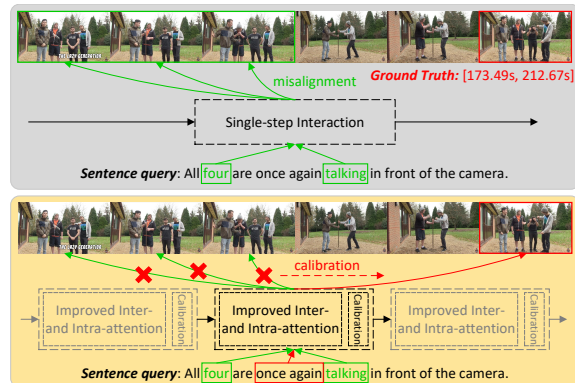

Figure 1: Illustration of our motivation. **Upper**: Previous methods are mainly based on a single-step interaction with attention, which is insufficient to reason the complicated multi-modal relations, thus may lead to the misalignment on semantics. **Bottom**: we develop an iterative network with improved attention mechanism and calibration module, which can progressively align accurate semantic.

2017), and temporal sentence grounding (Gao et al., 2017; Anne Hendricks et al., 2017). Among them, temporal sentence grounding is the most challenging task due to its complexity of multi-modal interactions and complicated context information. Given an untrimmed video, it aims to determine the segment boundaries including start and end timestamps that contain the interested activity according to a given sentence description.

Existing methods mainly focus on learning multimodal interaction by a single-step attention mechanism. Most of approaches (Yuan et al., 2019b; Chen and Jiang, 2019; Zhang et al., 2019a; Rodriguez et al., 2020; Chen et al., 2020; Liu et al., 2020a,b) utilize a simple co-attention mechanism to learn the inter-modality relations between word-frame pairs for aligning the semantic information. Besides, some approaches (Chen et al., 2019; Zhang et al., 2019b; Liu et al., 2021) employ a single-step self-attention to explore the contextual information among intra-modality to correlate relevant frames or words.

---

[*]Corresponding author.

Although these methods achieve promising results, they are severely limited by two issues. 1) These single-step methods only consider the inter- or intra-modal relation once, which is insufficient to learn the complicated multi-modal interaction that needs multi-step reasoning. Besides, the misalignment between inter-modality or the wrong-attention across intra-modality caused by such single-step attention will directly degenerate the performance on the boundary results. As shown in Figure 1, the task targets to localize the query "*All four are **once again** talking in front of the camera*" in the video. It is hard for single-step methods to directly pay more attention on phrase "***once again***", easily leading to the misalignment problem and the wrong grounding result. 2) Nowhere-to-attend problem is generally happened in TSG task, in which the background frames do not match any word in the sentence, and the basic attention may generate the wrong attention weights in these cases.

In this paper, we develop a novel Iterative Alignment Network (IA-Net) for temporal sentence grounding, which addresses the above problems by an end-to-end framework within multi-step reasoning. Specifically, we introduce an iterative matching scheme to explore both inter- and intra-modal relations progressively with an improved attention based inter- and intra-modal interaction module. In this module, we first pad the multi-modal features with learnable parameters to tackle the nowhere-to-attend problem, and enhance the basic co-attention mechanism into a parallel manner that can provide multiple attended features for better capturing the complicated inter- and intra-modal relations. Then, to refine and calibrate the misaligned attention happened in early reasoning step, we develop a calibration module following each attention module to refine the alignment knowledge during the iterative process. By stacking multiple such improved interaction modules, our IA-Net provides effective attention to iteratively reason the complicated relations between the vision and language features step-by-step, providing more accurate segment boundaries.

Our main contributions are three-fold:

- We propose an iterative framework for temporal sentence grounding to progressively align the complicated semantics between vision and language.

- We formulate the proposed iterative matching method with an improved co-attention mechanism to utilize learable paddings to address nowhere-to-attend problem with deep latent clues, and a calibration module to refine or calibrate the alignment knowledge of inter- and intra-modal relations during the reasoning process.

- Extensive experiments are performed to examine the effectiveness of the proposed IA-Net on three datasets (ActivityNet Captions, TACoS, and Charades-STA), in which we achieve the state-of- the-art performances.

2 Related Work

Temporal sentence grounding. Temporal sentence grounding (TSG) is a new task introduced recently (Gao et al., 2017; Anne Hendricks et al., 2017). Formally, given an untrimmed video and a natural sentence query, the task aims to identify the start and end timestamps of one specific video segment, which contains activities of interest semantically corresponding to the given sentence query. To interact video and sentence features, some works align the semantics of video with language by a recurrent neural network (RNN). (Chen et al., 2018) design a recurrent module to temporally capture the evolving fine-grained frame-by-word interactions between video and sentence. (Zhang et al.) propose to apply a bidirectional GRU instead of normal RNN for alignment. However, these RNNs can not align the semantics well in this task. As attention has proved its effectiveness on contextual correlation mining, amount of works tend to align relevant visual features with the query text description by an attention module. (Liu et al., 2018a) design a memory attention mechanism on query sentence to emphasize the visual features mentioned in the sentence. (Wang et al., 2020) use a soft attention on moment features based on the sentence feature. (Zhang et al., 2020) adopt a simple multiplication operation for visual and language feature fusion. Moreover, the visual-textual co-attention module is widely utilized to model the cross-modal interaction (Liu et al., 2018b; Yuan et al., 2019b; Chen et al., 2019, 2020; Rodriguez et al., 2020; Jiang et al., 2019; Qu et al., 2020; Nan et al., 2021), which performs effective and efficient in most of challenging scenes. There are also some works (Zhang et al., 2019b; Chen et al., 2019) adopt self-attention block to correlate frames or words in each modality for constructing scene meaning.

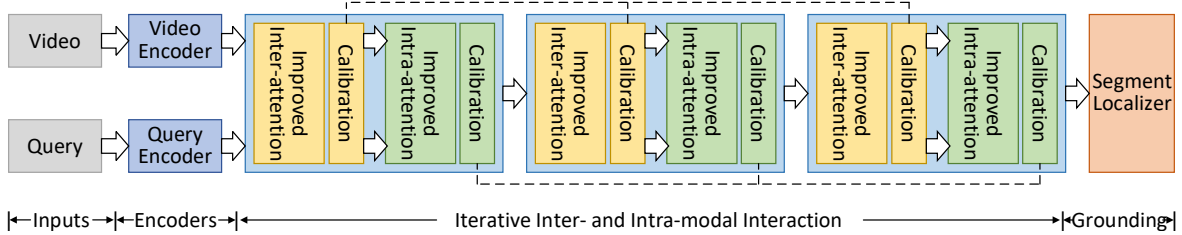Although these attention based methods have

Figure 2: The architecture of our proposed IA-Net. The iterative process is devised for multi-modal interaction.

made great progress in TSG, they are severely limited by such single-step attention mechanism. Motivated by this, we introduce an iterative alignment scheme to explore fine-grained inter- and intra-modal relations. For complicated correlation capturing, we pad the multi-modal features with learnable parameters and enhance the co-attention with multi heads. For semantic misalignment, we additionally develop a calibration module to refine the alignment knowledge. Such iterative process helps our model align more accurate semantic.

Attention mechanism. Attention has achieved great success in various tasks, such as image classification, machine translation, and visual question answering. (Vaswani et al., 2017) propose the Transformer to capture the long-term dependency with multi-headed architecture. Although it attracts great interests from multi-modal retrieval community, it only consider sentence-guided attention on video frames or the video-guided attention on sentence words with complex computation. Compared to it, co-attention mechanism (Lu et al., 2016; Xiong et al., 2016) is proposed to jointly reason about frames and words attention with light weights, which is more suitable for addressing the real-world temporal sentence grounding task. In this paper, we consider the nowhere-to-attend cases in TSG task that frame/word is irrelevant to the whole sentence/video, and address it by utilizing a learnable paddings during the process of attention map generation. We also improve the basic co-attention mechanism into a parallel manner like Transformer which provides multiple latent attended features for better correlation mining.

3 The Proposed Method

The TSG task considered in this paper is defined as follows. Given an untrimmed reference video $\mathcal{V}$ and a sentence query $\mathcal{Q}$, we need to predict the start and end timestamps (τ_s, τ_e), where the segment in $\mathcal{V}$ from time point τ_s to τ_e corresponds to the same

semantic as $\mathcal{Q}$.

In this section, we introduce our framework IA-Net as shown in Figure 2. Our model consists of three main components: video and query encoders, iterative inter- and intra-modal interaction, and the segment localizer. Video and sentence are first fed to the encoders for extracting multi-modal features. Then we iteratively interact their features for semantic alignment. Specially, in each iterative step, we utilize a co-attention to align the inter-modal semantic and an another co-attention to correlate intra-modal instances in each modality. We improve the basic co-attention mechanism in a parallel manner, and devise two calibration modules following the inter- and intra-attention to refine and calibrate the knowledge of cross-modal alignment and self-modal correlation during the iterative interaction process. At last, we utilize a segment localizer to ground the segment boundaries.

3.1 Video and Query Encoders

Video encoder. For video encoding, we first extract the clip-wise features by a pre-trained C3D network (Tran et al., 2015), and then add a positional encoding (Vaswani et al., 2017) to take positional knowledge. Considering the sequential characteristic in video, a bi-directional GRU (Chung et al., 2014) is further utilized to incorporate the contextual information in time series. The output of this encoder is $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$, which encodes the context in video.

Query encoder. For query encoding, we first extract the word embeddings by the Glove word2vec model (Pennington et al., 2014), and also use the positional encoding and bi-directional GRU to integrate the sequential information. The final feature representation of the input sentence is denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N} \in \mathbb{R}^{N \times D}$.

3.2 Improved Inter-modal Interaction

The improved inter-modal interaction module is based on co-attention mechanism to capture the im-

9304

portance between each pair of visual clip and word features. To tackle nowhere-to-attend problem and calibrate the misalignment knowledge, we improve the co-attention in a parallel manner with learnable paddings and devise a calibration module followed by it. Details are shown in Figure 3.

Nowhere-to-attend and parallel attention. Previous co-attention based works in TSG (Yuan et al., 2019b; Chen et al., 2019, 2020) formally compute the attention maps by directly calculating the inner product between V, Q. However, it often occurs at the creation of an attention map that there is no particular frame or word that the model should attend, especially for the background frames that do not match any word in the sentence. This will lead to the wrong attention on the mismatched frame-word pairs. To deal with such cases, we add K elements to both sentence words and video clips to additionally serve for no-attention instances. In details, we incorporate two learnable matrices $[v_{\oslash_1}, ..., v_{\oslash_k}], [q_{\oslash_1}, ..., q_{\oslash_k}] \in \mathbb{R}^{K \times D}$ to V, Q as $\widetilde{V} = [v_1, ..., v_T, v_{\oslash_1}, ..., v_{\oslash_k}] \in \mathbb{R}^{(T+K) \times D}, \widetilde{Q} = [q_1, ..., q_N, q_{\oslash_1}, ..., q_{\oslash_k}] \in \mathbb{R}^{(N+K) \times D}$, respectively. Besides, we also enhance co-attention into multiple attention heads to capture complicated relations in different latent space, and use their average as the attention result.

To generate H number of attention maps, we first linearly project the D-dimensional features of $\widetilde{V}, \widetilde{Q}$ into multiple lower D^--dimensional spaces, where $D^- = D/H$. We take the h-th attention map ($h \in H$) as example:

$$\widetilde{V}^h = \text{Linear}(\widetilde{V}; \Theta_{V,h}), \ \widetilde{Q}^h = \text{Linear}(\widetilde{Q}; \Theta_{Q,h}), \quad (1)$$

where $\text{Linear}(\cdot)$ denotes a fully-connected layer with parameter Θ. Then, we compute the attention map by inner product with row-wise normalization as:

$$A_V^h = \text{Softmax}(\frac{\widetilde{V}^h(\widetilde{Q}^h)^T}{\sqrt{D^-}}) \in \mathbb{R}^{(T+K) \times (N+K)}, \quad (2)$$

$$A_Q^h = \text{Softmax}(\frac{\widetilde{Q}^h(\widetilde{V}^h)^T}{\sqrt{D^-}}) \in \mathbb{R}^{(N+K) \times (T+K)}. \quad (3)$$

We take average fusion of multiple attended features, which is equivalent to averaging H number of attention maps as:

$$A_V = \frac{1}{H}\sum_{h=1}^{H} A_V^h, \ A_Q = \frac{1}{H}\sum_{h=1}^{H} A_Q^h. \quad (4)$$

At last, we can get the V, Q-grounded alignment features M_V, M_Q, in which each element captures related semantics shared by the whole Q, V to each v_t, q_n:

$$M_V = (A_V \widetilde{Q})[1 : T, :] \in \mathbb{R}^{T \times D}, \quad (5)$$

$$M_Q = (A_Q \widetilde{V})[1 : N, :] \in \mathbb{R}^{N \times D}. \quad (6)$$

Calibration module. After receiving the alignment features M_V, M_Q and the multi-modal features V, Q, to refine the alignment knowledge for the next interaction step, we aim to update each modal features V, Q by aggregating them with the corresponding alignment features M_V, M_Q with a gate function dynamically. In details, we first generate a fusion feature for each modality to enhance its semantics by:

$$R_V = \text{Tanh}(W_{R_V} V + U_{R_V} M_V + b_{R_V}), \quad (7)$$

$$R_Q = \text{Tanh}(W_{R_Q} Q + U_{R_Q} M_Q + b_{R_Q}), \quad (8)$$

where W, U, b are the learnable parameters. To select the discriminative information and filter out incorrect one, a gating weight can be formulated as follows:

$$Z_V = \text{Sigmoid}(W_{Z_V} V + U_{Z_V} M_V + b_{Z_V}), \quad (9)$$

$$Z_Q = \text{Sigmoid}(W_{Z_Q} Q + U_{Z_Q} M_Q + b_{Z_Q}). \quad (10)$$

At last, the calibrated output of the current inter-modal interaction module can be obtained by:

$$\widehat{V} = Z_V \odot V + (1 - Z_V) \odot R_V, \quad (11)$$

$$\widehat{Q} = Z_Q \odot Q + (1 - Z_Q) \odot R_Q, \quad (12)$$

where $\odot$ denotes the element-wise multiplication.

The developed gate mechanism has two main contributions: 1) The information of each modality can be refined by itself and the enhanced semantic features shared with R. It helps to filter out trivial information maintained in V, Q, and calibrate the misaligned attention by re-considering its individual shared semantics. 2) The contextual information from alignment features M_V, M_Q summarize the contexts regard to each instances in cross-modal features Q, V, respectively. After the
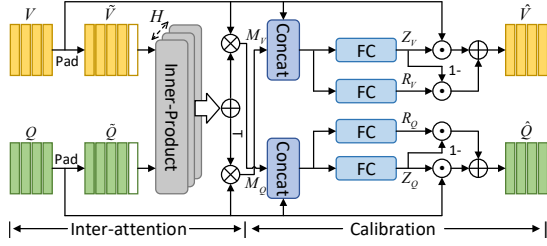
Figure 3: Illustration of our inter-modal interaction.

Figure 4: Illustration of our intra-modal interaction.

gating process, the contextual information maintained in $\widehat{V}, \widehat{Q}$ will also assist to determine the shared semantics in latter attention procedure. It will progressively enhance the interaction among inter-modal features and thus benefit the representation learning.

3.3 Improved Intra-modal Interaction

The output visual clip and sentence word features of the inter-modal interaction module have encoded cross-modal relations between clips and words. With such contextual cross-modal information, we implement intra-attention on each modality to correlate the relevant instances for composing the scene meaning. Different from inter-attention, there is no nowhere-to-attend problem as video or sentence has strong temporal relations in itself. A calibration module is also utilized for self-relation refinement as shown in Figure 4.

3.4 Iterative Alignment with The Improved Inter- and Intra-modal Interaction Block

In this section, we introduce how to integrate the improved inter- and intra-modal interaction modules to enable the iterative alignment for temporal sentence grounding. The inter-modal interaction helps aggregate features from the other modality to update the clip and word features according to the cross-modal relations. The clip and word features would be updated again with the information within the same modality via the intra-modal interaction. We use one inter-modal interaction module followed by one intra-modal interaction module to form a basic improved interaction block (**IIB**) in our proposed IA-Net framework as:

$$\widehat{V}_l, \widehat{Q}_l = \mathbf{IIB}(\widehat{V}_{l-1}, \widehat{Q}_{l-1}), \qquad (13)$$

where l is the block number. Multiple blocks could be further stacked thanks to the calibration module for alignment refinement, helping to reason for the accurate segment boundaries.
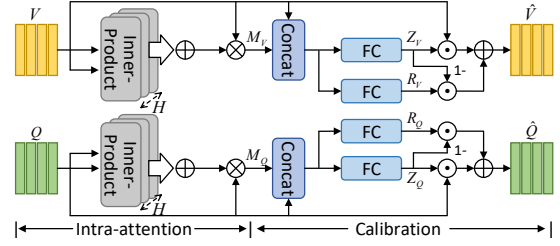
3.5 Segment Localizer and Loss Function

After multiple interaction blocks, we utilize a cosine similarity function (Mithun et al., 2019) on $\widehat{V}, \widehat{Q}$ to generate a new video-aware sentence representation $\widehat{Q}'$ which has the same T-dimensional features like $\widehat{V}$. We fuse two modal features as $\boldsymbol{f}_t = [\widehat{\boldsymbol{v}}_t, \widehat{\boldsymbol{q}}'_t], \boldsymbol{F} = \{\boldsymbol{f}_t\}_{t=1}^T \in \mathbb{R}^{T \times 2D}$ by concatenation. To predict the target video segment, similar to (Yuan et al., 2019a), we pre-define multi-size candidate moments $\{(\hat{\tau}_s, \hat{\tau}_e)\}$ on each frame t, and adopt multiple full connection (FC) layers to process features $\boldsymbol{f}_t$ to produce the confidence scores $\{cs\}$ of all windows and predict corresponding temporal offsets $\{(\hat{\delta}_s, \hat{\delta}_e)\}$. The final predicted moments of time t can be presented as $\{(\hat{\tau}_s + \hat{\delta}_s, \hat{\tau}_e + \hat{\delta}_e)\}$.

Training. We first compute the Intersection over Union score o between each candidate moment $(\hat{\tau}_s, \hat{\tau}_e)$ with ground truth (τ_s, τ_e). If the o is larger than a threshold value λ, this moment is viewed as positive sample, reverse as the negative sample. Thus we can obtain N_{pos} positive samples and N_{neg} negative samples in total (N_{total}). We adopt an alignment loss to align the predicted confidence scores and IoU:

$$\mathcal{L}_{align} = -\frac{1}{N_{total}} \sum o\log(cs) + (1-o)\log(1-cs). \qquad (14)$$

We also devise a boundary loss for N_{pos} positive samples to promote exploring the precise start and end points as:

$$\mathcal{L}_b = \frac{1}{N_{pos}} \sum \mathcal{S}(\hat{\delta}_s - \delta_s) + \mathcal{S}(\hat{\delta}_e - \delta_e), \quad (15)$$

where $\mathcal{S}$ represents the smooth L1 function. We adopt α to control the balance of the alignment loss and boundary loss:

$$\mathcal{L} = \mathcal{L}_{align} + \alpha \mathcal{L}_b. \qquad (16)$$

Testing. We rank all candidate moments according to their predicted confidence scores, and then "Top-

Method	ActivityNet Captions						TACoS					
	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.1	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.1	R@5, IoU=0.3	R@5, IoU=0.5
TGN	45.51	28.47	-	57.32	43.33	-	41.87	21.77	18.90	53.40	39.06	31.02
CTRL	47.43	29.01	10.34	75.32	59.17	37.54	24.32	18.32	13.30	48.73	36.69	25.42
ACRN	49.70	31.67	11.25	76.50	60.34	38.57	24.22	19.52	14.62	47.42	34.97	24.88
CBP	54.30	35.76	17.80	77.63	65.89	46.20	-	27.31	24.79	-	43.64	37.40
SCDM	54.80	36.75	19.86	77.29	64.99	41.53	-	26.11	21.17	-	40.16	32.18
ABLR	55.67	36.79	-	-	-	-	34.70	19.50	9.40	-	-	-
GDP	56.17	39.27	-	-	-	-	39.68	24.14	13.50	-	-	-
CMIN	63.61	43.40	23.88	80.54	67.95	50.73	32.48	24.64	18.05	62.13	38.46	27.02
2DTAN	59.45	44.51	26.54	85.53	77.13	61.96	**47.59**	37.29	25.32	70.31	**57.81**	45.04
DRN	-	45.45	24.36	-	77.97	50.30	-	-	23.17	-	-	33.36
IA-Net	**67.14**	**48.57**	**27.95**	**87.21**	**78.99**	**63.12**	47.18	**37.91**	**26.27**	**71.75**	57.62	**46.39**

Table 1: Performance compared with the state-of-the-art TSG models on ActivityNet Captions and TACoS dataset.

n (Rank@n)" candidates will be selected with non maximum suppression.

4 Experiments

4.1 Datasets

ActivityNet Captions. ActivityNet Captions (Krishna et al., 2017) contains 20k untrimmed videos with 100k descriptions from YouTube. The videos are 2 minutes on average, and the annotated video clips have much larger variation, ranging from several seconds to over 3 minutes. Following public split, we use 37,417, 17,505, and 17,031 sentence-video pairs for training, validation, and testing respectively.

TACoS. TACoS (Regneri et al., 2013) is widely used on TSG task and contain 127 videos. The videos from TACoS are collected from cooking scenarios, thus lacking the diversity. They are around 7 minutes on average. We use the same split as (Gao et al., 2017), which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing.

Charades-STA. Charades-STA is built on the Charades dataset (Sigurdsson et al., 2016), which focuses on indoor activities. In total, there are 12408 and 3720 moment-query pairs in the training and testing sets respectively.

4.2 Experimental Settings

Evaluation Metric. Following previous works (Gao et al., 2017; Yuan et al., 2019a; Zhang et al., 2020), we adopt "R@n, IoU=m" as our evaluation metrics. The "R@n, IoU=m" is defined as the percentage of at least one of top-n selected moments having IoU larger than m.

Implementation Details. We define continuous 16 frames as a clip and each clip overlaps 8 frames

with adjacent clips, and apply C3D (Tran et al., 2015) to encode the videos on ActivityNet Captions, TACoS, and I3D (Carreira and Zisserman, 2017) on Charades-STA. We set the length of video feature sequences to 200 for ActivityNet Captions and TACoS datasets, 64 for Charades-STA dataset. As for sentence encoding, we utilize Glove word2vec (Pennington et al., 2014) to embed each word to 300 dimension features. The hidden state dimension of Bi-GRU networks is set to 512. During segment localization, we adopt convolution kernel size of [16, 32, 64, 96, 128, 160, 192] for ActivityNet Captions, [8, 16, 32, 64, 128] for TACoS, and [16, 24, 32, 40] for Charades-STA. We set the stride of them as 0.5, 0.125, 0.125, respectively. We set the high-score threshold λ to 0.45, and the balance hyper-parameter α to 0.001 for ActivityNet Captions, 0.005 for TACoS and Charades-STA. We train our model with an Adam optimizer with leaning rate 8×10^{-4}, 3×10^{-4}, 4×10^{-4} for Activity Captions, TACoS and Charades-STA, respectively.

4.3 Comparison to State-of-the-Art Methods

Compared methods. We compare our proposed model with the following baseline methods on the TSG task: TGN (Chen et al., 2018), CTRL (Gao et al., 2017), ACRN (Liu et al., 2018a), CBP (Wang et al., 2020), SCDM (Yuan et al., 2019a), ABLR (Yuan et al., 2019b), GDP (Chen et al., 2020), CMIN (Zhang et al., 2019b), 2DTAN (Zhang et al., 2020), and DRN (Zeng et al.). These methods interact multi-modal features only in a single step process.

Analysis. As shown in Tables 1 and 2, we compare our IA-Net with all above methods on three datasets. It shows that IA-Net performs among the best in various scenarios on all three benchmark datasets across different criteria and ranks

| Method | Charades-STA | | | |
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CTRL	23.63	8.89	58.92	29.57
ACRN	20.26	7.64	71.99	27.79
CBP	36.80	18.87	70.94	50.19
GDP	39.47	18.49	-	-
2DTAN	39.81	23.25	79.33	51.15
SCDM	54.44	33.43	74.43	58.08
IA-Net	**61.29**	**37.91**	**89.78**	**62.04**

Table 2: Performance compared with the state-of-the-art TSG models on Charades-STA dataset.

Method	Run-Time	Model Size
CTRL	2.23s	**22M**
ACRN	4.31s	128M
2DTAN	0.57s	232M
IA-Net	**0.11s**	68M

Table 3: Efficiency comparison run on TACoS dataset.

the first or the second in all cases. On ActivityNet Captions, we outperform DRN by 3.59% and 12.82% in the strict metrics "R@1, IoU=0.7" and "R@5, IoU=0.7". We also brings 1.41% and 1.16% improvements compared to 2DTAN. On TACoS dataset, the cooking activities take place in the same kitchen scene with some slightly varied cooking objects, thus it is hard to localize such fine-grained activities. Compared to the top ranked method 2DTAN, our model still achieves the best results on "R@1, IoU=0.5" and "R@5, IoU=0.5", which validates that IA-Net is able to localize the moment boundary more precisely. On Charades-STA, we outperform the SCDM by 6.85%, 4.48%, 15.35% and 3.96% in all metrics. The main reasons for our proposed model outperforming the competing models lie in two folds. First, compared to methods like GDP and CMIN which utilize basic attention to interact multi-modal features, our method provides an improved attention mechanism to address "nowhere-to-attend" problem. We also devise a distillation module to refine and calibrate the alignment knowledge. Second, previous works all adopt a single interaction process with no tolerance on the attention mistake. Thanks to the designed distillation module across the whole framework, our IA-Net can stack multiple interaction blocks to progressively reasoning for the segment boundaries in an accurate direction.

4.4 Model Efficiency Comparison

To further investigate the efficiency of our IA-Net, we conduct the comparison on TACoS dataset with other released methods. All experiments are run on one NVIDIA TITAN-XP GPU. As shown in Table 3, "Run-Time" denotes the average time to localize one sentence in a given video, "Model Size" denotes the size of parameters. It can be observed that our IA-Net achieves the fastest run-time with the relatively smaller model size. Since CTRL and

ACRN need to sample candidate segments with various sliding windows, they need a quite time-consuming matching procedure. 2DTAN adopts a convolution architecture to generate a large 2D temporal map, which contains a large number of parameters across the convolution layers. Compared to them, our IA-Net is lightweight with a few parameters of the linear layers, leading to relatively smaller model size, thus is faster than 2DTAN.

4.5 Ablation Study

We perform extensive ablation studies on the ActivityNet Captions dataset. The results are shown in Table 4.

Effectiveness of each component. To evaluate each component in our improved interaction block, we design four strong baselines: model A only contains the improved inter-attention module for multi-modal interaction; model B adds the calibration module after each inter-attention module of model A; model C adds an improved intra-attention module to model B; model D is the IA-Net, which has additional calibration module after each intra-attention module of model C. For fair comparison, we stack the same number of interaction blocks in all four baselines for evaluation. From the table, we can find that both two calibration modules in each interaction block of IA-Net can bring significant improvement (A→B, C→D) by refining the alignment knowledge. The intra-attention can also improve the performance (B→C) by capturing the contextual information among intra-modality.

How to choose the padding size? We investigate the performance on different padding size K, which is originally introduced to deal with "nowhere-to-attend" problem. Although $K = 1$ has the ability to guide the frame or words to attend nothing, such limited latent space can not meet complicated relations between different videos and sentences. As shown in Table 4, we found that the use of $K > 1$ improves performance to a certain extent, and $K = 3$ yields the best.

How to choose the number of attention maps? Different number of parallel attention maps will
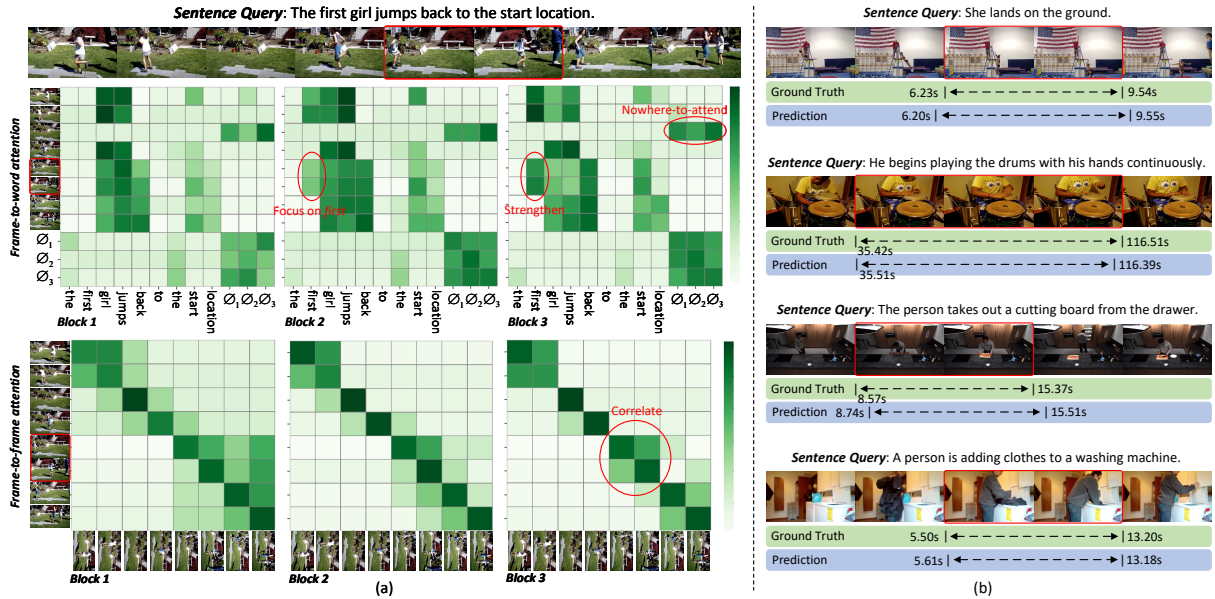
Figure 5: (a) Visualization on the inter- and intra-attention of different interaction blocks. (b) Qualitative results.

Component	Details	R@1, IoU=0.7	R@5, IoU=0.7
Full model	A	21.76	56.60
	B	23.57	58.33
	C	25.62	61.24
	D	**27.95**	**63.12**
Padding size K	0	25.81	59.07
	1	27.03	61.61
	2	27.69	63.07
	3	**27.95**	**63.12**
	4	27.88	62.94
Number (H) of attention maps	1	25.89	61.13
	2	27.14	62.39
	4	**27.95**	**63.12**
	8	27.66	63.05
Number (l) of stacked blocks	1	25.17	59.76
	2	26.88	61.63
	3	27.95	**63.12**
	4	**27.97**	63.03

Table 4: Ablation study on ActivityNet Captions.

guide the model attend to different relationships in several latent spaces. Following (Vaswani et al., 2017), we implement 1, 2, 4, 8 number of attention maps for experiments. We find that $H = 4$ achieves the best performance.

Choice of the number of stacked interaction blocks. Our improved interaction block contains calibration module for refining the alignment knowledge. The results in table indicates that the model with more than 3 blocks will not bring more improvement.

4.6 Qualitative Results

We visualize the fused attention maps in Figure 5 (a). For frame-to-word attention, it can be ob-

served that the first interaction block fail to focus on the word "first". With the help of the calibration module, the attention is progressively calibrated on the word "first" in the following blocks. Besides, the third frame pays more attention on the padding elements as it does not match any words in the query, which indicates that our IA-Net addresses the "nowhere-to-attend" problem well. For frame-to-frame attention, our model correlates the relevant frames more precisely in deeper blocks. The qualitative results are shown in Figure 5 (b).

5 Conclusion

In this paper, we have studied the problem of temporal sentence grounding, and proposed a novel Iterative Alignment Network (IA-Net) in an end-to-end fashion. The core of our network is the multi-step reasoning process with the improved inter- and intra-modal interaction module which is designed in two aspects: 1) we pad the multi-modal features with learnable parameters for capturing more complicated correlation in deep latent space; 2) we develop a calibration module to refine and calibrate the alignments knowledge from early steps. By stacking multiple such interaction modules, our IA-Net can progressively capture the fine-grained interactions between two modalities, providing more accurate video segment boundaries. Extensive experiments on three challenging benchmarks demonstrate the effectiveness and efficiency of the proposed IA-Net.

6 Acknowledgments

References

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 International Conference on Computer Vision (ICCV)*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NIPS)*.

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 (ICCV)*.

Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2020a. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1841–1851.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11235–11244.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020b. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018a. Attentive moment retrieval in videos. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018b. Cross-modal moment localization in videos. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *The IEEE Winter Conference on Applications of Computer Vision (WACV).*

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 (CVPR).*

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV).*

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

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 (ICCV).*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS).*

Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of International Conference on Machine Learning (ICML).*

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019a. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems (NIPS).*

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019b. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the ACM International Conference on Multimedia (ACM MM).*

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).*

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*