Multi-Scale Self-Contrastive Learning with Hard Negative Mining for Weakly-Supervised Query-based Video Grounding

Shentong Mo  
Carnegie Mellon University  
Pittsburgh, United States  
shentonm@andrew.cmu.edu

Daizong Liu  
Peking University  
Beijing, China  
dzliu@stu.pku.edu.cn

Wei Hu*  
Peking University  
Beijing, China  
forhuwei@pku.edu.cn

Abstract

Query-based video grounding is an important yet challenging task in video understanding, which aims to localize the target segment in an untrimmed video according to a sentence query. Most previous works achieve significant progress by addressing this task in a fully-supervised manner with segment-level labels, which require high labeling cost. Although some recent efforts develop weakly-supervised methods that only need the video-level knowledge, they generally match multiple pre-defined segment proposals with query and select the best one, which lacks fine-grained frame-level details for distinguishing frames with high repeatability and similarity within the entire video. To alleviate the above limitations, we propose a self-contrastive learning framework to address the query-based video grounding task under a weakly-supervised setting. Firstly, instead of utilizing redundant segment proposals, we propose a new grounding scheme that learns frame-wise matching scores referring to the query semantic to predict the possible foreground frames by only using the video-level annotations. Secondly, since some predicted frames (i.e., boundary frames) are relatively coarse and exhibit similar appearance to their adjacent frames, we propose a coarse-to-fine contrastive learning paradigm to learn more discriminative frame-wise representations for distinguishing the false positive frames. In particular, we iteratively explore multi-scale hard negative samples that are close to positive samples in the representation space for distinguishing fine-grained frame-wise details, thus enforcing more accurate segment grounding. Extensive experiments on two challenging benchmarks demonstrate the superiority of our proposed method compared with the state-of-the-art methods.

1. Introduction

Query-based video grounding has attracted increasing attention due to its wide spectrum of applications in video understanding [5, 11, 24, 30]. This task aims to determine the start and end timestamps of a target segment in an untrimmed video that contains an activity semantically corresponding to a given sentence description, as shown in Figure 1. Most previous works [6, 10, 19, 20, 36, 39] have achieved significant performance by addressing query-based video grounding in a fully-supervised manner, which however requires a large amount of segment-level annotations (location of the target segment in the video according to the semantic of the matched query). Such manual annotation is quite labor-intensive and time-consuming, thus limiting the wide applicability of query-based video grounding.

Recently, some weakly-supervised works [8, 14, 25, 26, 47] have been proposed to alleviate the above issue by only leveraging the video-level knowledge of matched video-query pairs without detailed segment labels. These methods generally pre-define multiple segment proposals, and employ video-level annotations as supervision to learn the segment-query matching scores for selecting the best one. However, the generated segment proposals are redundant and contain many negative (i.e., false) samples, resulting in inferior effectiveness and efficiency of the models. Further, as for the positive (i.e., correct) proposals covering the accurate foreground frames, they are of high similarity [41] and require more sophisticated intra-modal recognition capabilities to distinguish. Especially for the boundary frames which exhibit similar visual appearance to the foreground frames in a certain segment, some of them are background frames that are hard to recognize. Once such segment is selected as the best one, the grounding performance will be degenerated due to the background noise.

To this end, we propose a novel Multi-scale Self-Contrastive Learning (MSCL) paradigm with hard negative mining strategy for weakly-supervised query-based video grounding, aiming to learn fine-grained frame-wise seman-
tic matching by progressively sampling harder negative-positive frames for discriminative feature learning. In particular, instead of relying on redundant segment proposals for matching and selection, we propose to learn more fine-grained frame-wise matching scores to predict whether each frame is the foreground frame. Once the scores of successive frames are larger than a learnable threshold, they are taken to construct the predicted segment, leading to more efficient grounding. The threshold is acquired from the frame-wise scores learned and enhanced by our developed multi-scale self-contrastive learning paradigm. We achieve such fine-grained frame-wise representation learning by the following twofold novelties.

Firstly, since we only resort to video-level annotations with no access to the frame-level knowledge, we propose frame-wise matching score prediction to estimate the score of each frame matched to the video-level annotations in a weakly-supervised manner, as well as the frame-wise matching weight via an attention mechanism in order to choose possible foreground frames with query semantics. Secondly, in order to improve the frame-wise score prediction by enhancing the frame-wise representations, we propose a self-contrastive learning with multi-scale hard negative mining strategy, which especially discriminates frames adjacent to the ground truth target segment—referred to as hard negative samples as shown in Figure 1. In particular, we dynamically set a range according to the previously estimated scores so as to select positive and negative samples, where samples with scores within the range serve as negative frames and those above the range as positive ones. The range is progressively updated to exploit harder negative samples that are more similar to the positive ones, which leads to more discriminative features learning. Note that, our contrastive learning strategy is quite different from the previous vanilla ones [28, 43, 47] in this task, since they all utilize a one-step algorithm to define the constant negative samples with coarse frame-level representations. Compared to them, we employ a multi-step process to iteratively mine the negative samples in a coarse-to-fine manner, defining harder negative samples and thus leading to more discriminative frame-wise representation learning. Further, we explore hard negative samples from different hierarchies—local frame-scale and nonlocal segment-scale, thus learning multi-scale intrinsic features.

Specifically, given the input video and query, we first encode both visual and textual features, and align their semantics by a video-query attention mechanism to learn the cross-modal interactions. Next, we predict frame-wise matching scores referring to the interacted features for foreground frame (inside the target segment) localization, as well as the frame-wise contribution weights for their importance estimation, which are aggregated to compute the overall semantic dependency between the video-query pair. Further, we enhance frame-wise representations by learning more discriminative features via the proposed multi-scale self-contrastive learning strategy, where both frame-scale
and segment-scale hard negative sampling are deployed in a coarse-to-fine manner.

Our main contributions are summarized as follows:

- We propose a novel self-contrastive learning framework for weakly-supervised query-based video grounding, which predicts fine-grained frame-wise matching scores referring to the query semantics for more accurate segment localization.

- We propose a multi-scale hard negative mining in the self-contrastive learning to learn discriminative frame-wise representations by adaptively sampling hard negatives in the frame-scale and the segment-scale respectively, which captures both local and nonlocal intrinsic patterns.

- Extensive experiments demonstrate that the proposed MSCL model outperforms the state-of-the-art method significantly over two challenging benchmarks.

2. Related Work

Fully-supervised query-based video grounding. Most existing works address the video grounding task in a fully-supervised manner, where both the annotations of video-sentence pairs and corresponding segment boundaries are given. Traditional methods [6,10] utilize the proposal-based framework that samples video segment proposals through dense sliding windows and subsequently integrate query with these proposal representations via a matrix operation. To further mine the cross-modal interaction more effectively, some works [15–18,21–23,36,37,39,42,45,46] integrate the sentence representation with those pre-defined segment proposals individually, and then evaluate their matching relationships. The proposal with the highest matching score is selected as the target segment. Although the proposal-based methods can achieve significant performance, they severely rely on the quality of the proposals and are very time-consuming. Instead of utilizing the segment proposals, recent proposal-free methods [1,2,27,40] directly regress the temporal locations of the target segment. Specifically, they either regress the start/end timestamps based on the entire video representation [1,27], or predict at each frame to determine whether this frame is a start or end boundary [2,40]. These works are much more efficient than the proposal-based ones, but achieve relatively lower performance. However, both of proposal-based and proposal-free methods heavily rely on a large amount of human annotations that are hard to collect in practice.

Weakly-supervised query-based video grounding. As manually annotating temporal boundaries of target moments is time-consuming, recent research attentions have been shifted to developing weakly-supervised video grounding models [4,26,32,33], which only require video-level annotations. [26] proposed the first weakly-supervised model to learn a joint embedding space for video and query representations. [8] develop a two-stream structure to measure the moment-query consistency and conduct moment selection simultaneously. Although above methods have achieved promising performance, they are two-stage approaches that utilize multi-scale sliding windows to generate moment candidates, therefore suffering from inferior effectiveness and efficiency. To address this issue, [14,25,47] further improve the segment-sentence matching accuracy, and score all the moments sampled at different scales in a single pass. [38] employ a reinforcement learning framework to refine the segment boundary. However, almost all of the existing methods rely on the segment proposals for matching and selection, which fail to capture and distinguish more fine-grained details among visually similar frames for acquiring more accurate segment boundaries.

Contrastive Learning. Contrastive learning [3,9] is a self-supervised learning paradigm that has demonstrated its effectiveness in many tasks, such as image classification, object detection and point cloud classification. Previous works [28,43,47] also showed promising results of contrastive learning in video grounding. Typically, [28] proposed a dual contrastive learning loss function by utilizing video-level samples for video-to-video and video-to-query representation learning. A Counterfactual Contrastive Learning framework [47] is designed to distinguish video-level embeddings between counterfactual positive and negative samples for hard negative sampling. Our model differs significantly from these methods: 1) We leverage the input single video only to perform contrastive learning over frames and segments, without resorting to different instances of videos as in previous works. 2) We propose multi-scale hard negative sampling at the frame scale and the segment scale to iteratively capture both local and non-local intrinsic feature representations.

3. Methodology

3.1. Overview

We focus on weakly-supervised query-based video grounding. Given an untrimmed video and the language query, the goal is to localize the start and end time of the temporal moment corresponding to the language query. As illustrated in Figure 2, the proposed MSCL model mainly consists of four modules:

- Multi-modal encoding. Given the multi-modal input, we first employ video and query encoders to encode both visual and textual features, and then interact the cross-modal information for semantic alignment.
• **Frame-wise matching score prediction.** After generating query-specific video representations by the cross-modal interaction, we predict frame-wise matching scores and frame-wise matching weights referring to the aligned multi-modal features for choosing the possible foreground frames within the video.

• **Multi-scale self-contrastive learning.** In order to improve the frame-wise score prediction by enhancing the frame-wise representations, we perform self-contrastive learning at both the frame-scale and the segment-scale with progressive hard negative mining to distinguish more fine-grained frame-wise details, thus enforcing more accurate segment grounding.

• **Segment localization.** At the inference time, we first construct possible segments by choosing the consecutive frames with scores higher than the threshold, and then select the best segment by comparing the average scores of the internal frames within each segment.

We elaborate on the four modules in order as follows.

### 3.2. Multi-Modal Encoding

**Video and query encoders.** For each video, following previous works, we first employ a pre-trained C3D model [34] to extract its frame-level features $V = \{v_i\}_{i=1}^n \in \mathbb{R}^{n \times D_v}$, where $n$ is the number of frames and $D_v$ is the feature dimension. For each query, we deploy the GloVe model [29] to obtain word-level embeddings $Q = \{q_i\}_{i=1}^m \in \mathbb{R}^{m \times D_q}$, where $m$ is the number of the words and $D_q$ is the feature dimension. Then both video and query features are projected into the same latent space by two fully-connected layers before being fed into the modality-specific encoders $f_v(\cdot)$, $f_q(\cdot)$ to generate the final visual representations $\tilde{V}$ and query embeddings $\tilde{Q}$.

That is, $\tilde{V} = f_v(V')$, $\tilde{Q} = f_q(Q')$. Here, $f_v(\cdot)$, $f_q(\cdot)$ share weights and consist of four convolution layers, followed by a multi-head attention layer [35].

**Video-query interaction.** We further apply a video-query attention mechanism to learn the cross-modal interactions, where we calculate the similarity score $S \in \mathbb{R}^{n \times m}$ between video and query features, and use the SoftMax operation along the row and column to generate $S_v$ and $S_q$, respectively. Next, we compute the video-to-query ($V$) and query-to-video ($Q$) attention contexts [44] as:

$$V = S_v \cdot \tilde{Q} \in \mathbb{R}^{n \times D}, \quad Q = S_q \cdot \tilde{V}^T \cdot \tilde{V} \in \mathbb{R}^{m \times D}.$$  

Then a single feed-forward layer FFN (composed of multiple linear layers) is applied to generate the interacted output.
features \( V^q \in \mathbb{R}^{n \times D} \) as:

\[
V^q = \text{FFN}(\tilde{V}; \mathcal{V}; \tilde{V} \odot \mathcal{V}; \tilde{V} \odot \mathcal{Q}),
\]

where \( \odot \) denotes the Hadamard product.

### 3.3. Frame-wise Matching Score Prediction

In the weakly-supervised setting, we only have access to the knowledge of the matched video-query pair without corresponding detailed segment-level annotations. In order to determine which frame is matched with the query semantics and how much the frame contributes to the final grounding, we introduce a score-based self-supervised branch to predict frame-wise matching scores and frame-wise matching weights for choosing the most possible foreground frames. Specifically, we devise a frame score head \( h_s(\cdot) \) and a frame weight head \( h_w(\cdot) \) to predict the corresponding matching score \( s_i \) and weight \( w_i \) for each frame \( i \), respectively. Here, both \( h_s(\cdot), h_w(\cdot) \) are composed of three linear layers. The \( S = \{s_i\}_{i=1}^n \) and \( W = \{w_i\}_{i=1}^n \) are formulated as:

\[
S = \text{Sigmoid}(h_s(V^q)); W = \text{Softmax}(h_w(V^q)).
\]

Then, for the \( k \)-th video and \( k \)-th query in each batch, their final semantic matching score \( \hat{s}_{k,k} \) is calculated as

\[
\hat{s}_{k,k} = \sum_{i=1}^n s_i \cdot w_i.
\]

In addition, we also estimate the similarity score (utilizing dot-product attention) between video features \( V \) and query features \( Q \) to measure their distance. The overall score objective is defined as:

\[
L_{\text{score}} = -\log \frac{\sum_{k=1}^K (\hat{s}_{k,k} + V_k \cdot Q_k)}{\sum_{k=1}^K \sum_{j=1}^K (\hat{s}_{k,j} + V_k \cdot Q_j)},
\]

where \( \hat{s}_{k,j} \) represents the overall video score corresponding to the \( j \)-th query features and \( k \)-th video features at the same batch. \( \hat{s}_{k,k} \) denotes the score of the matched video-query pair. \( K \) denotes the batch size. In this way, we maximize the overall score of video and query features from correct pairs while minimizing the score of false pairs. After getting the matching scores of all frames, we take them as pseudo labels to provide better supervisions for iteratively training the following contrastive learning module, and the learned discriminative features in turn further lead to more precise matching score prediction.

### 3.4. Multi-scale Self-contrastive Learning

In order to discriminate the frame-wise representations for more accurate prediction of matching scores, we propose a multi-scale self-contrastive learning paradigm with hard negative mining to capture more discriminative frame-wise representations in a coarse-to-fine manner. Specifically, we iteratively mine hard negative samples that are close to positive samples in the representation space with a multi-step strategy. In each step, we first choose the positive and negative frames according to their predicted frame-wise scores at frame-scale, and then consider one positive segment with the highest segment score while taking the other segments as negative samples at frame-scale. Then, we perform both frame-scale and segment-scale contrastive learning to learn more discriminative fine-grained frame-wise details. The updated frame-wise features in turn provide more accurate matching score to mine harder negative samples in the next step of learning. By performing multi-scale self-contrastive learning with such an iterative strategy, our model is able to enforce more accurate segment grounding. We will illustrate the details of both frame- and segment-scale negative mining of each step in the following.

#### Frame-scale

In order to mine hard negative frames that are close to positive frames in the representation space, we iteratively assign a lower bound \( b_1 \) and an upper bound \( b_u \) to select positive and negative frames. The lower bound \( b_1 \) and the upper bound \( b_u \) are defined from frame-wise scores as:

\[
b_1 = b_1^0 \cdot \delta^{-e_0}, \quad b_u = \frac{1}{n} \sum_{i=1}^n s_i,
\]

where \( \delta \) is the increasing step and \( e_0 \) denote the current epoch and the updated cycle of epoch respectively. \( b_1^0 \) is the initial value of \( b_1 \). We set \( b_1^0 = e^{-8}, e_0 = 50, \delta = 10 \) during the training, that is, we increase \( b_1 \) exponentially by 10 every 50 epoch after the warm-up stage.

Accordingly, we consider frames with scores greater than \( b_u \) as positive frames, and other frames with scores ranging from \( b_1 \) to \( b_u \) as negative frames. The loss function of the frame-scale contrastive learning is defined as:

\[
L_{\text{fra}} = -\log \frac{\sum_{k=1}^K V_k \cdot Q_k \cdot p_k^f}{\sum_{k=1}^K V_k \cdot Q_k \cdot n_k^f},
\]

where \( p_k^f \) and \( n_k^f \) denote the binary index mask of positive
and negative frames at batch index $k$, respectively. The entries of corresponding indices are 1 and others are 0.

**Segment-scale.** In order to enforce more accurate segment grounding predictions, we locate the predicted segments $\{g_t\}_{t=1}^T$ with consecutive indices and calculate the segment score by averaging the scores of internal frame located in the segment. Then we consider the segment with the highest segment score as the positive segment and other segments as negative samples. The loss function of the segment-scale contrastive learning is formulated as:

$$L_{\text{seg}} = -\log \frac{\sum_{k=1}^{K} \mathbf{v}_k \cdot \mathbf{q}_k \cdot \mathbf{p}_k^g}{\sum_{k=1}^{K} \mathbf{v}_k \cdot \mathbf{q}_k \cdot \mathbf{n}_k^g}, \quad (7)$$

where $\mathbf{p}_k^g$ and $\mathbf{n}_k^g$ denote the binary index mask of frames located in positive and negative segments at batch index $k$, respectively. The entries of corresponding indices are 1 and the others are 0.

The overall objective of our model is minimized in an end-to-end manner and formulated as:

$$L = L_{\text{score}} + \lambda_{\text{fra}} \cdot L_{\text{fra}} + \lambda_{\text{seg}} \cdot L_{\text{seg}}, \quad (8)$$

where $\lambda_{\text{fra}}$ and $\lambda_{\text{seg}}$ denote the weighting hyper-parameters of the frame loss and segment loss, respectively. In the experiments, we set $\lambda_{\text{fra}} = 10$ and $\lambda_{\text{seg}} = 5$.

The overall algorithm of our training approach is summarized in Algorithm 1, where we utilize an iterative strategy to gradually mine the hard negative samples. In order to enforce the model predict accurate positive samples with higher confidence, we first warm-up our model in the first 50 epochs without the procedure of hard negative sampling. Then, we iteratively mine the hard negative samples with high similarity to the positive ones.

### 3.5. Segment Localization

At the inference time, we select the segment with the highest segment score as the final prediction. Specifically, we extract all possible segments by choosing consecutive frames with scores higher than the upper bound $b_n$, and calculate the segment score by averaging the scores of internal frames located in the segment. Then we take the segment with the highest segment score as the final output.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

**Charades-STA.** The Charades-STA dataset [7] is built based on the Charades [31] dataset, which contains 6,672 videos of indoor activities and involves 16,128 query-video pairs. There are 12,408 pairs used for training and 3,720 used 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-Caption.** The ActivityNet-Caption dataset [13] contains 20,000 videos with 100,000 queries, where 37,421 query-video pairs are used for training and 34,536 are used for testing. The average duration of the videos is 1 minute and 50 seconds. On average, each video in ActivityNet-Caption has 3.65 annotated moments and each annotated moment lasts for 36 seconds.

**Evaluation metrics.** Following previous works, we adopt the metrics “R@n, IoU=m” to evaluate our model, where “R@n, IoU=m” presents the proportion of the top $n$ moment candidates with IoU larger than $m$. Specifically, we set $n$ as 1, 5 and set $m$ as 0.3, 0.5, 0.7 in Charades-STA dataset and 0.1, 0.3, 0.5 in ActivityNet-Caption dataset.

### 4.2. Experimental Settings

To make a fair comparison with previous methods like [47], we extract video features from the pre-trained C3D network [34] and query features from the 300-d Glove embedding [29]. We train the model for 200 epochs with the batch size of 16. We use a warm-up training without hard negative sampling for 50 epochs. The dimension of encoded features is set to 512. Our model is optimized by Adam [12] with an initial learning rate of 0.01 and linear decay of learning rate. All experiments are conducted on single NVIDIA GeForce RTX 3090 GPU.

### 4.3. Comparison with State-of-the-art Methods

**Charades-STA.** We compare our method against the current state-of-the-art methods under weakly-supervised settings on Charades-STA dataset in Table 1. As can be seen, we achieve the best performance over all baselines.

| Method | Charades-STA | ActivityNet-Caption |
|--------|-------------|---------------------|
|        | R@1 | IoU=0.3 | IoU=0.5 | IoU=0.7 | R@1 | IoU=0.3 | IoU=0.5 | IoU=0.7 |
| TGA    | 32.14 | 19.94 | 8.84 | 80.58 | 65.52 | 33.51 | - | - | - |
| CTF    | 39.80 | 27.30 | 12.90 | - | - | - | 74.20 | 44.30 | 23.60 |
| ReLoCLNet | - | - | - | - | - | - | - | - | - |
| SCN    | 42.96 | 23.58 | 9.97 | 95.56 | 71.80 | 38.87 | - | - | - |
| MARN   | - | 31.94 | 14.81 | - | - | - | 71.45 | 55.69 | - |
| RTBPN  | 60.04 | 32.36 | 13.24 | 97.48 | 71.85 | 41.18 | 73.73 | 49.77 | 29.63 |
| VGN+CCL | - | 33.21 | 15.68 | - | - | - | 71.45 | 47.01 | 29.95 |
| Ours   | 58.92 | 43.15 | 23.49 | 98.02 | 81.23 | 48.45 | 75.61 | 55.05 | 38.23 |

**Table 1.** Comparison results on the Charades-STA and ActivityNet-Caption datasets.
4.4. Ablation Study

In this part, we conduct extensive ablation studies on each module in our MSCL including frame-wise matching score prediction and frame-/segment-scale hard negative sampling, the effect of batch size, and the hyper-parameters. Unless specified, we perform all ablation studies on the Charades-STA benchmark.

Effect of each module. In order to understand how each module in our MSCL affects the final performance, we explore the effect of each proposed loss as shown in Table 2. Our model with frame-wise matching score prediction outperforms the baseline by 6.72%, 1.18%, and 0.94% in terms of three criteria, which shows the effectiveness of this module. Introducing frame-scale and segment-scale hard negative sampling separately boosts the performance of our model with frame-wise matching score prediction only. Furthermore, by adding frame-scale and segment-scale hard negative sampling together, we observe the highest increasing range of 7.74%, 21.79%, and 13.71%. This demonstrates the superiority of our multi-scale hard negative sampling over baselines.

Analysis of frame and segment loss. Furthermore, we conduct extensive experiments to explore the weighting hyper-parameters of frame and segment loss used in multi-scale hard negative sampling in Table 3. Specifically, we take the value of $\lambda_{fra}$ and $\lambda_{seg}$ from (1, 5, 10) for different control settings. With the increase of $\lambda_{fra}$, the performance of our model decreases since we fail to consider the more global information of segments in the whole video. However, with the increase of $\lambda_{seg}$, more global information of segments is introduced such that the results of our MSCL are improved. As can be seen, our MSCL reaches the best performance when $\lambda_{fra}=10$ and $\lambda_{seg}=5$, which implies the importance of balancing the weight of the frame-wise and segment-wise hard negative sampling during the training.

Robustness to the batch size. In this part, we analyze the effect of the batch size on the final performance of our MSCL, as shown in Table 4, where we set the batch size as 4, 8, 16, 32, 48. From Table 4, we observe that our MSCL with the batch size of 16 achieves the best results in terms of all metrics. Meanwhile, the performance of our model does not change too much with the altering of the batch size. This further validates the robustness of our MSCL to the choice of the batch size. In other words, we do not need a large batch size that is desired in previous contrastive learning based methods [43,47] for weakly-supervised video grounding.

4.5. Visualization Results

In this section, we provide more detailed visualization results on how our MSCL predicts more accurate segment grounding results given frame score curves. Qualitative examples of hard negative sampling and foreground frame predictions on two benchmarks are visualized to validate the superiority of our MSCL.

Frame-wise matching scores. In order to better understand the effectiveness of the segment localization in our MSCL, we plot the frame-wise score curves with respect to the frame index among the positive frames in Figure 3. As can be seen, many hard negative samples with high frame scores appeared in the training process. With the multi-scale hard negative sampling, our MSCL predicts the segment with the highest score greater than the upper bound $b_u$ as the final output, which matches the target segment. This shows the importance of the segment localization in more accurate segment grounding.

Hard Negative Mining. In Figure 4, we also visualize the hard negative samples at epoch=0, 50, 100, 150. We observe the hard negative samples with high scores are progressively closer to the positive frames, which validates the effectiveness of our multi-scale hard negative sampling.
Table 4. Exploration study on the effect of batch size.

| Batch Size | IoU=0.3 R@1 | IoU=0.5 R@1 | IoU=0.7 R@1 | IoU=0.3 R@5 | IoU=0.5 R@5 | IoU=0.7 R@5 |
|------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 4          | 58.53±0.08  | 42.81±0.06  | 23.19±0.04  | 97.88±0.05  | 80.97±0.04  | 48.12±0.04  |
| 8          | 58.81±0.05  | 43.03±0.03  | 23.38±0.02  | 97.96±0.04  | 81.14±0.04  | 48.36±0.03  |
| 16         | 58.92±0.03  | 43.15±0.01  | 23.49±0.01  | 98.02±0.02  | 81.23±0.02  | 48.45±0.01  |
| 32         | 58.91±0.02  | 43.12±0.01  | 23.45±0.01  | 97.98±0.01  | 81.19±0.01  | 48.42±0.01  |
| 48         | 58.85±0.02  | 43.09±0.02  | 23.41±0.01  | 97.92±0.02  | 81.13±0.02  | 48.36±0.02  |

Figure 3. Visualization of frame score curves used for segment localization. Red dotted lines denote the upper bound $b_u$. The ground-truth of frame indices are 49-60, 53-62, 6-23, and 26-37.

Figure 4. An illustration of the dynamic process of hard negative sampling (Gray Shadow) at epoch=0, 50, 100, 150. GT denotes the ground-truth (red indexes denote the segment boundaries of ground-truth). It shows that our iterative mining strategy can mine harder negative samples with the step goes on, leading to more discriminative frame-wise representation learning.

samples of Charades-STA and ActivityNet-Caption benchmarks in Figure 5. By comparison, our MSCL achieves better performance than SCN [14] and VGN+CCL [47]. Particularly, we achieve more accurate results on the boundary of the ground-truth segment due to the effectiveness of our multi-scale hard negative sampling strategy.

5. Conclusion

In this work, we propose a novel multi-scale self-contrastive learning model for weakly-supervised query-based video grounding. Instead of utilizing redundant segment proposal for semantic matching, we predict frame-wise scores and weights for matching fine-grained frame-wise features with query semantics. In order to learn more discriminative frame-wise representations for predicting accurate frame-wise scores, we further introduce a multi-scale self-contrastive learning with multi-step hard negative mining strategy to progressively discriminate hard negative samples that are close to positive samples in the representation space. This iterative approach captures fine-grained frame-scale details as well as segment-scale semantics for distinguishing frames with high repeatability and similarity within the entire video. Experimental results show that our proposed model outperforms state-of-the-art methods on two challenging benchmarks.
Figure 5. Visualization of examples on Charades-STA and ActivityNet-Caption benchmarks. GT denotes the ground-truth.

References

[1] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In Proceedings of the American Association for Artificial Intelligence, 2019.

[2] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10551–10558, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Proceedings of International Conference on Machine Learning (ICML), 2020.

[4] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. arXiv preprint arXiv:2001.09308, 2020.

[5] Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok Choudhary. Temporal sequence modeling for video event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2227–2234, 2014.

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. Proceedings of the IEEE International Conference on Computer Vision (ICCV), page 5267–5275, 2017.

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5267–5275, 2017.

[8] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. Wsln: Weakly supervised natural language localization networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1481–1487, 2019.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9729–9738, 2020.

[10] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), page 1380–1390, 2018.

[11] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 499–515, 2018.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 706–715, 2017.

[14] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11539–11546, 2020.

[15] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. arXiv preprint, 2022.

[16] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In AAAI, 2022.

[17] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In COLING, 2020.
[18] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In EMNLP, 2021. 3
[19] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yuilai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In CVPR, 2021. 1
[20] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In ACM MM, 2020. 1
[21] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In AAAI, 2022. 3
[22] Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In EMNLP, 2021. 3
[23] Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. Exploring motion and appearance information for temporal sentence grounding. In AAAI, 2022. 3
[24] Jingzhou Liu, Wenhui Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 10900–10910, 2020. 1
[25] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), pages 156–171, 2020. 1, 3
[26] Niluthpol Chowdhury Mithun, Sujay Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 11592–11601, 2019. 1, 3
[27] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 3
[28] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Intervenitional video grounding with dual contrastive learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2765–2775, 2021. 2, 3
[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014. 4, 6
[30] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1049–1058, 2016. 1
[31] Gunnar A Sigurdsson, Gýl Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In European Conference on Computer Vision (ECCV), pages 510–526. Springer, 2016. 6
[32] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. arXiv preprint arXiv:2003.07048, 2020. 3
[33] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pages 2083–2092, 2021. 3
[34] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2015. 4, 6
[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS), pages 5998–6008, 2017. 4
[36] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12168–12175, 2020. 1, 3
[37] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 3
[38] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In Proceedings of the 28th ACM International Conference on Multimedia, pages 1283–1291, 2020. 3
[39] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 1, 3
[40] 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), 2020. 3
[41] Yawen Zeng, Da Cao, Xiaoichi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2215–2224, 2021. 1
[42] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 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), pages 1247–1257, 2019. 3
[43] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In Proceedings of the 44th International ACM SIGIR Conference.
[44] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6543–6554, 2020.

[45] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12870–12877, 2020.

[46] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 655–664, 2019.

[47] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems (NIPS), 33:18123–18134, 2020.