Hierarchical Deep Residual Reasoning for Temporal Moment Localization

Ziyang Ma, Xianjing Han, Xuemeng Song*, Yiran Cui, Liqiang Nie
Shandong University, Shandong, China
ziyang.ma@mail.sdu.edu.cn,{hanxianjing2018,sxmustc,conyrol120,nieliqiang}@gmail.com

ABSTRACT

Temporal Moment Localization (TML) in untrimmed videos is a challenging task in the field of multimedia, which aims at localizing the start and end points of the activity in the video, described by a sentence query. Existing methods mainly focus on mining the correlation between video and sentence representations or investigating the fusion manner of the two modalities. These works mainly understand the video and sentence coarsely, ignoring the fact that a sentence can be understood from various semantics, and the dominant words affecting the moment localization in the semantics are the action and object reference. Toward this end, we propose a Hierarchical Deep Residual Reasoning (HDRR) model, which decomposes the video and sentence into multi-level representations with different semantics to achieve a finer-grained localization. Furthermore, considering that videos with different resolution and sentences with different length have different difficulty in understanding, we design the simple yet effective Res-BiGRUs for feature fusion, which is able to grasp the useful information in a self-adapting manner. Extensive experiments conducted on Charades-STA and ActivityNet-Captions datasets demonstrate the superiority of our HDRR model compared with other state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; *Novelty in information retrieval.*

ACM Reference Format:
Ziyang Ma, Xianjing Han, Xuemeng Song*, Yiran Cui, Liqiang Nie. 2021. Hierarchical Deep Residual Reasoning for Temporal Moment Localization. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3469877.3490595

1 INTRODUCTION

The explosive increasing of the video data in our daily life makes the processing and automatic understanding to the video significantly necessary. Due to its huge application requirements in fields of security (e.g. surveillance) and network media (e.g. fragment retrieval), the task of Temporal Moment Localization (TML) [2, 19, 21, 26] that aims at localizing the start and end points of the activity described by the sentence query in an untrimmed video, attracts the attention of plenty of researchers. It is worth noting that compared with traditional video understanding tasks, such as the action detection that detects actions from a limited pre-defined set [27, 37, 44], TML is more flexible, since it supports the natural language based query. Accordingly, the task of TML is more challenging.

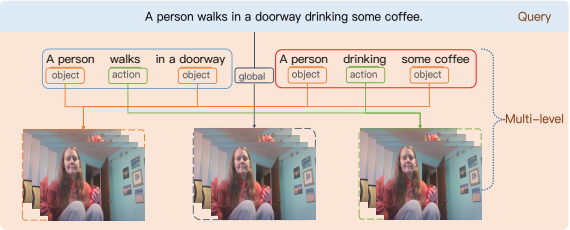

Figure 1: Illustration of the multi-level representations of the sentence and video.

Although existing studies have achieved promising progress, they mainly suffer from the following two limitations: 1) *lack the explicit semantic-oriented representation learning.* Since the representation of the sentence plays a pivotal role in TML, several efforts [5, 8, 12, 17, 36, 43] have investigated the fine-grained (e.g., word-level and phrase-level) representations of the sentence, for localizing the target moments in video. Nevertheless, they neglect the fact that the key words in the sentence query referring to the target moment are usually actions and objects references. In other words, they overlook the potential of the actions and objects in the sentence in localizing the target moment. And 2) *lack the adaptive multi-modal fusion.* As two modalities are involved in the task of TML, one essential problem of TML is how to effectively fuse the representations of the two modalities to facilitate the localization of the target moment. Although existing methods have achieved great success with either the graph neural network [19, 40, 42, 45] or the iterative attention [19, 26, 39], they neglect that the information embedded in different videos and sentences may have different levels of difficulty to be understood. Intuitively, shorter and high-quality videos are easier to be analysed, compared with the longer and low-quality ones. Analogously, the longer and complex sentences are also more difficult to be reasoned compared with the shorter ones. For example, the sentence query "A man begins changing as he goes upstairs" is harder to be understood by the model than "A man is eating". Therefore, it is necessary to devise an adaptive fusion scheme to handle the real-world application scenario, where both simple and complex samples are possible as the input.

*Xuemeng Song is the corresponding author.

MMAsia '21, December 1–3, 2021, Gold Coast, Australia

ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3490595

Toward address the above limitations, as shown in Figure 1, we propose to explore the multi-level representations of the video and sentence, including not only the coarse global representation, but also the fine-grained semantic-oriented (action- and object-oriented) representations. Concretely, we introduce a Hierarchical Deep Residual Reasoning (HDRR) model to tackle the task of TML, as shown in Figure 2, where we represent the video and sentence at different levels to thoroughly capture the correlation between video clips and the sentence query, and design the Res-BiGRUs to boost the adaptive fusion of the two modalities. In particular, we employ the pre-trained BERT to perform Semantic Role Labeling (SRL) to the sentence and get the global, action, and object level representations. To learn the matching degree between the video clips and the sentence query, we devise the Res-BiGRUs to adaptively fuse the two modalities and avoid the information loss caused by the representation smoothing. Based on the fused multi-level representations, we hierarchically localize the video moment according to the sentence query.The main contributions of our work are threefold:

- We propose a Hierarchical Deep Residual Reasoning (HDRR) model, which decomposes the video and sentence into multi-level representations with different semantics to strengthen the temporal moment localization with hierarchical reasoning.
- Considering that different videos and sentences may have different levels of difficulty to be reasoned, we design the Res-BiGRUs for the self-adaptive feature fusion, which also avoids the information loss caused by the representation smoothing as a byproduct.
- Extensive experiments on the Charades-STA and ActivityNet-Captions datasets demonstrate the superiority of our model over other state-of-the-art methods. We will release the code[1] to benefit other researchers.

2 RELATED WORK

2.1 Temporal Moment Localization

The task of TML in untrimmed videos was first introduced by Gao *et al.* [11] and Hendricks *et al.*[1]. Early works on TML mainly focus on mining the correlation between the video clip and sentence representations [1, 11, 20, 21, 39]. For example, Liu *et al.* [20] enhanced the sentence representation by combining contextual video features with the attention mechanism. Although early methods obtain compelling progress, they overlook the fine-grained interaction between the video and the sentence. To address this issue, some later methods [5, 8, 12, 17, 36, 43] shed light on the fine-grained understanding of the two modalities. For example, Ge *et al.* [12] added predefined actions feature in the process of matching sentence and video. One key limitation of these methods is that the predefined actions are limited and can hardly cover all kinds of actions occurred in testing phase. Besides, they ignored the objects information, which also plays a pivotal role in localizing the target moment. Recently, Mun *et al.* [23] considered the semantic entities such as actions and objects in the sentence, but they entangled the interaction between different semantic entities with the video, which may require more search space to capture fine-grained localization. Beyond that, in our work, we propose the multi-grained moment localization, where the video and sentence is represented with multi-levels, i.e., the global level and the fine-grained action/object level.

[1]https://github.com/ddlBoJack/HDRR

2.2 Semantic Role Labeling

Semantic Role Labeling (SRL) raised by Charles J. Fillmore [10] is the task of determining the latent predicate argument structure of a sentence and hence facilitate the question answering, like who did what to whom. Traditional SRL systems [4] are mainly based on the syntactic analysis. However, complete syntactic analysis requires all the syntactic information contained in the sentence, and minor errors in syntactic analysis may lead to severe errors in the sentence reasoning. The block-based SRL approach solves SRL as a sequence tagging problem, which is usually solved with BERT. Simple BERT[28] is an implementation of the BERT based model, which is currently the state-of-the-art single model for English PropBank SRL [24]. Inspired by the success of SRL in sentence structure analysis, in this paper, we employ the BERT-based SRL model [28] to extract the actions and objects in the sentence query and hence facilitate the hierarchical moment localization.

3 METHOD

In this section, we first give the problem formulation, then detailed the proposed Hierarchical Deep Residual Reasoning model.

3.1 Problem Formulation

Formally, suppose we have an untrimmed video V and a sentence query S. Owing to that the target moment usually involves multiple continuous units of the video, we split the untrimmed video and represent it with $V = \{v_t\}_{t=1}^{T}$, where v_t is the t-th unit of the video and T denotes the total number of the units. The sentence query is represented as $S = \{s_l\}_{l=1}^{L}$, where s_l is the l-th word in the sentence and L is the length of the sentence. Overall, the task is to determine the start and end points (ξ^s, ξ^e) of the target moment in the video corresponding to the sentence query.

3.2 Hierarchical Deep Residual Reasoning

3.2.1 Multi-level Sentence Encoder. We first present the multi-level sentence encoder used for representing the sentence query, from not only the coarse-grained global level, but also fine-grained action and object levels.

Global Level Feature Extraction. Regarding the global-level encoding of the sentence, similar with previous studies [19, 21, 26, 39], we use GloVe [25] to obtain the word embedding $\mathbf{s}_l$ for each word in the sentence. Then we use BiGRU to encode the the contextual information among the words as follows,

$$\begin{cases} \overrightarrow{\mathbf{h}}_l^s = \overrightarrow{GRU}^s\left(\mathbf{s}_l, \overrightarrow{\mathbf{h}}_{l-1}^s\right), \\ \overleftarrow{\mathbf{h}}_l^s = \overleftarrow{GRU}^s\left(\mathbf{s}_l, \overleftarrow{\mathbf{h}}_{l+1}^s\right), \\ \mathbf{s}_l^g = f^s\left(\overrightarrow{\mathbf{h}}_l^s \| \overleftarrow{\mathbf{h}}_l^s\right), \end{cases} \tag{1}$$

where $\overrightarrow{\mathbf{h}}_l^s$ and $\overleftarrow{\mathbf{h}}_l^s$ denote the l-th hidden state vectors generated by the forward and the backward GRU, respectively. $\|$ denotes the

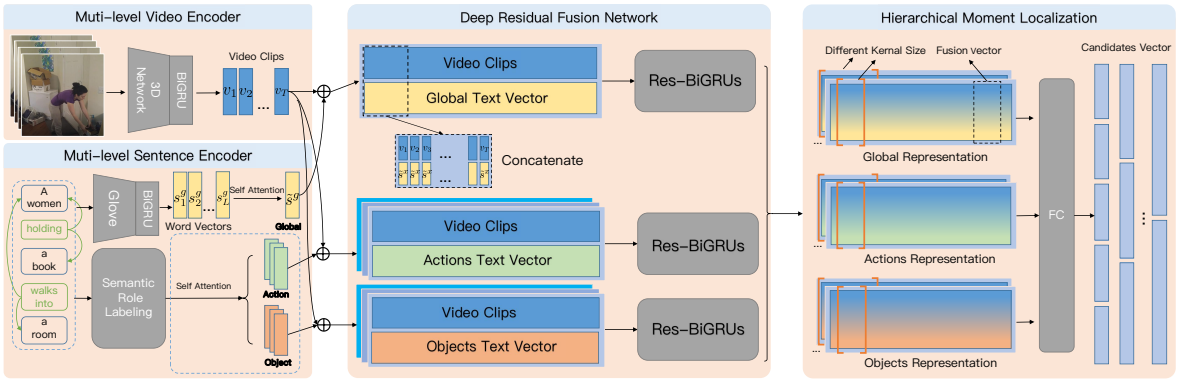

Figure 2: Pipeline of the proposed Hierarchical Deep Residual Reasoning. Firstly, the video and the sentence are encoded in multi-level layers, then multi-layer matching is carried out, and our Res-BiGRU module is used for self-adapting fusion. Finally, our hierarchical temporal localization is obtained through hierarchical convolution.

concatenation operation and f^s denotes the fully-connected layer followed with ReLU activation function to fuse the hidden states for each word derived by the two GRUs. $\mathbf{s}_l^g \in \mathbb{R}^{d_s}$ refers to the latent representation of the g-th word, and d_s is the dimension of the representation. Finally, the global representation of the sentence is denoted as $\mathbf{S}^g = [\mathbf{s}_1^g, \mathbf{s}_2^g, \cdots, \mathbf{s}_L^g], \mathbf{S}^g \in \mathbb{R}^{L \times d_s}$.

Semantic Level Feature Extraction. To explore the fine-grained information of sentences, similar to [9], we use the pre-trained SRL BERT [28] to obtain the semantic role of each word in the sentence, which first decomposes the sentence into multiple frames and then analyses the semantic role of words in each frame. For example, with the SRL BERT, the sentence "A woman holding a book walks into a room" can yield two frames: "a woman (object) / holding (action) / a book (object)" and "a woman (object) / walks into (action) / a room (object)", where the objects "a woman" and "a book" are correlated by the action "holding", and the objects "a woman" and "a room" are associated with the action "walks into", respectively.

Formally, we use the SRL BERT to get the semantic role for the given sentence S and derive the object and action level representations denoted as $\mathbf{S}^o = \{\mathbf{s}_1^o, \mathbf{s}_2^o, \cdots, \mathbf{s}_{L_o}^o\}$ and $\mathbf{S}^a = \{\mathbf{s}_1^a, \mathbf{s}_2^a, \cdots, \mathbf{s}_{L_a}^a\}$, where L_a and L_o are the total number of the words with action and object role, respectively. Note that to ease the calculation, we pad the non-action and non-object words in the sentence with 0s to regularize the dimension of $\mathbf{S}^o$ and $\mathbf{S}^a$ to $\mathbb{R}^{L \times d_s}$, respectively.

Attentive Representation Learning. To highlight the prominent information of the representation at each level, we further employ the multi-head self-attention [31] that has turned out to be effective in capturing the the useful contextual information, to derive the final sentence representations as follows,

$$\begin{cases} \mathrm{h}_i = \mathrm{Softmax}\left(\dfrac{\mathbf{S}^x \mathbf{W}_i^{Q,x} \left(\mathbf{S}^x \mathbf{W}_i^{K,x}\right)^\top}{\sqrt{d_h}}\right) \mathbf{S}^x \mathbf{W}_i^{V,x}, \\ \tilde{\mathbf{s}}^x = \mathbf{W}_O^x \mathrm{Concat}\left(h_1, \ldots, \mathrm{h}_n\right), \end{cases} \tag{2}$$

where $\mathbf{S}^x \in \mathbb{R}^{L \times d_s}, x \in \{g, a, o\}$ represents the representations of the sentence at different levels, including the global, and semantic levels. $\mathbf{W}_i^{Q,x}, \mathbf{W}_i^{K,x}, \mathbf{W}_i^{V,x} \in \mathbb{R}^{d_s \times d_h}$ are the learnable parameters of the i-th head. $d_s = d_h \times n$, where d_h is the size of the output feature for each head and n is the number of the parallel heads. The output of each head h_i is concatenated and projected by a linear transformation $\mathbf{W}_O^x \in \mathbb{R}^{1 \times L}$ to generate the final different level sentence representations $\tilde{\mathbf{s}}^x \in \mathbb{R}^{1 \times d_s}, x \in \{g, a, o\}$.

3.2.2 Multi-level Video Encoder. To facilitate the correspondence modeling between the video clip and the sentence, we also characterize the video at different levels, i.e., the global, action, and object levels.

Global Level Encoding. We first use the pre-trained 3D network [3, 30, 33] to extract the visual feature for the video, denoted as $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_T]$, where $\mathbf{v}_t$ is the visual feature of the t-th unit. Similar with the global level encoding of the sentence, we also use BiGRU to excavate the contextual information among the video unit sequence, which can be formulated as follows,

$$\mathbf{V}^g = BiGRU_v(\mathbf{V}) \tag{3}$$

where $BiGRU$ denotes the BiGRU network following the same architecture of Equation 1. $\mathbf{V}^g = [\mathbf{v}_1^g, \mathbf{v}_2^g, \cdots, \mathbf{v}_T^g]$ represents the obtained global level representation of the video, where $\mathbf{v}_t^g$ stands for the global level representation of the t-th video unit.

Semantic Level Encoding. Apparently, the semantic level representations of video can be hardly explicitly extracted like that of the sentence. Thus, we employ two fully-connected layers to highlight the action and object information in the video, respectively. Formally, we have,

$$\mathbf{v}_t^x = f^x(\mathbf{v}_t^g), x \in \{a, o\} \tag{4}$$

where $f^x, x \in \{a, o\}$ is a fully-connected layer, in which a represents for "action" and o represents for "object". $\mathbf{v}_t^x, x \in \{a, o\}$ refers to the action- and object-oriented representations of the t-th video unit. Ultimately, the action- and object-oriented representations of the whole video can be denoted as $\mathbf{V}^a = [\mathbf{v}_1^a, \mathbf{v}_2^a, \cdots, \mathbf{v}_T^a]$ and $\mathbf{V}^o = [\mathbf{v}_1^o, \mathbf{v}_2^o, \cdots, \mathbf{v}_T^o]$, respectively.

3.2.3 Deep Residual Fusion Network. In order to adaptively fuse the representations of the video and sentence at each level, we devise the Res-BiGRUs based fusion network, which is the stacking of multiple BiGRU layer, as shown in Figure 3. The deep structure of the Res-BiGRUs ensures the sufficient exploitation to the difficult samples, while the residual design is employed to automatically maintain the characteristics among the connected BiGRU layers and thus avoids the representation smoothness caused by the deep

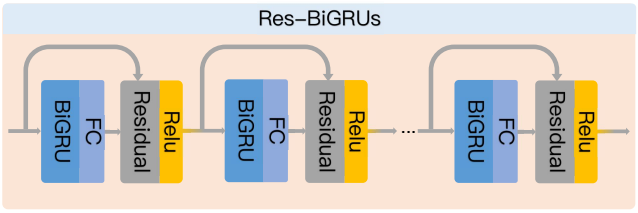

Figure 3: The illustration of Res-BiGRUs.

structure [16]. In particular, we fuse each level representation of the video and sentence, i.e., $\mathbf{V}^x = [\mathbf{v}_1^x, \mathbf{v}_2^x, \cdots, \mathbf{v}_T^x]$ and $\tilde{\mathbf{s}}^x$, $x \in \{g, a, o\}$, and feed them into the Res-BiGRUs as follows,

$$\hat{\mathbf{F}}^x = \text{Res-BiGRUs}^x(\mathbf{F}_0^x):$$
$$\begin{cases} \mathbf{F}_0^x &= [\mathbf{v}_1^x \| \tilde{\mathbf{s}}^x, \mathbf{v}_2^x \| \tilde{\mathbf{s}}^x, \cdots, \mathbf{v}_T^x \| \tilde{\mathbf{s}}^x] \\ \mathbf{H}_m^x &= f_m^x(BiGRU_m^x(\mathbf{F}_{m-1}^x)), \\ \mathbf{F}_m^x &= ReLU(\mathbf{H}_m^x + \mathbf{F}_{m-1}^x), m = 1, 2, \cdots, M \end{cases} \tag{5}$$

where Res-BiGRUs^x denotes the Res-BiGRUs for the x level representation. $BiGRU_m^x$ and f_m^x, $x \in \{g, a, o\}, m \in \{1, 2, \cdots, M\}$ stand for the m-th BiGRU layer and the m-th fully-connected layer in the Res-BiGRUs of the fusion network $\mathcal{F}^x$, respectively. $\mathbf{H}_m^x$ and $\mathbf{F}_m^x$ denote the hidden representation and output of the m-th BiGRU layer. Finally, we obtain the fused representation of the video and sentence as $\hat{\mathbf{F}}^x = \mathbf{F}_M^x \in \mathbb{R}^{L \times d_f}$, where d_f is the dimension of the fused representation.

3.2.4 Hierarchical Moment Localization. We conduct the hierarchical moment localization based on the multi-level fused video and sentence representation $\hat{\mathbf{F}}^x, x \in \{g, a, o\}$. In particular, for each level, we first localize the coarse moment candidates of various lengths by the convolutional neural network. Then we refine the boundary of the moment candidates by predict their offsets to adjust the boundary.

Candidate Ranking. Following previous studies [1, 5, 8, 11, 12, 20, 36, 40], we employ the 1D-Convolutional neural network, which is able to jointly generate a series of fixed-length (depending on the filter size) moment candidates as well as their ranking scores with the 1-D convolution operation. In particular, we introduce 3 1D-CNN based ranking networks to derive the moment candidate list based on their global, action, and object level representations, respectively. Formally, we have

$$\{(t_k^{x,s}, t_k^{x,e}, r_k^x)\}_{k=1}^K = sigmoid(Rank^x(\hat{\mathbf{F}}^x)), \tag{6}$$

where $Rank^x, x \in \{g, a, o\}$ is the 1D-CNN based ranking network. For all the three ranking networks, we use Q filters with the same set of sizes $\{w_1, w_2, \cdots, w_Q\}$, where w_q is the size the q-th filter, based on which we can obtain a set of moment candidates of the size w_q. Since the Q filters of all the ranking networks share the same set of sizes, the set of moments candidates of different lengths, denoted by the start and end points $(t_k^{x,s}, t_k^{x,e})$'s, derived from the representations of three different levels are the same. Namely, $(t_k^{g,s}, t_k^{g,e}) = (t_k^{a,s}, t_k^{a,e}) = (t_k^{o,s}, t_k^{o,e})$. For simplicity, we use (t_k^s, t_k^e) to denote the unified start and end points of k-th moment candidate. r_k^x is the corresponding ranking score of the moment candidate (t_k^s, t_k^e) at the x level.

Ultimately, we learn the final ranking score r_k of the k-th moment candidate by multiple levels as follows,

$$r_k = f_h(r_k^g \| r_k^a \| r_k^o), \tag{7}$$

where f_h is a fully-connected layer.

Boundary Regression. To refine the moment localization, we further predict the offsets of candidate moments to adjust their boundaries. Similar with the ranking network, we also employ the 1D-Convolutional Neural Nework to predict the offsets for the candidate moments. It is worth noting that we only use the global representation to adjust the boundary, since it is more comprehensive compared with other representations.

$$\begin{cases} \{d_k^s\}_{k=1}^K = Offset^s(\hat{\mathbf{F}}^g), \\ \{d_k^e\}_{k=1}^K = Offset^e(\hat{\mathbf{F}}^g), \end{cases} \tag{8}$$

where $Offest^s$ and $Offest^e$ are the offset determination network regarding the start point and end point of the moment, respectively. In particular, both $Offest^s$ and $Offest^e$ are implemented with the 1D-convolutional neural network. d_k^s and d_k^e refer to the start and end offset of the k-th moment candidate, respectively.

Ultimately, the predicted refined boundary $(\hat{\xi}_k^s, \hat{\xi}_k^e)$ of the k-th moment candidate can be represented as follows,

$$\begin{cases} \hat{\xi}_k^s = t_k^s + d_k^s, \\ \hat{\xi}_k^s = t_k^e + d_k^e. \end{cases} \tag{9}$$

3.2.5 Training. As for the optimization, we employ the alignment loss and regression loss to encourage the precious moment localization of the model.

Alignment Loss. Similar with [5, 40], we use alignment loss to promote the model to assign a high ranking score to the candidate moment that has a large overlap with the ground truth. In particular, we use Intersection over Union (IoU) to represent the alignment degree between the candidate moment (t_k^s, t_k^e) and the ground truth moment (ξ_k^s, ξ_k^e), which can be denoted as IoU_k. The alignment loss $\mathcal{L}_{aln}$ is formulated as,

$$\mathcal{L}_{aln} = -\frac{1}{K}\sum_{k=1}^{K} IoU_k \log(r_k) + (1 - IoU_k)\log(1 - r_k), \tag{10}$$

where K is the total number of the candidate moments.

Regression Loss. To boost the boundary adjustment, we use the regression loss $\mathcal{L}_{reg}$ to promote the generated candidate moments close to the ground truth, which is formulated as follows,

$$\mathcal{L}_{reg} = smooth_{L_1}(\xi^s - \hat{\xi}_k^s) + smooth_{L_1}(\xi^e - \hat{\xi}_k^e), \tag{11}$$

where $smooth_{L_1}$ denotes Smooth L_1 loss function. Pay attention that we only calculate the loss of the candidate moment $(\hat{\xi}_k^s, \hat{\xi}_k^e)$ with the specific k, which has the highest IoU with the ground truth moment (ξ^s, ξ^e). Here we obtain the final loss function, which is defined as follows,

$$\mathcal{L} = \mathcal{L}_{aln} + \alpha\mathcal{L}_{reg}, \tag{12}$$

where α is a trade-off hyper-parameter.

Table 1: Performance comparison on Charades-STA dataset in terms of R@1, IoU@0.3, R@1, IoU@0.5 and R@1, IoU@0.7. "-" indicates that the corresponding result is unavailable.

		Charades-STA		
Feature	Method	R@1, IoU@0.3	R@1, IoU@0.5	R@1, IoU@0.7
C3D	MCN	-	17.46	8.01
	CTRL	-	23.63	8.89
	ABLR	-	24.36	9.01
	SM-RL	-	24.36	11.17
	SAP	-	27.42	13.36
	ACL	-	29.39	12.23
	QSPN	54.70	35.60	15.80
	DEBUG	54.95	37.39	17.69
	RWM	-	36.70	13.74
	CBP	54.30	36.80	18.87
	GDP	54.54	39.47	18.49
	TripNet	-	38.29	16.07
	TSP-PRL	-	37.39	17.69
	PMI	55.48	39.73	19.27
	HDRR	**62.37**	**43.04**	**21.32**
Twostream	RWM	-	37.23	17.72
	TSP-PRL	-	45.30	24.73
	HDRR	**68.33**	**54.06**	**27.31**
I3D	MAN	-	46.63	22.72
	ExCL	65.10	44.10	23.30
	SCDM	-	54.44	33.43
	LGI	72.96	59.46	**35.48**
	HDRR	**73.44**	**59.46**	34.11

Table 2: Performance comparison on ActivityNet-Captions dataset in terms of R@1, IoU@0.3 and R@1, IoU@0.5. "-" indicates that the corresponding results are not available.

		ActivityNet-Captions	
Feature	Method	R@1, IoU@0.3	R@1, IoU@0.5
C3D	MCN	21.37	9.58
	CTRL	28.70	14.00
	ACRN	31.29	16.17
	TGN	43.81	27.93
	QSPN	45.30	27.70
	TripNet	48.42	32.19
	ABLR	55.67	36.79
	RWM	53.00	36.90
	CBP	54.30	35.76
	SCDM	54.80	36.75
	GDP	56.17	39.27
	DEBUG	55.91	39.72
	TSP-PRL	56.08	38.76
	LGI	58.52	41.51
	2D-TAN	59.45	**44.51**
	PMI	59.69	38.28
	HDRR	**61.10**	43.20

4 EXPERIMENTS

4.1 Datasets

To evaluate our proposed HDRR model, we conducted experiments on two benchmark datasets.

Charades-STA [11]: This dataset is built based on the Charades dataset [29], which contains 6,672 videos of indoor activities and 16,128 query-video pairs. There are 12,408 pairs for training and 3,720 for testing. The average duration of each video is 29.76 seconds. Each video has 2.4 annotated moments and each annotated moment lasts for 8 seconds on average.

ActivityNet-Captions [18]: This dataset contains 20K videos with 100K queries, where 37,421 query-video pairs are used for training and 34,536 for testing. The average duration of the videos is 110 seconds. On average, each video has 3.65 annotated moments and each annotated moment lasts for 36 seconds.

4.2 Implementation Details

Evaluation Metric. Similar with [11], we adopted the metric "R@m, IoU@n", which represents the proportion of the top m moment candidates with IoU larger than n, for evaluation. Following the mainstream test setting, we set m as 1 for both datasets, while n as 0.3, 0.5, 0.7 for Charades-STA dataset, and 0.5, 0.7 for ActivityNet-Captions dataset.

Video Feature. As for the Charades-STA dataset, we used I3D network [3], C3D network [30], and Twostream network [33] for the feature extraction. The extracted dimensions are 1,024, 4,096, and 8,192 respectively. As for the ActivityNet-Captions dataset, we used the widely used features of 500-D extracted by C3D network. As the average duration of the video in the ActivityNet-Captions dataset is longer than that in the Charades-STA dataset, we split each video in the Charades-STA and ActivityNet-Captions datasets into 75 and 200 units, respectively.

Text Feature. For the sentence presentation, we obtained the word embedding of 300-D by the pre-trained GloVe [25]. We used the pre-trained BERT [28] to conduct the semantic role labeling. The maximum length of the sentence in Charades-STA and ActivityNet-Captions datasets are set as 10 and 50, respectively.

Training settings. A single NVIDIA Titan XP and a single NVIDIA RTX 2080Ti are used to train our model. We used Adam optimizer and set the learning rate as 0.003 for Charades-STA dataset and 0.0003 for ActivityNet-Captions dataset. The batch size is set to 128 in both datasets. The depth M of the Res-BiGRUs is set as 3. Similar with the existing work [26], the filter size in the 1D-convolutional neural network (i.e., $Rank^x$, $Offset^s$, and $Offset^e$) is set as [6, 12, 24, 48, 72] for Charades-STA dataset and [16, 32, 64, 96, 128, 160, 192] for ActivityNet-Captions dataset. The trade-off parameter α used for balancing the two losses is set as 0.001.

4.3 Performance Comparison

We compared our HDRR with the following state-of-the-art methods, including proposal-based methods, proposal-free methods, and reinforcement-learning-based methods. For proposal-based methods that generate dense proposals to localize the moment and perform location regression to adjust the boundary, we adopt CTRL [11], MCN [1], ACRN [20], TGN [5], SAP [8], QSPN [36], MAN [40], ACL [12], as well as the latest work CBP [32], 2D-TAN[41], SCDM [38], and PMI [7] as the baselines. Regarding the proposal-free methods that predict the results directly without the candidate boxes, we choose ABLR [39], ExCL [13], DEBUG [22], as well as the latest work GDP [6], and LGI [23] as the baselines. Pertaining to the Reinforcement-learning based methods, where the reinforcement learning is used, we select RWM [15], SM-RL [34], TripNet [14], as well as some latest work TSP-PRL [35] as the baseline.

Table 3: Results of the ablation studies on the Charades-STA dataset based on the I3D features.

Method	R@1, IoU@0.3	R@1, IoU@0.5	R@1, IoU@0.7
HDRR-w/o-Action	70.00	57.23	31.77
HDRR-w/o-Object	71.99	58.80	31.72
HDRR-w/o-Action&Object	69.76	55.27	30.13
HDRR-w/o-Res-BiGRUs	68.60	51.34	25.97
HDRR-w/o-self-attention	72.45	58.33	32.77
HDRR	**73.44**	**59.46**	**34.11**

Tables 1 and 2 show the performance comparison among our HDRR and the state-of-the-art methods on two datasets. The experiment results of the baseline methods are referred by their papers. From these two Tables, we have the following observations: 1) Our HDRR has superiority over other methods among most scenarios, demonstrating the effectiveness and robustness of our model. 2) Our HDRR largely surpasses the baseline methods on the Charades-STA dataset with the two-stream feature. For instance, HDRR surpasses TSP-PRL by the margin of 8.76% with respect to R@1, IoU@0.5. The possible reason may be that the two-stream feature is more powerful in capturing the actions and objects information of the video, which hence boosts the multi-level video representation learning in Section 3.2.2 and thus improves the localization performance.

4.4 Ablation Study

We conducted the ablation study to demonstrate the effects of different components of our HDRR with the following derivations:

- **HDRR-w/o-Action**: We removed the action level representation of the video and sentence to verify the effect of the action level representation.
- **HDRR-w/o-Object**: We removed the object level representation of the video and sentence to investigate the effect of the object level representation.
- **HDRR-w/o-Action&Object**: We removed both the action and object level representations of video and sentence.
- **HDRR-w/o-Res-BiGRUs**: To evaluate the effectiveness of the proposed Res-BiGRU module, we replaced the Res-BiGRUs with a fully-connected layer in HDRR.
- **HDRR-w/o-Residual**: We removed the residual connections in Res-BiGRUs to test the effect of the residual incorporation.

Table 3 presents the results of the ablation study, from which we can draw the following conclusions: 1) Our HDRR surpasses HDRR-w/o-Action, HDRR-w/o-Object, and HDRR-w/o-Action&Object, which indicates that both of the action and object level representations are able to provide useful information to boost the moment localization. 2) Our HDRR achieves better performance than HDRR-w/o-Res-BiGRUs, demonstrating the effectiveness of the designed Res-BiGRUs. The deep and residual structure of the Res-BiGRUs is able to facilitate the adaptive integration between the video and sentence, and enhance the moment localization. And 3) our HDRR surpasses HDRR-w/o-self-attention, demonstrating the effectiveness of the multi-head self-attention in capturing the prominent information of the sentence and enhancing the understanding of it.

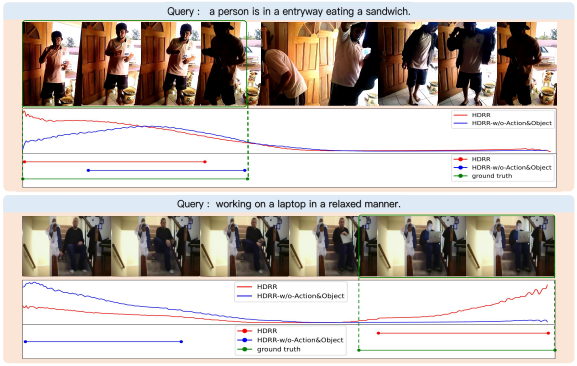

Figure 4: Visualization of the ranking score distributions as well as the localization results of the query-video pairs.

4.5 Result Visualization

To get an intuitive understanding of our model, we illustrated the ranking score distributions and the final localization of the query-video pairs, with two examples of HDRR and HDRR-w/o-Action&Object in Figure 4. In the first example, with the enhancement to the representation of the action "eating" and the object "sandwich", our HDRR pays more attention to the beginning of the video and localizes the moment closer to the ground truth. In the second example, we found that both the ranking score distribution of the beginning and the end candidates are higher, which may caused by the content similarity of the beginning and the end of the video. Owing to the attention on the object information (i.e., "laptop"), our HDRR achieves more reasonable localization results than HDRR-w/o-Action&Object. These two examples demonstrate the effectiveness of the hierarchical moment localization based on the multi-level representations of our HDRR.

5 CONCLUSION

In this work, we propose a Hierarchical Deep Residual Reasoning (HDRR) model to solve the problem of temporal moment localization, where the hierarchical matching is conducted on the video and sentence based on their multi-level representations. In addition, we design the simple yet effective Res-BiGRUs for feature fusion, which is able to adaptively grasp useful information in exploring deeper understanding of the two modalities, alleviating the problem of model design caused by different difficulty of data understanding. Through the methods above, the two modalities achieve finer-grained exploration and interaction. Extensive experiments conducted on the Charades-STA and ActivityNet-Captions datasets demonstrate the promising performance of our HDRR. In the future, we will dedicate to the investigation of more appropriate action and object extraction manner in the video, to match the multi-level text representations.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:U1936203; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; CCF-Baidu Open Fund, No.: CCF-BAIDU OF2020019; Young creative team in universities of Shandong Province, No.:2020KJN012.

REFERENCES

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

[2] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. 2020. STRONG: Spatio-temporal Reinforcement Learning for Cross-modal Video Moment Localization. In *ACM International Conference on Multimedia*. 4162–4170.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and The Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[4] Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Conference on Computational Natural Language Learning*. 152–164.

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally Grounding Natural Sentence in Video. In *Conference on Empirical Methods in Natural Language Processing*. 162–171.

[6] 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 *AAAI Conference on Artificial Intelligence*. 10551–10558.

[7] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yugang Jiang. 2020. Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos. In *European Conference on Computer Vision*. 333–351.

[8] Shaoxiang Chen and Yugang Jiang. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. In *AAAI Conference on Artificial Intelligence*. 8199–8206.

[9] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10638–10647.

[10] Charles J Fillmore et al. 1976. Frame Semantics and the Nature of Language. In *Conference on the Origin and Development of Language and Speech*. 20–32.

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

[12] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision*. 245–253.

[13] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Conference of the North American Chapter of the Association for Computational Linguistics*. 1984–1990.

[14] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. 2019. Tripping Through Time: Efficient Localization of Activities in Videos. *arXiv preprint arXiv:1904.09936* (2019).

[15] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos. In *AAAI Conference on Artificial Intelligence*. 8393–8400.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal Video Moment Retrieval with Spatial and Language-Temporal Attention. In *International Conference on Multimedia Retrieval*. 217–225.

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *IEEE International Conference on Computer Vision*. 706–715.

[19] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *ACM International Conference on Multimedia*. 4070–4078.

[20] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 15–24.

[21] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *ACM International Conference on Multimedia*. 843–851.

[22] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In *Conference on Empirical Methods in Natural Language Processing*. 5147–5156.

[23] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10810–10819.

[24] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational linguistics* (2005), 71–106.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[26] 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 *ACM International Conference on Multimedia*. 4280–4288.

[27] Alexander Richard and Juergen Gall. 2016. Temporal Action Detection Using a Statistical Language Model. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3131–3140.

[28] Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv preprint arXiv:1904.05255* (2019).

[29] 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 *European Conference on Computer Vision*. 510–526.

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

[31] 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*. 5998–6008.

[32] Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI Conference on Artificial Intelligence*. 12168–12175.

[33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[34] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *IEEE Conference on Computer Vision and Pattern Recognition*. 334–343.

[35] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video. In *AAAI Conference on Artificial Intelligence*. 12386–12393.

[36] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *AAAI Conference on Artificial Intelligence*. 9062–9069.

[37] Xitong Yang, Xiaodong Yang, Mingyu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. 2019. STEP: Spatio-temporal Progressive Learning for Video Action Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 264–272.

[38] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *International Conference on Neural Information Processing Systems*. 536–546.

[39] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression. In *AAAI Conference on Artificial Intelligence*. 9159–9166.

[40] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1247–1257.

[41] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI Conference on Artificial Intelligence*. 12870–12877.

[42] Zongmeng Zhang, Xianjing Han, Xuemeng Song, Yan Yan, and Liqiang Nie. 2021. Multi-Modal Interaction Graph Convolutional Network for Temporal Language Localization in Videos. *IEEE Transactions on Image Processing* (2021).

[43] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal Interaction Networks for Query-based Moment Retrieval in Videos. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 655–664.

[44] Jiaxing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. 2018. Step-by-step Erasion, One-by-one Collection: A Weakly Supervised Temporal Action Detector. In *ACM International Conference on Multimedia*. 35–44.

[45] Yuan Zhou, Mingfei Wang, Ruolin Wang, and Shuwei Huo. 2020. Graph Neural Network for Video-Query based Video Moment Retrieval. *arXiv preprint arXiv:2007.09877* (2020).