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, xianjing.han, xuemeng.song, yiran.cui, liqiang.nie@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, with different length having different difficulty in understanding.

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

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.

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.
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\footnote{https://github.com/ddlBoJack/HDRR} 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.

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 \(S = \{s_j\}_{j=1}^{L}\), where \(s_j\) is the \(j\)-th word in the sentence and \(L\) is the length of the sentence. Overall, the task is to determine the start and end points \((\xi^g, \xi^f)\) 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 \(s_j\) for each word in the sentence. Then we use BiGRU to encode the the contextual information among the words as follows,

\[
\begin{align*}
\overline{h}_j^g &= \overline{\text{GRU}}^g\{s_j, \overline{h}_{j-1}^g\}, \\
\overline{h}_j^h &= \overline{\text{GRU}}^h\{s_j, \overline{h}_{j+1}^h\}, \\
\overline{s}_j^g &= f^g(\overline{h}_j^g || \overline{h}_j^h),
\end{align*}
\]

where \(\overline{h}_j^g\) and \(\overline{h}_j^h\) denote the \(l\)-th hidden state vectors generated by the forward and the backward GRU, respectively. || denotes the
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. $s^f_1 \in \mathbb{R}^{d_f}$ refers to the latent representation of the $g$-th word, and $d_i$ is the dimension of the representation. Finally, the global representation of the sentence is denoted as $S^g = [s^g_1, s^g_2, \ldots, s^g_L]$, $S^a \in \mathbb{R}^{L \times d_a}$.

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 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 $S^o = \{o_1, o_2, \ldots, o_L\}$ and $S^a = \{a_1, a_2, \ldots, a_L\}$, where $L_o$ and $L_a$ 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 $S^o$ and $S^a$ to $\mathbb{R}^{L \times d}$, 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 useful contextual information, to derive the final sentence representations as follows:

$$
\begin{align*}
\hat{h}_i &= \text{Softmax} \left( S^o W^Q \cdot S^a W^K \cdot \frac{1}{\sqrt{d_i}} \right) S^a W^K \cdot \chi, \\
\hat{s}^g &= W^g \text{Concat}(h_1, \ldots, h_n),
\end{align*}
$$

where $S^g \in \mathbb{R}^{L \times d_a}$, $x \in \{g, a, o\}$ represents the representations of the sentence at different levels, including the global, and semantic levels. $W^Q \in \mathbb{R}^{d_a \times d_a}$, $W^K \in \mathbb{R}^{d_a \times d_a}$, $W^K \in \mathbb{R}^{d_a \times d_a}$ are the learnable parameters of the $i$-th head, $d_i = d_a \times n$, where $d_a$ 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 $W^g \in \mathbb{R}^{d_a \times d}$ to generate the final different level sentence representations $\hat{s}^g \in \mathbb{R}^{L \times d_a}$, $x \in \{g, a, o\}$.

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 $V = [v_1, v_2, \ldots, v_T]$, where $v_t$ 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,

$$
V^g = \text{BiGRU}_G(V)
$$

where $BiGRU$ denotes the BiGRU network following the same architecture of Equation 1. $V^g = [v^g_1, v^g_2, \ldots, v^g_T]$ represents the obtained global level representation of the video, where $v^g_t$ 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,

$$
v^x_t = f^x(v^g_t), x \in \{a, o\}
$$

where $f^x$, $x \in \{a, o\}$ is a fully-connected layer, in which $a$ represents for “action” and $o$ represents for “object”. $v^x_t, x \in \{a, o\}$ 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 $V^x = [v^x_1, v^x_2, \ldots, v^x_T]$ and $V^o = [v^o_1, v^o_2, \ldots, v^o_T]$, 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
which is able to jointly generate a series of fixed-length (depending on the filter size) moment candidates of different sizes, the set of moments candidates of different lengths by the convolutional neural network. Then we refine the boundary adjustment, we use the offset determination network to adjust the boundary, since it is more comprehensive compared with other representations.

### 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 Network 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.

$$L_{\text{reg}} = \text{smooth}_{L1}\left(\hat{\xi}_K^s - \tilde{\xi}_K^s\right) + \text{smooth}_{L1}\left(\hat{\xi}_K^e - \tilde{\xi}_K^e\right)$$

where smooth$_{L1}$ denotes the smooth $L1$ loss function. Pay attention to promote the generated candidate moments close to the ground truth, which is formulated as follows:

$$L_{\text{reg}} = \alpha L_{\text{reg}}$$

where $\alpha$ 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.

| 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        |
|           | ABLR       | -            | 29.39        | 12.23        |
|           | QSPN       | 54.70        | 35.60        | 15.80        |
|           | DEBUG      | 54.95        | 37.39        | 17.69        |
|           | CBP        | 54.30        | 36.70        | 13.74        |
|           | SCDM       | 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        |

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@n, 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 [5], 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 30, 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., RankK, Off sets, and Off sets+) 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 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.

| 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        |
|           | HDPR       | 61.10        | 43.20        |

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.

| 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        |
|           | HDPR       | 61.10        | 43.20        |
Table 3: Results of the ablation studies on the Charades-STA dataset based on the 3D 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.

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.
