2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428369

DIVING INTO THE RELATIONS: LEVERAGING SEMANTIC AND VISUAL STRUCTURES FOR VIDEO MOMENT RETRIEVAL

Ziyue Wu[1], *Junyu Gao*[2,3], *Shucheng Huang*[*1] *and Changsheng Xu*[2,3,4]

[1]Jiangsu University of Science and Technology
[2]The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4]PengCheng Laboratory, Shenzhen, China
wuziyuewzy@gmail.com, schuang@just.edu.cn, {junyu.gao, csxu}@nlpr.ia.ac.cn

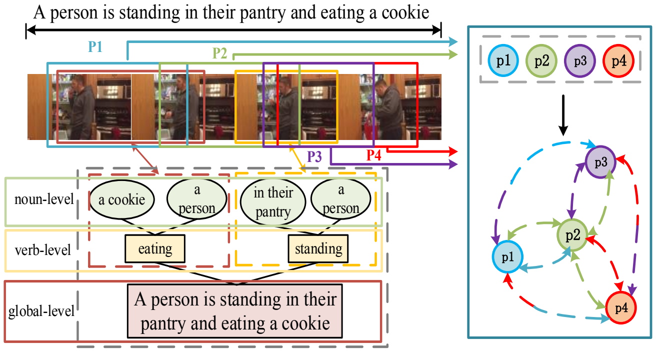
ABSTRACT

Existing dominant approaches for video moment retrieval task are to learn semantic correlation between a given query and the video. However, these methods rarely explore the fine-grained semantic structure and comprehensive visual structure, leading to insufficient utilization of textual and visual relations. In this paper, we propose a unified framework for video moment retrieval, which considers to simultaneously encode semantic and visual structures. Specifically, a semantic role tree is built to reveal the fine-grained semantic information by generating hierarchical textual embeddings. Then the semantic structure is adopted to facilitate the visual structure learning with a contextual attention-based proposal interaction module. Finally, we adaptively aggregate and obtain the visual-semantic matching information through a multi-level fusion strategy to select the best matching moment proposal. Extensive experiments on two popular benchmarks (Charades-STA and ActivityNet Captions) show that our proposed method achieves state-of-the-art performance. Codes are available in the Supplementary Material.

Index Terms**—** Video moment retrieval, semantic role tree, visual structure modeling

1. INTRODUCTION

We address natural language-based moment retrieval in untrimmed videos. Given an untrimmed video and a natural language query, video moment retrieval (VMR) is to localize the correct temporal segment that is semantically aligned with the given query. It has various applications such as robotic navigation, video entertainment, and autonomous driving, to name a few [1, 2, 3, 4, 5]. Despite much progress has been achieved in recent years [6, 7, 8, 9, 10, 11, 12, 13], VMR remains difficult due to the harsh nature of videos and texts, including complex temporal relations, fine-grained semantic structures, and huge cross-modal gap between visual and textual features [11, 14, 15, 16]. The current dominant approaches for video moment retrieval is to learn the semantic correlation between the query and the video. To this end, numerous cross-modality alignment strategies are designed such as cross-attention [1, 2], recurrent neural networks [17, 18], semantic conditioned dynamic modulation [11], and 2D temporal adjacent network [14].


A person is standing in their pantry and eating a cookie
P1
P2
P3
P4
noun-level
verb-level
global-level
a cookie
a person
in their pantry
a person
eating
standing
A person is standing in their pantry and eating a cookie
p1
p2
p3
p4

(a) (b)

Fig. 1: (a) Semantic structure. A query consists of multiple semantic levels including global-level, verb-level and noun-level. (b) Visual Structure. Different proposals may have relations to others in content or context.

Although achieving favorable performance, most current methods do not take full advantage of the fine-grained and comprehensive relation information in both semantic and visual structures: **(1)** Many existing VMR approaches only encode the semantic information of the query in a global manner [9, 19, 10, 20, 14, 12, 13], i.e., embedding the texts into a global vector representation by using LSTM or other sequential models, but ignore the intrinsic and fine-grained structure of the sentence. Obviously, as shown in Figure 1(a), a

[*]Shucheng Huang(schuang@just.edu.cn) is the corresponding author.

This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0100604, in part by the National Natural Science Foundation of China under Grants 62072286, 61720106006, 6207245561721004, 61772244, 61832002, 62002355, U1836220, and U1705262, and in part by the Key Research Program of Frontier Sciences of CAS under Grant QYZDJSSWJSC039, and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province, Grant NO. KYCX20_3164.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.

query sentence (e.g., "A person is standing in their pantry and eating a cookie.") corresponding to a specific video moment has multiple semantic levels including global-level (the whole sentence), verb-level (actions, behaviors), and noun-level (objects, entities). The multiple semantic levels involves complicated interactions between them, which are actually organized as a tree structure, indicating that a textual query can be effectively grounded onto the video by properly aligning different semantic levels with the corresponding video parts. Although some previous methods attempt to utilize the semantic structure in either a partial [10] or an implicit fashion [7], they fail to capture the fine-grained and explicit semantic structure of the query. **(2)** Intuitively, humans usually takes the relation-aware context of the visual structure into consideration and select the correct temporal moment. As a result, the visual structure is also important for temporal localization [21], which can provide more cues to facilitate the recognition of each moment proposal. In VMR, the visual structure is typically regarded as the interaction among different moment proposals, as shown in Figure 1 (b). However, most traditional VMR models neglect to exploit the proposal-proposal relations. Although a few approaches explore the visual structures by using graph neural networks [8] or self-attention [22], they do not consider the fine-grained semantic structure of the query for guiding the proposal-proposal relation modeling. In fact, both the semantic and visual structures can enhance and complement each other in the VMR task. Here, a fine-grained semantic structure can make the relation learning of moment proposals more accurate, and a comprehensive visual structure can effectively match relevant semantic details from the query.

Motivated by the above observations, in this paper, we propose a novel video moment retrieval method as depicted in Figure 2, which can perform fine-grained and comprehensive relation modeling by jointly leverage both semantic and visual structures in an end-to-end manner. Specifically, for semantic structure, we decompose a query sentence into three different semantic levels by building a semantic role tree, where an attention-based message passing module is designed to propagate semantic information across the hierarchical structure. The learned fine-grained semantic information is then adopted to facilitate the visual structure learning with a contextual attention-based proposal interaction module. Finally, the multi-level visual-semantic matching information is adaptively aggregated to select the best matching moment proposal.

The contributions of this work are three-fold: **(1)** We propose a joint semantic and visual structure modeling approach for video moment retrieval, which can get benefit from the fine-grained and comprehensive relation learning. **(2)** A multi-level proposal interaction learning module is designed, leading the semantic and visual structures enhance and complement each other in a unified framework. **(3)** We conduct comprehensive experiments on two popular VMR benchmarks and achieve state-of-the-art performance.

2. RELATED WORK

Existing VMR methods can be grouped into two categories: one-stage and two-stage. Most one-stage approaches [23, 6, 7] aim to build a proposal-free strategy and directly regress temporal locations of target moment. Zeng et al. propose the Dense Regression Network (DRN) [6] that aims to regress the distance from each frame to the boundary frame of target moment. Mun et al. design a local-global video-text interaction algorithm [7], which uses a sequential query attention module (SQAN) and exploits the semantic information from local to global without considering the explicit semantic structure construction. Most two-stage methods [20, 8, 9, 10, 11, 14] firstly sample some moment proposals from the video and then rank them relying on the similarity between proposals and the query. To improve the quality of the proposals, Zhang et al. design the Moment Alignment Network (MAN) [8] to model the complex temporal relations by using a graph neural network. Xu et al. propose the Query-Guided Segment Proposal Network (QSPN) [9] which designs a multilevel approach to combine visual features and textual features with an auxiliary video captions task. Ge et al. propose the Activity Concepts based Localizer (ACL) [10] which matches activity concepts information with generated proposals.

Although these existing methods achieve impressive performance, most of them do not make full use of the fine-grained and comprehensive relation information in both semantic and visual aspects. On the one hand, although MAC and SQAN exploit the semantic structure of a sentence, they are implemented in a partial [10] or an implicit manner. On the other hand, a few methods like MAN or VSLNet explore the visual structures by using GNN or self-attention, however, they do not consider the fine-grained semantic structure of the query for guiding the proposal-proposal relation modeling.

3. OUR APPROACH

Given an untrimmed video V and a query sentence S, our task aims to localize the start and end timestamps of a video segment (i.e., a temporal bounding box $b = (t_s, t_e)$ meaning a video segment starting at t_s and ending at t_e), which is best matching to the given query.

3.1. General scheme

We design a novel method for video moment retrieval, named SV-VMR, which jointly models the fine-grained and comprehensive relations by using both semantic and visual structures. Specially, as shown in Figure 2, an off-the-shelf semantic role parsing toolkit [24] is employed to extract different semantic levels in a query. Then, we construct a semantic role tree and perform message passing on it. We obtain semantic-aware features of each moment candidate via the proposed multi-level proposal interaction learning. Finally, a BCE (Binary Cross-Entropy) loss is used to optimize the whole framework.

3.2. Semantic structure modeling via semantic role tree

For a given query sentence S, we employ the off-the-shelf semantic role parsing toolkit [24] to obtain verbs and noun

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.

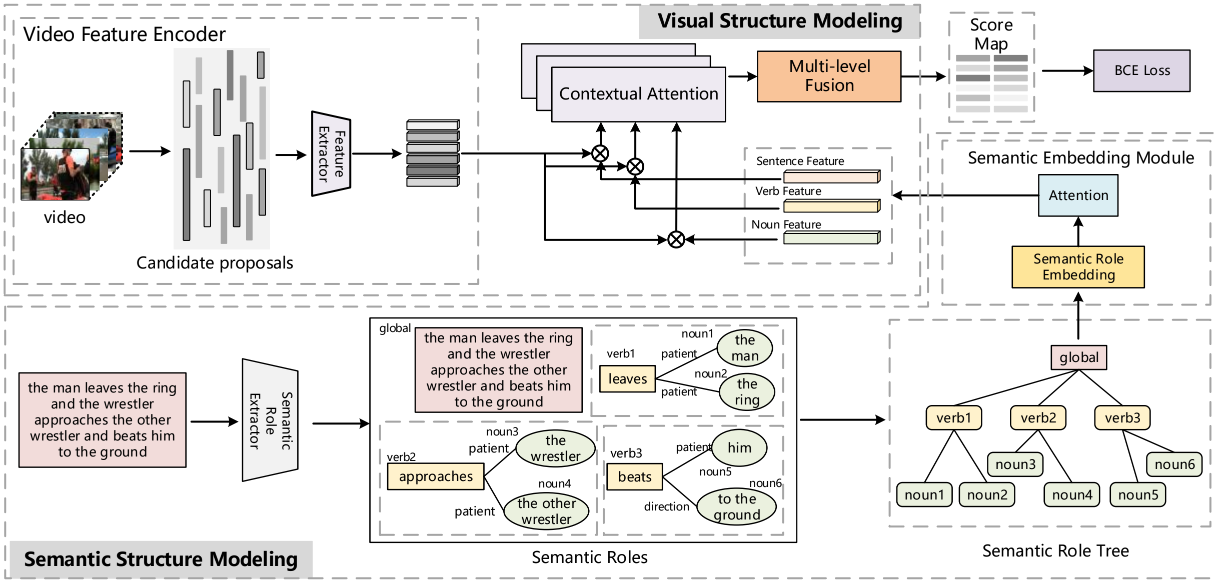


Fig. 2: Overview of our proposed SV-VMR framework. SV-VMR mainly consists of two components including the semantic structure modeling and the visual structure modeling. We first build a semantic role tree from the given query. Then we fuse visual features with different semantic level features. The interaction between different proposals is explored by the contextual attention. Finally we rank all proposals by a multi-level fusion module and optimize the whole framework with a BCE loss.

phrases. Then we set these phrases as nodes in our semantic role tree as shown in Figure 2, where we consider the whole sentence as the root node. All the verb nodes are connected with the root node, and noun nodes are set as leaf nodes. If a noun is related to a verb in the semantic aspect, we connect the noun node with the verb node. For example, in the sentence "The man leaves the ring and the wrestler approaches the other wrestler and beats him to the ground", we set the whole sentence as the root node which is connected with the verb nodes "leaves", "approaches" and "beats". The phrases "the man", "the ring", "to the ground" and "the wrestler" are connected with "leaves", "approaches" and "beats" as noun nodes respectively. For the sentence $S = \{s_1, ..., s_L\}$ with L words, we first employ a bidirectional LSTM (Bi-LSTM) to obtain a sequence of contextual-aware word embeddings $\{\mathbf{w}_1, ..., \mathbf{w}_L\}$. Then we utilize an attention-based strategy [25] to generate different semantic embeddings. After that we obtain three semantic-level embeddings $\mathbf{g}^{(s)}$, $\mathbf{g}^{(v)}$ and $\mathbf{g}^{(n)}$, representing sentence embedding, verb embeddings and noun embeddings respectively. Note that $\mathbf{g}^{(v)} = \{\mathbf{g}_1^{(v)}, ..., \mathbf{g}_{N_v}^{(v)}\}$ and $\mathbf{g}^{(n)} = \{\mathbf{g}_1^{(n)}, ..., \mathbf{g}_{N_n}^{(n)}\}$, N_v and N_n mean the number of verb nodes and noun nodes.

Attention-based message passing. Considering the three semantic levels, we adopt an attention-based message passing module to propagate semantic information across the hierarchical structure as follows:

$$\begin{aligned} \mathbf{e}_i^{(r)} &= \sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot (\mathbf{W}^r \mathbf{g}_j^{(r)}) \\ \alpha_{ij} &= \frac{\exp((\mathbf{W}_1 \mathbf{g}_i^{(r)})^{\mathrm{T}} \cdot (\mathbf{W}_2 \mathbf{g}_j^{(r)}))}{\sum_{j \in \mathcal{N}_i} \exp((\mathbf{W}_1 \mathbf{g}_i^{(r)})^{\mathrm{T}} \cdot (\mathbf{W}_2 \mathbf{g}_j^{(r)}))} \end{aligned} \tag{1}$$

where $\mathcal{N}_i$ means the neighbors of the i-th node including itself, $r \in \{s, v, n\}$, $\mathbf{e}^{(r)} = \text{MeanPooling}(\{\mathbf{e}_i^{(r)}\}_{i=1}^{N_r})$. We obtain the semantic features denoted as $\mathbf{e}^{(s)}$, $\mathbf{e}^{(v)}$ and $\mathbf{e}^{(n)}$, which means sentence features, verb features, and noun features, respectively. These features depict the information from different levels, and play an important role in modeling the visual structure. Note that $\mathbf{e}^{(r)} \in \mathbb{R}^{d^E}$ and d^E is the dimension of the semantic embeddings.

3.3. Visual structure modeling via multi-level proposal interaction learning

Video feature encoder. For a given untrimmed video V, this component extracts the features of input video frames and encodes the features into a set of proposals. We follow a sparse sampling strategy [14] to obtain moment proposals. We denote these proposals as $P = \{p_m\}_{m=1}^{M}$, where p_m represents a proposal and M means the number of all proposals. Then we utilize a pretrained 3D CNN model (e.g., I3D [26]) to extract features from every proposal p_m :

$$\mathbf{U} = \{\mathbf{u}_1, ..., \mathbf{u}_m, ..., \mathbf{u}_M\} = \mathbf{Encoder}(\{p_m\}_{m=1}^{M}) \tag{2}$$

where $\mathbf{u}_m$ is the feature of proposal p_m, $\mathbf{u}_m \in \mathbb{R}^D$, D is the dimension of the feature.

Cross-model feature fusion. With both the query features and the video features, we first fuse the proposal features $\mathbf{U} = \{\mathbf{u}_1, ..., \mathbf{u}_M\}$ with the transformed features from different semantic levels $\mathbf{E} = \{\mathbf{e}^{(s)}, \mathbf{e}^{(v)}, \mathbf{e}^{(n)}\}$ as follows:

$$\begin{aligned} \widetilde{\mathbf{e}}^{(r)} &= \text{ReLU}(\mathbf{W}^g \mathbf{e}^{(r)} + \mathbf{b}^g) \\ \mathbf{f}_m^{(r)} &= \widetilde{\mathbf{e}}^{(r)} \odot \mathbf{u}_m \end{aligned} \tag{3}$$

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.

where $\mathbf{W}^g \in \mathbb{R}^{D \times d^E}$ and $\mathbf{b}^g$ are the parameters of the fully connected layer, $\odot$ is Hadamard product. $\mathbf{f}_m^{(r)}$ is the fused proposal feature. Let $\mathbf{F}^{(r)} = \{\mathbf{f}_1^{(r)}, ..., \mathbf{f}_m^{(r)}\} \in \mathbb{R}^{M \times D}$.

Contextual attention-based proposal interaction. In general, a video is corresponding to various semantic information, which means different proposals may contain one or more related visual concepts in the temporal dimension. As a result, each proposal may have relations to others in content or context. To exploit the relations between the generated proposals [21, 27], we utilize a contextual attention-based module to learn the interaction between different proposals, which aims to capture global and local context information:

$$\begin{aligned}
&\text{CT-Attn}(\mathbf{Q}_F, \mathbf{K}_F, \mathbf{V}_F) = \text{Softmax}(\frac{\mathbf{Q}_F \mathbf{K}_F^\top}{\sqrt{D}})\mathbf{V}_F \\
&\text{head}_i = \text{CT-Attn}(\mathbf{Q}_F = \mathbf{F}\mathbf{W}_i^Q, \mathbf{K}_F = \mathbf{F}\mathbf{W}_i^K, \mathbf{V}_F = \mathbf{F}\mathbf{V}_i^Q) \\
&\widetilde{\mathbf{F}} = \text{MultiHead}(\mathbf{F}) = \text{Concat}(\text{head}_1, ..., \text{head}_H)\mathbf{W}^O
\end{aligned} \tag{4}$$

where $\mathbf{Q}_F$, $\mathbf{K}_F$ and $\mathbf{V}_F$ are the calculated queries, keys and values as input. Here, for simplicity, we omit the superscript of $\mathbf{F}$. $\mathbf{W}_i^Q \in \mathbb{R}^{D \times d_f}$, $\mathbf{W}_i^K \in \mathbb{R}^{D \times d_f}$, $\mathbf{W}_i^V \in \mathbb{R}^{D \times d_f}$, and i denotes the i-th head. H means that the attention is calculated H times. Note that d_f is the number of the output channels. After that, we compute the values for all heads and concatenate them together. Then we use a linear projection matrices $W^O \in \mathbb{R}^{HD \times d_f}$ to obtain the projected multi-head feature $\widetilde{\mathbf{F}}$.

Multi-level fusion. In this section, we design a multi-level module to fuse information from different semantic levels. Then we capture an appropriate representation to do the final prediction as follows:

$$\begin{aligned}
&\widetilde{\mathbf{F}}' = \gamma_s \widetilde{\mathbf{F}}^{(s)} + \gamma_v \widetilde{\mathbf{F}}^{(v)} + \gamma_n \widetilde{\mathbf{F}}^{(n)} \\
&q_1, ..., q_m = \mathbf{Predictor}(\widetilde{\mathbf{F}}')
\end{aligned} \tag{5}$$

where three learnable parameters γ_s, γ_v and γ_n are used to control the balance of different semantic-level features $\widetilde{\mathbf{F}}^{(s)}$, $\widetilde{\mathbf{F}}^{(v)}$, $\widetilde{\mathbf{F}}^{(n)}$. Finally we utilize a predictor to score all proposals with their features $\widetilde{\mathbf{F}}'$. $\{q_1, ..., q_m\}$ are the scores of the proposals. Here, the predictor is a fully connected layer.

3.4. Learning and inference

Considering that the moment proposals have certain length, we select to compute the IoU score o_m for each proposal with the ground truth moment. Furthermore, two thresholds o_{min} and o_{max} are set to calculate the soft label $y_m = \frac{o_m - o_{min}}{o_{max} - o_m}$ for each proposal p_m. Note that if $y_m \leq 0$, we let $y_m = 0$, and we set $y_m = 1$ if $y_m \geq 1$. With the soft labels, we train our model by a binary cross entropy loss as:

$$\mathcal{L} = -\frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} y_m \log q_m + (1 - y_m) \log(1 - q_m) \tag{6}$$

where N means the number of training video-text pairs.

During inference, we rank all proposals relying on the predicted scores and obtain the final retrieval results.

4. EXPERIMENTS

In this section, we evaluate our proposed SV-VMR method for video moment retrieval on two widely-used public datasets: Charades-STA [20] and ActivityNet Captions [28]. The extensive experiments and ablation studies demonstrate the effectiveness of the proposed method.

4.1. Implementation Details

We first follow [14] to generate moment proposals. For visual feature extraction, we employ VGG, C3D [29] and I3D [26] for Charades-STA and C3D [29] for ActivityNet Captions dataset. The feature dimensions d^f and D is set to 512 and d^E is set to 1024. For query encoding, we set the word embedding size as 300 and initialize it with the pretrained Glove embeddings. Then a three layers bi-directional LSTM with 512 hidden units serves for query encoding. The maximum numbers of verb nodes and noun nodes are set to 4 and 6 respectively. We maintain all word tokens after tokenization and truncate all text queries that have maximum 20 words for both datasets. For the contextual attention-based proposal interaction, we apply a multi-head module with 3 heads. The scaling thresholds o_{max} and o_{min} are set to 0.5 and 1.0. We utilize Adam with a learning rate of 1×10^{-4} and a batch size of 32 for optimization.

4.2. Comparison with state-of-the-art methods

Compared methods. DRN (CVPR 20) [6], 2D-TAN (AAAI 20) [14], VSLNet (ACL 20) [2], SM-RL (CVPR 19) [12], ACL (WACV 19) [10], R-W-M (AAAI 19) [19], T-to-C (AAAI 19) [9], SAP (AAAI 19) [30], MAN (CVPR 19) [8], ExCL (EMNLP 19) [23], SCDM (NeurIPS 19) [11], CTRL (ICCV 17) [20]. The comparison results are shown in Table 1 and Table 2 respectively. The best performance is highlighted in **bold** and the second best underline.

Metrics. Following Gao et al. [20], we adopt the evaluation metrics R@n,IoU=m to measure the ability of our approach where we calculate the percentage of at least one of the top-n predicted moments which have Intersection over Union (IoU) larger than m. Specifically, we set n∈{1, 5} with m∈{0.5, 0.7} for Charades-STA dataset and n∈{1, 5} and m∈{0.3, 0.5, 0.7} for ActivityNet Caption dataset.

Charades-STA. We utilize three types of visual features to prove the effectiveness of our model on Charades-STA. For C3D and VGG features, our method outperforms the state-of-the-art methods on most metrics. Compared with T-to-C and MAN, our method obtains the performance gains of (2.49%, 4.18%) and (2.36%, 4.24%) on the important metric R@1,m={0.5, 0.7}, which means our proposed method can locate a video moment proposal more precisely. When using I3D features, our method performs better than other methods on the metric R@1,m=0.5 and R@5,m=0.7 and also obtain comparable results on other metrics. Compared with the state-of-the-art two-stage methods, e.g., 2D-TAN and SCDM, our methods achieves better performance on the important metric R@{1, 5}, IoU=0.5 by gains of (10.07%, 3.98%) and

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.

Table 1: Comparison results on Charades-STA.

Method	Features	R@1		R@5	
		IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
CTRL	C3D	23.63	8.89	58.92	29.52
SM-RL	C3D	24.36	11.17	61.25	32.08
ACL	C3D	30.48	12.20	64.84	35.13
R-W-M	C3D	36.70	-	-	-
T-to-C	C3D	35.60	15.80	79.40	**45.40**
Ours	C3D	**38.09**	**19.98**	**84.05**	40.44
SAP	VGG	27.42	13.36	66.37	38.15
MAN	VGG	41.24	20.54	83.21	**51.85**
2D-TAN	VGG	39.70	23.31	80.32	51.26
Ours	VGG	**43.60**	**24.78**	**83.58**	50.22
ExCL	I3D	44.10	22.40	-	-
DRN	I3D	53.09	31.75	**89.06**	**60.05**
SCDM	I3D	54.44	33.43	74.43	58.08
VSLNet	I3D	54.19	**35.22**	-	-
2D-TAN*	I3D	45.48	25.83	85.03	52.39
Ours	I3D	**55.55**	32.75	89.01	56.18

* We re-train 2D-TAN by using the I3D features with the official code [14].

(1.11%, 14.58%). Although DRN and VSLNet obtain better results than ours on the metric m=0.7, both of them adopt a temporal location regression strategy while our method only uses fixed proposals. We leave this strategy as the future work. Note that the recent high-performance method SQAN [7] achieves the best performance of 59.46% on the metric R@1,m=0.5, however, it can only regress one temporal location for VMR while our SV-VMR can rank multiple moment proposals in a more practical setting. In addition, our proposed method outperforms SQAN on the ActivityNet Captions dataset.

Table 2: Comparison results on ActivityNet Captions.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
CTRL	-	14.00	-	-	-	-
ACRN	-	16.17	-	-	-	-
TGN	45.51	28.47	-	57.32	43.33	-
R-W-M	-	36.90	-	-	-	-
T-to-C	-	27.70	13.60	-	71.85	45.96
DRN	-	**45.45**	24.36	-	**77.97**	50.30
SCDM	54.80	36.75	19.86	77.29	64.99	41.53
2D-TAN	59.45	44.51	26.54	85.53	77.13	61.96
Ours	**61.39**	45.21	**27.32**	**85.98**	77.10	**63.44**

Comparison on ActivityNet Captions. Table 2 reports the video moment retrieval results of various methods. We use C3D features for fair comparisons. We can observe consistent findings compared with Charades-STA dataset. Our proposed SV-VMR outperforms the state-of-the-arts such as DRN, SCDM, and 2D-TAN on most metrics.

4.3. Ablation Studies

In this section, we will perform complete and in-depth ablation studies to evaluate the effect of each component. The experiments are performed on the Charades-STA dataset.

Importance of the semantic structure. To investigate the influence of the semantic role tree in our framework, as shown in Table 3, the impact of progressively adding one type of semantic level information, from top to down, is presented. We can find that when only global-level information(i.e., the whole sentence) is used, the designed baseline achieves the result of (46.57%, 26.08%) on the important metrics R@1, m={0.5, 0.7}. When we add verb-level and noun-level information, the results are improved by (2.03%, 0.75%) and (8.98%, 2.45%). The results demonstrate that each semantic level has a positive effect for the accuracy of VMR task and they cooperate each other to achieve favorable results.

Importance of the visual structure. To verify the importance of our contextual attention-based module, we design a baseline SV-VMR(w/o. CT-Attn) that abandons the contextual attention-based modules in our model and just pass the fused proposal features (Eq. (4)) into our predictor. From Table 4 we can find that SV-VMR(full) outperforms the SV-VMR(w/o. CT-Attn) by absolute gains of (3.70%, 2.99%) on R@1 metric and (2.80%, 2.50%) on R@5 metric. The above results validate the importance of modeling the visual structure and implicate that our contextual attention-based module is useful for the VMR task by adaptively considering the relations among moment proposals.

Table 3: Importance of semantic level information.

sentence	verbs	nouns	R@1		R@5	
			IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
√	×	×	46.57	26.08	86.56	53.82
√	√	×	48.60	27.20	87.31	54.22
√	√	√	55.55	32.75	89.01	56.18

Effect of the multi-head attention. To evaluate the effect of the multi-head module in the contextual attention, we implement a baseline SV-VMR(w/o. multi-head) by replacing the multi-head module with a single-head one in our framework. It is obvious in Table 4 that SV-VMR(w/o. multi-level) obtains comparable results on R@5, m={0.5, 0.7} and it gets a lower performance for the important metric R@1, m={0.5, 0.7} compared with SV-VMR(full).

Effect of the multi-level fusion. The parameters γ_s, γ_v and γ_n are learned to control the balance between the weights of different fused semantic-level features in our model. To explore the influence of the adaptive fusion strategy, we design a baseline SV-VMR(w/o. γ), which simply averages the three semantic-level features as the final features to conduct prediction. From the results in Table 4, we can find that SV-VMR(full) gets a higher accuracy on all metrics compared with SV-VMR(w/o. γ), which demonstrates the importance of the adaptive fusion strategy.

Table 4: Ablation results on Charades-STA dataset.

Methods	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
SV-VMR(w/o. CT-Attn)	51.85	29.76	86.21	53.68
SV-VMR(w/o. multi-head)	52.64	28.17	88.12	54.84
SV-VMR(w/o. γ)	53.25	29.27	86.88	54.73
SV-VMR(full)	55.55	32.75	89.01	56.18

5. CONCLUSIONS

We propose a novel method named SV-VMR for the VMR task, which jointly leverage both semantic and visual struc-

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.

tures in an end-to-end manner. Specifically, a semantic role tree is build to extract the fine-grained semantic information, which is then adopted to facilitate the visual structure learning with a contextual attention-based proposal interaction module. Finally, our model aggregate the multi-level visual-semantic matching information to retrieve target moment accurately. The experimental results demonstrate the effectiveness of our proposed model. In the future, we will consider richer structural information for VMR such as the spatial-temporal structure and knowledge structure. Other strategies can also be utilized to further boost the performance such as proposal regression and re-captioning.

6. REFERENCES

[1] 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 *AAAI*, 2020, pp. 10551–10558.

[2] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou, "Span-based localizing network for natural language video localization," in *ACL*, 2020.

[3] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "Learning to model relationships for zero-shot video classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "Graph convolutional tracking," in *CVPR*, 2019.

[5] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu, "Deep relative tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1845–1858, 2017.

[6] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan, "Dense regression network for video grounding," in *CVPR*, 2020, pp. 10287–10296.

[7] Jonghwan Mun, Minsu Cho, and Bohyung Han, "Local-global video-text interactions for temporal grounding," in *CVPR*, 2020, pp. 10810–10819.

[8] 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 *CVPR*, 2019, pp. 1247–1257.

[9] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *AAAI*, 2019, vol. 33, pp. 9062–9069.

[10] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *WACV*. IEEE, 2019, pp. 245–253.

[11] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *NeurIPS*, 2019, pp. 534–544.

[12] Weining Wang, Yan Huang, and Liang Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *CVPR*, 2019, pp. 334–343.

[13] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo, "Localizing natural language in videos," in *AAAI*, 2019, vol. 33, pp. 8175–8182.

[14] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *AAAI*, 2020, vol. 34, pp. 12870–12877.

[15] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling," in *ACM MM*, 2018.

[16] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *AAAI*, 2019.

[17] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua, "Temporally grounding natural sentence in video," in *EMNLP*, 2018, pp. 162–171.

[18] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *SIGIR*, 2019, pp. 655–664.

[19] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *AAAI*, 2019, vol. 33, pp. 8393–8400.

[20] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia, "Tall: Temporal activity localization via language query," in *ICCV*, 2017, pp. 5267–5275.

[21] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan, "Graph convolutional networks for temporal action localization," in *ICCV*, 2019, pp. 7094–7103.

[22] 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 *ECCV*. Springer, 2020, pp. 156–171.

[23] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann, "ExCL: Extractive Clip Localization Using Natural Language Descriptions," in *NACCL-HLT, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 1984–1990, Association for Computational Linguistics.

[24] Peng Shi and Jimmy Lin, "Simple bert models for relation extraction and semantic role labeling," *arXiv preprint arXiv:1904.05255*, 2019.

[25] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *CVPR*, 2020, pp. 10638–10647.

[26] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[27] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu, "Multi-modality cross attention network for image and sentence matching," in *CVPR*, 2020, pp. 10941–10950.

[28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[30] Shaoxiang Chen and Yu-Gang Jiang, "Semantic proposal for activity localization in videos via sentence query," in *AAAI*, 2019, vol. 33, pp. 8199–8206.

Authorized licensed use limited to: Nanyang Technological University. Downloaded on October 11,2021 at 06:21:28 UTC from IEEE Xplore. Restrictions apply.