WEAKLY-SUPERVISED MOMENT RETRIEVAL NETWORK FOR VIDEO CORPUS
MOMENT RETRIEVAL

Sunjae Yoon∗ Dahyun Kim∗ Ji Woo Hong Junyeong Kim
Kookhui Kim Chang D. Yoo

Korea Advanced Institute of Science and Technology (KAIST)

ABSTRACT
This paper proposes Weakly-supervised Moment Retrieval Network (WMRN) for Video Corpus Moment Retrieval (VCMR), which retrieves pertinent temporal moments related to natural language query in a large video corpus. Previous methods for VCMR require full supervision of temporal boundary information for training, which involves a labor-intensive process of annotating the boundaries in a large number of videos. To leverage this, the proposed WMRN performs VCMR in a weakly-supervised manner, where WMRN is learned without ground-truth labels but only with video and text queries. For weakly-supervised VCMR, WMRN addresses the following two limitations of prior methods: (1) Blurry attention over video features due to redundant video candidate proposals generation, (2) Insufficient learning due to weak supervision only with video-query pairs. To this end, WMRN is based on (1) Text Guided Proposal Generation (TGPG) that effectively generates text guided multi-scale video proposals in the prospective region related to query, and (2) Hard Negative Proposal Sampling (HNPS) that enhances video-language alignment via extracting negative video proposals in positive video sample for contrastive learning. Experimental results show that WMRN achieves state-of-the-art performance on TVR and DiDeMo benchmarks in the weakly-supervised setting. To validate the attainments of proposed components of WMRN, comprehensive ablation studies and qualitative analysis are conducted.

Index Terms— Multi-modal video corpus moment retrieval, Weakly-supervised learning

1. INTRODUCTION
Understanding quintessential meaning between visual information and natural language appears to be a desiderata in high-level vision-language tasks. Umpteen endeavors for enlinking visual context and language semantics has contributed to the expansion of vision-language tasks, including image/video captioning [1], image/video retrieval [2, 3], image/video question answering [4, 5]. Among these tasks, Video Corpus Moment Retrieval (VCMR) is a task to localize a moment in large video corpus, which includes identifying relevant video in multiple videos and searching for a specific moment in the identified video.

Previous works [6, 7] for VCMR require full supervision for training, which includes annotating temporal boundaries for a large number of videos, which is labor intensive process. To leverage this, the proposed WMRN performs VCMR in a weakly-supervised manner, where WMRN is learned without ground-truth temporal boundaries but only with the video and text query. For weakly-supervised VCMR (wVCMR), our proposed WMRN addresses the following two main challenges: (1) Blurry attention over video features due to redundant video candidate proposals generation, (2) Insufficient learning due to weak supervision only with video-query pairs.

One of the key challenges in wVCMR is to mitigate a redundancy among unnecessary video candidate proposals. A bunch of proposals containing video segment is generated to find a moment corresponding to given query, which result in a burden for network to perform an attention on the proposals and entail heavy computations. The other key challenge is to cope with insufficient learning without ground-truth temporal boundaries. Although contrastive learning has recently been applied for promoting interpretability under weak su-
Rachel: Hey, I'm having dinner with my dad. Do you wanna come? Rachel invites Phoebe to have dinner with her and her father.

Fig. 2: Illustration of WMRN for VCMR. WMRN is composed of the following components: (a) Text Guided Proposal Generation, (b) Hard Negative Proposal Sampling. (best viewed in zoom)

2. RELATED WORKS

2.1. Video Moment Retrieval

Moment retrieval task has evolutionarily grown up into general format of retrieval. Input modalities extended from a word and short plot to sentences and video corpus including interplay among multiple characters [7, 8]. Jie et al. [6] recently propose VCMR which incorporates video retrieval and single video moment retrieval (SVMR). Despite the respectful efforts for retrieval tasks, annotating temporal boundaries is still in demand. To deal with this issue, weakly-supervised methods have been proposed. Mithun et al. [2] first proposed weakly-supervised framework which performs SVMR without boundary annotations. The methods in [9, 10] proposed to learn video-language alignment by embedding semantics into joint space. Although revolutionary ways of using weak supervision have been proposed, they do not fully explore weakly-supervised manner in a video corpus level.

2.2. Video Proposal Generation

Video proposals contain a subset of the video and consist of candidates for moment retrieval. As, in weakly-supervised manner, learning from boundary information is not available, generating multi-scale proposals and finding the most pertinent one are important. Ma et al. [10] contributes to selecting surrogate proposals in early stage. Although previous works [9, 11] proposed efficient proposal selection, reducing redundant proposals and generating ones are still challenging.

3. METHOD

Figure 2 gives a schematic of Weakly-supervised Moment Retrieval Network (WMRN) composed of two main modules: Text Guided Proposal Generation (TGPG) and Hard Negative Proposal Sampling (HNPS). The WMRN is designed based on following two insights: (1) Effectively generate multi-scale proposals pertinent to given query reduces the redundant computations and enhances attention performance and, (2) extract hard negative samples in contrastive learning to contribute to high-level interpretability. TGPG increases the number of candidate proposals in proximity of moments related to query, while decreases proposals in unrelated moments. HNPS selects negative samples in positive videos, which contribute to enhanced contrastive learning.

3.1. Input Representation

As previous work [7], we are given video, subtitles and a single sentence query as 

\[ V = \{ v_i \}_{i=1}^{N_v} \], \[ S = \{ s_i \}_{i=1}^{N_s} \] and \[ q \],

where \( N_v \) and \( N_s \) is the number of frames and subtitles in a single video. For text encoder, we extract contextualized token features from pre-trained on RoBERTa [12], where the tokens for subtitle and query are denoted as \( W_{vi} = \{ w_{iq}^i \}_{j=1}^{L_i} \) and \( W_q = \{ w_{iq}^j \}_{j=1}^{L_q} \). The \( L \) and \( N_q \) are the number of tokens in the \( i \)-th subtitle \( s_i \) and query \( q \). For video encoder, 2D features from ResNet [13] pre-trained on ImageNet [14] and 3D features from SlowFast [15] pre-trained on Kinetics [16] are concatenated after L2-normalization, which is denoted \( E^v \in \mathbb{R}^{N_v \times d} \). We also define the video features aligned...
with subtitles as $V_s = \{v_j^s\}_{j=1}^K$, where the $K$ is the number of frames in $s_i$. The final video, subtitle and query features are represented in $d$-dimensional space as follows:

$$E_{v_i}^w = LN(V_{s_i} + PE(V_{s_i})) \in \mathbb{R}^{K \times d},$$ (1)

$$E_{w_i}^s = LN(W_{s_i} + PE(W_{s_i})) \in \mathbb{R}^{L \times d},$$ (2)

$$E_{q}^s = LN(W_q + PE(W_q)) \in \mathbb{R}^{N_q \times d},$$ (3)

where PE is positional encoding [17] and LN is layer normalization [18]. $E_{v_i}^w$ is query token features and in subtitle $s_i$, the video features $E_{v_i}^s$ are attended by subtitle token features $E_{w_i}^s$ using hierarchical encoder [7]. The hierarchical encoder is composed of two Transformers [17] performing local and global context matching, which is called Cross-modal Transformer and Temporal Transformer. Cross-modal Transformer takes concatenated video-subtitle features $E_{v_i}^{w_s} = [E_{v_i}^{w}, E_{w_i}^s]$ as inputs and outputs cross attended video and subtitle features $[C_{v_i}^w, C_{w_i}^s]$. After collecting all the attended visual features $C_{v_i}^w = \{C_{v_i}^w\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times d}$, Temporal Transformer outputs temporally attended video features $T_{v}^w \in \mathbb{R}^{N_v \times d}$ from cross attended features $C_{v_i}^w$ and original features $E_{v_i}^w$ as:

$$(c_{v_i}; c_{w_i}) = \text{MultiHead}(E_{v_i}^w; E_{w_i}^s),$$ (4)

$$T_{v}^w = \text{MultiHead}(E_{v}^w + C_{v}^w; E_{v}^w + C_{v}^w).$$ (5)

### 3.2. Text Guided Proposal Generation

Text Guided Proposal Generation (TGPG) in Figure 2 effectively generates multi-scale video proposals in the prospective region related to query via similarity of query and video-subtitle, TGPG is performed from two key properties: (1) Similarity Integral Curve, (SIC) (2) Multi-scale Proposals (MP). We first calculate average pooling $A_{v_i}^w$ over $E_{v_i}^w$ and frame-level cosine similarities $S_{vq}$ between $T_{v}^w$ and $A_{v_i}^w$ as:

$$A_{v_i}^w = \text{Pool}_{\text{avg}}(E_{v_i}^w) \in \mathbb{R}^d,$$ (6)

$$S_{vq} = \text{cosine}(T_{v}^w W_k^k, A_{v_i}^w W_k^k) \in \mathbb{R}^{N_v}.$$ (7)

$W_k \in \mathbb{R}^{d \times d}$ is learnable parameters. The negative values in $S_{vq}$ are clipped to zero. The final $I_{SIC}$ is defined by integrating similarities $S_{vq}$ along the frame axis as follows:

$$I_{SIC}(n) := \sum_{i=1}^{n} S_{vq}^{i}, (1 \leq n \leq N_v).$$ (8)

Based on $I_{SIC}$, we generate new multi-scale text guided $N_p$ proposals $P^w = \{P^w_j\}_{j=1}^{N_p}$ by defining $y_k$ and $x_k$ like below:

$$y_k = I_{SIC}(N_v/N_p)k, (0 \leq k \leq N_p),$$ (9)

$$x_k = I_{SIC}^{-1}(y_k).$$ (10)

In figure 3, the $y_k$ is $k$-th location in y-axis, dividing the length $I_{SIC}(N_v)$ by $N_p$. The $x_k$ is output of inverse function $I_{SIC}^{-1}$ with input $y_k$. The proposal $P^w_j$ is generated by taking an average of $T_{v}^w$ along frames in interval $[x_j, x_{j+1}]$.

$$P^w_j = \frac{\sum_{i=x_j}^{x_{j+1}} T_{v}^w_i}{x_{j+1} - x_j}, (0 \leq j \leq N_p - 1).$$ (11)

The $M$ is the size of combining proposals and $P^M$ represents a larger area of the video. Final text guided multi-scale proposal $P = P^w \cup P^1 \cup \cdots \cup P^M$ produces a large number of various proposals in proximity of high similarity between query and video, shown in Multi-scale Proposals of Figure 2.

### 3.3. Hard Negative Proposal Sampling

Hard Negative Proposal Sampling (HNPS) extracts hard negative proposal in positive video which contributes to enhance the contrastive learning with margin $\Delta_c$. In original hinge loss $L_o$, negative proposal is selected in other videos $P^- \setminus P^w$ per batch. However, HNPS loss $L_h$ selects proposal with the lowest similarity in positive video $P^+$ as negative samples:

$$L_o = \max[0, \Delta_c - \max(P^+ A_{w_i}^w) + \min(P^- A_{w_i}^w)],$$ (13)

$$L_h = \max[0, \Delta_c - \max(P^+ A_{w_i}^w) + \min(P^+ A_{w_i}^w)],$$ (14)

$$L = \alpha L_o + \beta L_h,$$ (15)

where and $L_o$ contributes to video-level retrieval and $L_h$ contributes to moment-level retrieval and total loss $L$ is linear combination of $L_o$ and $L_h$ using hyperparameters $\alpha$ and $\beta$.

### 4. EXPERIMENTS

#### 4.1. Datasets

We experimented with two datasets: TVR [6], DiDeMo [19]. TV show Retrieval (TVR) [6] is video-subtitle dataset with...
Table 1: Performance comparison of WMRN to the related methods on TVR dataset. Model references: MEE [20]

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>IoU=0.7</th>
<th>R@1</th>
<th>R@10</th>
<th>R@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully</td>
<td>MEE+MCN [19]</td>
<td>0.42</td>
<td>2.98</td>
<td>10.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MEE+CAL [21]</td>
<td>0.39</td>
<td>2.98</td>
<td>11.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HERO [7]</td>
<td>6.21</td>
<td>19.34</td>
<td>36.66</td>
<td></td>
</tr>
<tr>
<td>Weakly</td>
<td>MEE+TGA [2]</td>
<td>0.24</td>
<td>1.57</td>
<td>5.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MEE+VLANet [10]</td>
<td>0.69</td>
<td>3.84</td>
<td>10.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WMRN (ours)</td>
<td>1.74</td>
<td>9.44</td>
<td>23.58</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Performance comparison of WMRN to the related methods on DiDeMo dataset.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>IoU=0.7</th>
<th>R@1</th>
<th>R@10</th>
<th>R@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully</td>
<td>HERO</td>
<td>2.14</td>
<td>11.43</td>
<td>36.09</td>
<td></td>
</tr>
<tr>
<td>Weakly</td>
<td>MEE+VLANet</td>
<td>0.33</td>
<td>2.13</td>
<td>9.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WMRN (ours)</td>
<td>0.62</td>
<td>3.09</td>
<td>12.96</td>
<td></td>
</tr>
</tbody>
</table>

109K queries and 21.8K videos from 6 TV shows across 3 genres (sitcom, medical, crime). The videos are aligned with subtitles and the average length is 76.2 seconds. The dataset is composed of 80% train, 10% val, 5% test-public and 5% test-private, where the test-public is used for a leaderboard. DiDeMo [19] is constructed for VMR task. It includes 41.2K queries and 10.6K unedited videos from Flickr. The videos are divided into 5-seconds moments, which produce 21 candidate moments in 30-second videos. The dataset is split into 80% train, 10% val and 10% test.

4.2. Quantitative Results

Table 1 and Table 2 summarize the experimental results on TVR and DiDeMo. We compare WMRN with previous methods in both fully-supervised and weakly-supervised manner. WMRN achieves the state-of-the-arts performance on both TVR and DiDeMo in weakly-supervised setting. MCN and CAL are incorporated with video retrieval system MEE to perform VCMR. TGA and VLANet are also incorporated with MEE to perform wVCMR. As DiDeMo in Table 2 uses predefined proposals, we apply only HNPS for the WMRN. The results indicate that text-guided proposals and sampling hard negative proposals can help to improve interpretability and boost performance in weakly-supervised setting.

4.3. Ablation studies

We experiment with several variants of WMRN to validate effectiveness of our key components. The second block of Table 3 provides the ablation results of TGPG and HNPS. In the WMRN without TGPG, the same multi-scale proposals in [10] were used, where we can see that TGPG contributes to high performance. The M is combining size of proposals and performance drops when a large number of proposals are combined, which results from proposal redundancy. The results from WMRN without HNPS also show performance drop but not as much as TGPG.

Fig. 4: Illustration of TGPG and HNPS. Red proposal is hard negative sample and dark yellow is top-1 proposal.

4.4. Qualitative Results

Figure 4 visualizes the TGPG and HNPS. The red colored proposal is selected as hard negative sample and HNPS was performed on the distant proposal from GT moment. The yellow colored proposals represents how similar the proposals are to a query. TGPG produces a multi-scale proposals in a proximity of GT and the most similar proposal largely overlaps with that of GT.

5. CONCLUSION

In this paper, we propose Weakly-supervised Moment Retrieval Network (WMRN) for Video Corpus Moment Retrieval in a weakly-supervised manner. For two main challenges in wVCMR: blurry attention over redundant video proposals and insufficient learning for weak-supervision, WMRN is proposed TGPG that generates text guided multi-scale proposals and HNPS that improves contrastive learning, which achieves high-level performance and interpretability.
6. REFERENCES


