VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval

Minuk Ma*, Sunjae Yoon*, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D. Yoo

Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{akalsdnr,dbstjswo505,junyeong.kim,yjlee22,sunghun.kang,cd_yoo}@kaist.ac.kr

Abstract. Video Moment Retrieval (VMR) is a task to localize the temporal moment in untrimmed video specified by natural language query. For VMR, several methods that require full supervision for training have been proposed. Unfortunately, acquiring a large number of training videos with labeled temporal boundaries for each query is a labor-intensive process. This paper explores a method for performing VMR in a weakly-supervised manner (wVMR): training is performed without temporal moment labels but only with the text query that describes a segment of the video. Existing methods on wVMR generate multi-scale proposals and apply query-guided attention mechanism to highlight the most relevant proposal. To leverage the weak supervision, contrastive learning is used which predicts higher scores for the correct video-query pairs than for the incorrect pairs. It has been observed that a large number of candidate proposals, coarse query representation, and one-way attention mechanism lead to blurry attention map which limits the localization performance. To address this issue, Video-Language Alignment Network (VLANet) is proposed that learns a sharper attention by pruning out spurious candidate proposals and applying a multi-directional attention mechanism with fine-grained query representation. The Surrogate Proposal Selection module selects a proposal based on the proximity to the query in the joint embedding space, and thus substantially reduces candidate proposals which leads to lower computation load and sharper attention. Next, the Cascaded Cross-modal Attention module considers dense feature interactions and multi-directional attention flows to learn the multi-modal alignment. VLANet is trained end-to-end using contrastive loss which enforces semantically similar videos and queries to cluster. The experiments show that the method achieves state-of-the-art performance on Charades-STA and DiDeMo datasets.

Keywords: Multi-modal learning, weakly-supervised learning, video moment retrieval

1 Introduction

Video moment retrieval (VMR) is a task to find a temporal moment in untrimmed video specified by a text description as illustrated in Figure 1. With the rising

* Both authors have equally contributed.

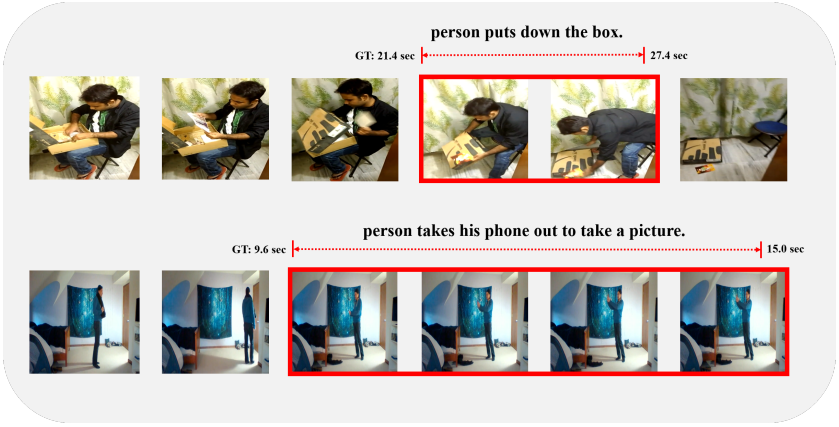


Fig. 1: Illustration of video moment retrieval task. The goal is to search the temporal boundary of the video moment that is most relevant to the given natural language query.

number of videos along with the need for a more detailed and refined search capability that demand a better understanding of the video, the task of Video Moment Retrieval is drawing appreciable attention.

A number of fully-supervised methods that learn from a set of videos with ground-truth time stamps corresponding to a given query have been proposed [3, 6, 23, 25]. For these methods, a large-scale video dataset that requires the laborious burden of temporally annotating the boundaries corresponding to each query is a sine qua non. In general, the performance of a fully-supervised method hinges on the quality of the dataset; however, for VMR, temporal boundaries are often ambiguous to annotate and may act as noise in the learning process.

Recently, weakly-supervised VMR (wVMR) [14, 12] that does not require the temporal boundary annotation for each query has been studied. To leverage the weak supervision, contrastive learning is applied such that higher scores are predicted for the correct video-query pairs than for incorrect pairs. This learning process improves the accuracy of the attention mechanism which plays a vital role in wVMR. Inspired by recent methods [14, 12], this paper addresses two critical challenges: (1) generating appropriate multi-scale video candidate proposals, and (2) learning the latent alignment between the text query and the retrieved video segment.

The first challenge is that the video segment proposals should be adequate in number to give high recall without excessive computational load, and the video segment should be of appropriate length to have high intersection-of-union (IoU) with ground truth. Previous methods [3, 6, 14, 12] greedily generated video candidate proposals using a pre-defined set of multi-scale sliding windows. As a consequence, these methods generally produce large number of multi-scale proposals which increase the chance of achieving high recall at the expense of high computational cost. When an attention mechanism is used thereafter to weigh

the proposals, the attention becomes blurry as there are too many proposals to attend.

The second challenge is to learn a similarity measure between video segment and text query without ground truth annotation. In [14], a text-to-video attention mechanism is incorporated to learn the joint embedding space of video and text query. More accurate multi-modal similarity could be attained with a text query representation that is more effective in interacting with video frame feature. Representing the text query as the last hidden feature of the Gated Recurrent Unit (GRU), as used in some previous methods [14, 12], is overly simplistic . In addition, applying one-way attention from query to video is not sufficient to bring out the most prominent feature in the video and query. Recent studies in Visual Question Answering [24, 5, 9] have explored the possibility of applying multi-directional attention flows that include both inter- and intra-modality attention. This paper devises an analogous idea for the problem of wVMR, and validate its effectiveness in retrieving the moment using the weak labels.

To rise to the challenge, this paper proposes a Video-Language Alignment Network (VLANet) for weakly-supervised video moment retrieval. As a first step, the word-level query representation is obtained by stacking all intermediate hidden features of GRU. Video is divided into overlapping multi-scale segment groups where the segments within each group share a common starting time. Then, the Surrogate Proposal Selection module selects one surrogate from each group which reduces the number of effective proposals for more accurate attention. To consider the multi-directional interactions between each surrogate proposal and query, the Cascaded Cross-modal Attention (CCA) module performs both intra- and inter-modality attention. The CCA module performs self-attention on each modality: video to video (V2V) and query to query (Q2Q), which considers the intra-modal relationships. Thereafter, the CCA module performs cross-modal attention from query to video (Q2V), video to query (V2Q) and finally attended query to attended video (Q2V). This cross-modal attention considers the inter-modal relationships that is critical in learning the multi-modal alignment. To leverage the weak labels of video-query pairs, VLANet is trained in an end-to-end manner using contrastive loss that enforces semantically similar videos and queries to cluster in the joint embedding space. The experiment results show that the VLANet achieves state-of-the-art performance on Charades-STA and DiDeMo datasets. Extensive ablation study and qualitative analyses validate the effectiveness of the proposed method and provide interpretability.

2 Related Work

2.1 Temporal Action Detection

The goal of temporal action detection is to predict the temporal boundary and category for each action instance in untrimmed videos. Existing works are divided into two groups: the fully-supervised and weakly-supervised. Zhao *et al.* [26] proposed a structured segment network that models the temporal structure

of each action instance by a structured temporal pyramid. Gao *et al.* [4] proposed Cascaded Boundary Regression which uses temporal coordinate regression to refine the temporal boundaries of the sliding windows. Lin *et al.* [11] proposed Boundary Sensitive Network that first classifies each frame as the start, middle, or end, then directly combines these boundaries as proposals.

In the weakly-supervised settings, however, only the coarse video-level labels are available instead of the exact temporal boundaries. Wang *et al.* [22] proposed UntrimmedNet that couples two components, the classification module, and the selection module, to learn the action models and reason about the temporal duration of action instances, respectively. Nguyen *et al.* [15] proposed a Sparse Temporal Pooling Network that identifies a sparse subset of key segments associated with the target actions in a video using an attention module and fuse the key segments using adaptive temporal pooling. Shou *et al.* [17] proposed AutoLoc that uses Outer-Inner-Contrastive loss to automatically discover the required segment-level supervision to train a boundary predictor. Liu *et al.* [13] proposed CleanNet that leverages an additional temporal contrast constraint so that the high-evaluation-score action proposals have a higher probability to overlap with the ground truth action instances.

2.2 Video Moment Retrieval

The VMR task is focused on localizing the temporal moment that is semantically aligned with the given natural language query. For this task, various supervised methods have been proposed [3, 6, 23, 25]. In Gao *et al.* [3] and Hendricks *et al.* [6], candidate moments are sampled using sliding windows of various lengths, and multi-modal fusion is performed to estimate the correlation between the queries and video moments. Xu *et al.* [23] proposed a model that integrates vision and language features using attention mechanisms and leverages video captioning as an auxiliary task. Zhang *et al.* [25] proposed Moment Alignment Network (MAN) that considers the relationships between proposals as a structured graph, and devised an iterative algorithm to train a revised graph convolution network.

Recently, the task was studied under the weakly-supervised setting [2, 14, 12]. Duan *et al.* [2] proposed to decompose weakly-supervised dense event captioning in videos (WS-DEC) into a pair of dual problems: event captioning and sentence localization. They proposed a cycle system to train the model based on the assumption that each caption describes only one temporal segment. Mithun *et al.* [14] proposed Text-Guided-Attention (TGA) model that learns a joint representation between video and sentence. The attention weight is used to retrieve the relevant moment at test time. Lin *et al.* [12] proposed Semantic Completion Network (SCN) that selects the top-K proposals considering exploration and exploitation, and measures the semantic similarity between the video and query. As an auxiliary task, SCN takes the masked sentence as input and predicts the masked words from visual representations.

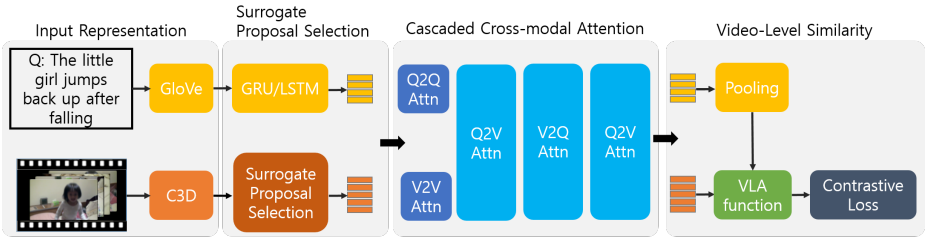


Fig. 2: Illustration of VLANet architecture. The Surrogate Proposal Selection module prunes out irrelevant proposals based on the similarity metric. Cascaded Cross-modal Attention considers various attention flows to learn multi-modal alignment. The network is trained end-to-end using contrastive loss.

3 Method

3.1 Method Overview

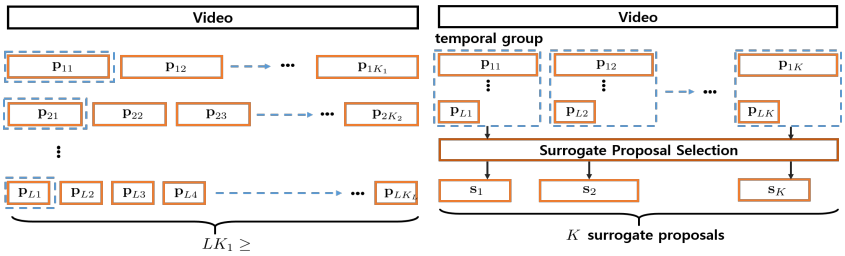
Figure 2 illustrates the overall VLANet architecture. The input text query is embedded using GloVe [16] after which each embedded representation is fed into a GRU [1]. In the meanwhile, the video is embedded based on C3D [21]. Video is divided into overlapping multi-scale segment groups where the proposals within each group share a common starting time. Given the video and query representations V and Q, the similarity c between video and query is evaluated by the Cascaded Cross-modal Attention (CCA) module. The learned attention weights by CCA are used to localize the relevant moment at test time. A video-query pair (V, Q) is positive if it is in the training data; otherwise, it is negative. The network is trained in an end-to-end manner using contrastive loss to enforce the scores of the positive pairs to be higher than those of the negative pairs. In practice, the negative pairs are randomly sampled in a batch.

3.2 Input representation

Query representation Gated Recurrent Unit (GRU) [1] is used for encoding the sentences. Each word of the query is embedded using GloVe and sequentially fed into a GRU. Prior methods [14] use only the final hidden feature of GRU to represent the whole sentence, which leads to the loss of information by excluding the interactions between frame- and word-level features of video and query. Motivated by recent works in visual question answering [5, 24], this paper uses all intermediate hidden features of the GRU. The query Q is represented as:

$$Q = [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_M] \tag{1}$$

where $\mathbf{w}_m \in \mathbb{R}^D$ denotes the m-th GRU hidden feature, and D is the dimension of the hidden feature. Each $\mathbf{w}_m$ is L2 normalized to output a unit vector.



(a) Multi-scale proposal generation (b) Surrogate Proposal Selection

Fig. 3: Comparison between the previous and the proposed proposal generation method. (a) generates large number of proposals of various lengths. (b) groups the proposals, and selects the surrogate proposals based on the proximity to the query.

Video representation Video is encoded using a C3D [21] model pre-trained on Sports-1M dataset [8] as in [3]. The feature was extracted at every 16 frames for Charades-STA. The VGG16 model [19] is used for frame-level feature extraction for DiDeMo dataset following [6]. Both C3D and VGG16 features were extracted from the penultimate fully-connected layer, which results in the feature dimension of 4096.

Video proposal generation As depicted in image Figure 3(a) previous methods [14, 12] generated proposals using multi-scale sliding windows. Meanwhile, as in Figure 3(b), VLANet organizes the multi-scale windows in segment groups such that within a group, all windows start at the same time instance. Each group will have the same number of windows of fixed scales. The interval between the starting times of each segment group is regular. With K segment groups and L multi-scale proposals, the total number of proposals is $K \cdot L$. Then, the video V is represented by:

$$V = \begin{bmatrix} \mathbf{p}_{11} & \mathbf{p}_{12} & \cdots & \mathbf{p}_{1K} \\ \mathbf{p}_{21} & \mathbf{p}_{22} & \cdots & \mathbf{p}_{2K} \\ \vdots & & \cdots & \\ \mathbf{p}_{L1} & \mathbf{p}_{L2} & \cdots & \mathbf{p}_{LK} \end{bmatrix} \tag{2}$$

where each $\mathbf{p}_{lk} \in \mathbb{R}^D$ denotes the proposal feature of the l-th scale in the k-th segment group, which is the average of the C3D features of the frames participating in the proposal. Fully-connected layers are used to resize the feature dimension of Q and V to D. L2 normalization is performed to make each $\mathbf{p}_{lk}$ a unit vector.

3.3 Surrogate Proposal Selection module

To reduce the large number of proposals, [12] proposed a sampling-based selection algorithm to prune out irrelevant proposals considering the exploration and exploitation. However, the method is trained using policy gradient algorithm [20] which suffers from high variance. Instead, as depicted in Figure 3(b), the Surrogate Proposal Selection module selects the best-matched proposals from each segment group based on the cosine similarity to the final hidden feature of the query. A surrogate proposal of the k-th segment group is defined as the proposal that has the largest cosine similarity to the final hidden feature of the query. The cosine similarity between each proposal and query is given by

$$\begin{bmatrix} \mathbf{p}_{11} \cdot \mathbf{w}_M & \mathbf{p}_{12} \cdot \mathbf{w}_M & \cdots & \mathbf{p}_{1K} \cdot \mathbf{w}_M \\ \mathbf{p}_{21} \cdot \mathbf{w}_M & \mathbf{p}_{22} \cdot \mathbf{w}_M & \cdots & \mathbf{p}_{2K} \cdot \mathbf{w}_M \\ \vdots & & \cdots & \\ \mathbf{p}_{L1} \cdot \mathbf{w}_M & \mathbf{p}_{L2} \cdot \mathbf{w}_M & \cdots & \mathbf{p}_{LK} \cdot \mathbf{w}_M \end{bmatrix} \tag{3}$$

where $\mathbf{w}_M$ is the final hidden feature of the query. It is empirically determined that the final hidden query feature is sufficient in pruning out irrelevant proposals at a low computational cost. The Surrogate Proposal Selection module pick the l'-th scale from each k-th segment group which is given by,

$$l' = \text{argmax}\left[\mathbf{p}_{1k} \cdot \mathbf{w}_M \;\; \mathbf{p}_{2k} \cdot \mathbf{w}_M \;\; \cdots \;\; \mathbf{p}_{Lk} \cdot \mathbf{w}_M\right], \tag{4}$$

$$\mathbf{s}_k = \mathbf{p}_{l'k} \tag{5}$$

where $\mathbf{s}_k$ is the surrogate proposal feature of the k-th segment group. In back-propagation, only the surrogate proposals $\mathbf{s}_k$'s contribute to the weight update which allows end-to-end learning. Then the video is represented by K surrogate proposal features:

$$\mathcal{V} = [\mathbf{s}_1 \;\; \mathbf{s}_2 \;\; \cdots \;\; \mathbf{s}_K] \tag{6}$$

where $\mathcal{V}$ is the updated video representation composed of the surrogate proposals.

3.4 Cascaded Cross-modal Attention module

Cascaded Cross-modal Attention (CCA) module takes the video and query representations as inputs, and outputs a compact attended video representation. Compared to text-guided attention (TGA) [14], CCA module considers more diverse multi-modal feature interactions including V2V, Q2Q, V2Q, and Q2V where each has its own advantages as described below.

Dense Attention The basic attention unit of CCA module is referred to as Dense Attention which calculates the attention between two multi-element features. Given $Y = [\mathbf{y}_1 \dots \mathbf{y}_M]^T \in \mathbb{R}^{M \times D}$ and $X = [\mathbf{x}_1 \dots \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$, the Dense Attention $A(X, Y) : \mathbb{R}^{N \times D} \times \mathbb{R}^{M \times D} \rightarrow \mathbb{R}^{N \times D}$ attends X using Y and is defined as follows:

$$\mathcal{E}(\mathbf{x}_n, Y) = \textstyle\sum_{m=1}^{M} \tanh(W_1 \mathbf{x}_n \cdot W_2 \mathbf{y}_m), \tag{7}$$

$$A(X, Y) = \text{Softmax}([\mathcal{E}(\mathbf{x}_1, Y)\ \mathcal{E}(\mathbf{x}_2, Y)\ \cdots\ \mathcal{E}(\mathbf{x}_N, Y)])X, \tag{8}$$

where W_1, W_2 are learnable parameters. Here, $\mathcal{E} : \mathbb{R}^D \times \mathbb{R}^{M \times D} \rightarrow \mathbb{R}$ is referred to as the Video-Language Alignment (VLA) function that performs the multi-modal alignment.

Self-attention Based on the Dense Attention defined above, the CCA module initially performs a type of self-attention that attends $\mathcal{V}$ and Q using $\mathcal{V}$ and Q respectively as given below,

$$\mathcal{V} \leftarrow A(\mathcal{V}, \mathcal{V}), \tag{9}$$

$$Q \leftarrow A(Q, Q). \tag{10}$$

The intra-attention allows each element of itself to be attended by its global contextual information. The attention from $\mathcal{V}$ to $\mathcal{V}$ is capable of highlighting the salient proposals by considering the innate temporal relationships. The attention from Q to Q updates the each word-level feature by considering the context of the whole sentence.

Cross modal attention Following self-attention defined above, the CCA module is used to cross-attend $\mathcal{V}$ and Q using Q and $\mathcal{V}$ respectively such that cross-modal attention is defined as follows:

$$\mathcal{V} \leftarrow A(\mathcal{V}, Q), \tag{11}$$

$$Q \leftarrow A(Q, \mathcal{V}). \tag{12}$$

The above attention is critical in learning the latent multi-modal alignment. It has been empirically observed that cross-modal attention applied in series several times until near-saturation can be conducive in producing better performance. Finally, a compact attended video representation $\mathbf{v}_{comp}$ is obtained by taking the sum of all elements of $\mathcal{V}$, and video-level similarity c is obtained by the VLA function between $\mathbf{v}_{comp}$ and Q as given below:

$$c = \mathcal{E}(\mathbf{v}_{comp}, Q). \tag{13}$$

The network is trained using the following contrastive loss:

$$\mathcal{L}_{contrastive} = max[0, \Delta - \mathcal{E}(\mathbf{v}_{comp}, Q^+) + \mathcal{E}(\mathbf{v}_{comp}, Q^-)] \tag{14}$$

where $\mathcal{E}$ is the VLA function defined above in section 3.4 and Δ is the margin. Q^+ and Q^- is positive and negative query features.

4 Experiment

4.1 Datasets

Charades-STA The Charades dataset was originally introduced in [18]. It contains temporal activity annotation and multiple video-level descriptions for each video. Gao *et al.* [3] generated temporal boundary annotations for sentences using a semi-automatic way and released the Charades-STA dataset that is for video moment retrieval. The dataset includes 12,408 video-sentence pairs with temporal boundary annotations for training and 3,720 for testing. The average length of the query is 8.6 words, and the average duration of the video is 29.8 seconds.

DiDeMo The Distinct Describable Moments (DiDeMo) dataset [6] consists of over 10,000 unedited, personal videos in diverse visual settings with pairs of localized video segments and referring expressions. The videos are collected from Flickr and each video is trimmed to a maximum of 30 seconds. The dataset includes 8,395, 1,065 and 1,004 videos for train, validation, and test, respectively. The videos are divided into 5-second segments to reduce the complexity of annotation, which results in 21 possible moments per video. The dataset contains a total of 26,892 moments with over 40,000 text descriptions. The descriptions in the DiDeMo dataset are natural language sentences that contain activities, camera movement, and temporal transition indicators. Moreover, the descriptions in DiDeMo are verified to refer to a single moment.

Evaluation Metric For Charades-STA, the evaluation metric proposed by [3] is adopted to compute "R@n, IoU=m". For the test set predictions, the recall R@n calculates the percentage of samples for which the correct result resides in the top-n retrievals to the query. If the IoU between the prediction and the ground truth is greater than or equal to m, the prediction is correct. The overall performance is the average recall on the whole test set.

For DiDeMo, the evaluation metric proposed by [6] is adopted. The evaluation metric is also R@n with different criteria for correct prediction. If the ground truth moment is in the top-n predictions, the prediction for the sample is counted as correct. The mIoU metric is computed by taking the average of the IoU between the predicted moment and the ground truth moment.

Table 1: Performance comparison of VLANet to the related methods on Charades-STA

Type	Method	R@1			R@5		
		IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
Baseline	Random	19.78	11.96	4.81	73.62	52.79	21.53
Fully	VSA-RNN [3]	-	10.50	4.32	-	48.43	20.21
	VSA-STV [3]	-	16.91	5.81	-	53.89	23.58
	CTRL [3]	-	23.63	8.89	-	58.92	29.52
	EFRC [23]	53.00	33.80	15.00	94.60	77.30	43.90
	MAN [25]	-	46.53	22.72	-	86.23	53.72
Weakly	TGA [14]	32.14	19.94	8.84	86.58	65.52	33.51
	SCN [12]	42.96	23.58	9.97	95.56	71.80	**38.87**
	VLANet (ours)	**45.24**	**31.83**	**14.17**	**95.70**	**82.85**	33.09

Table 2: Performance comparison of VLANet to the related methods on DiDeMo

Type	Method	R@1	R@5	mIoU
Baseline	Upper Bound	74.75	100	96.05
	Random	3.75	22.50	22.64
	LSTM-RGB-Local [6]	13.10	44.82	25.13
Fully	Txt-Obj-Retrieval [7]	16.20	43.94	27.18
	EFRC [23]	13.23	46.98	27.57
	CCA [10]	18.11	52.11	37.82
	MCN [6]	28.10	78.21	41.08
	MAN [25]	27.02	81.70	41.16
Weakly	TGA [14]	12.19	39.74	24.92
	VLANet (ours)	**19.32**	**65.68**	**25.33**

4.2 Quantitative result

Table 1 shows the performance comparison between VLANet and the related methods on Charades-STA. The first section indicates random baseline, the second section indicates fully-supervised methods, and the third section indicates weakly-supervised methods. VLANet achieves state-of-the-art performance on Charades-STA among weakly-supervised methods. It outperforms the random baseline, VSA-RNN, and VSA-STV by a large margin. Compared to the other fully-supervised methods such as CTRL and EFRC, its performance is comparable. Besides, compared to the other weakly-supervised methods TGA and SCN, VLANet outperforms others by a large margin.

Table 2 shows the performance comparison on DiDeMo. The first section contains the baselines, the second section contains fully-supervised methods, and the third section contains weakly-supervised methods. VLANet achieves state-of-the-art performance among the weakly-supervised methods. In the R@5 based test, especially, its performance is 25.94 higher than the runner-up model TGA.

Table 3: Performance of model variants and ablation study of VLANet on Charades-STA. The unit of stride and window size is frame.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
stride 4, window(176, 208, 240)	44.76	31.53	14.78	77.04	63.17	31.80
stride 6, window(176, 208, 240)	42.17	28.60	12.98	88.76	74.91	**34.70**
stride 8, window(176, 208, 240)	45.03	31.82	14.19	95.72	82.82	33.33
stride 6, window(128, 256)	42.39	28.03	13.09	94.70	73.06	30.69
stride 6, window(176, 240)	42.92	30.24	13.57	95.72	82.80	33.46
w/o cross-attn	43.41	30.08	13.23	95.72	82.41	33.06
w/o self-attn	42.31	30.81	15.38	95.38	80.02	33.76
w/o surrogate	35.81	25.30	12.26	80.61	64.57	31.31
full model	**45.03**	**31.82**	**14.19**	**95.72**	**82.82**	33.33

It is comparable to some fully-supervised methods such as CCA [1] and Txt-Obj-Retrieval. These indicate that even without the full annotations of temporal boundary, VLANet has the potential to learn latent multi-modal alignment between video and query, and to localizing semantically relevant moments.

4.3 Model variants and ablation study

Table 3 summarizes the performance of model variants and the ablation study conducted on VLANet. The first section shows the performance variation by varying stride and window sizes, and the second section shows the performance drop without core components. The strides and the sizes of the windows were determined by considering the average video length. The first three rows show that the network performs best with the stride of 8. While the proposals with stride 4 have finer granularity, the large number of proposals decreases the performance. The networks with three multi-scale proposals tend to achieve higher performance than the networks with two multi-scale proposals. This shows the importance of stride and the number of scales. After finding the best hyper-parameters of 'stride 8, window(176, 208, 240)' these values were fixed for the subsequent experiments and analyses. The network without cross-attention, self-attention show a decrease in performance, demonstrating the importance of the attention mechanisms. We generally notice a drop in performance with an increasing IoU metric. The drop is more drastic without cross-attention than without self-attention. This observation indicates that cross-modal attention has a larger influence on performance than self-attention. The performance of w/o surrogate is decreased significantly across all metrics. This indicates that selecting probable proposals in the early stage is critical to the performance.

[1] Here, CCA refers to a previous method [10], but not Cascaded Cross-modal Attention proposed in this paper.

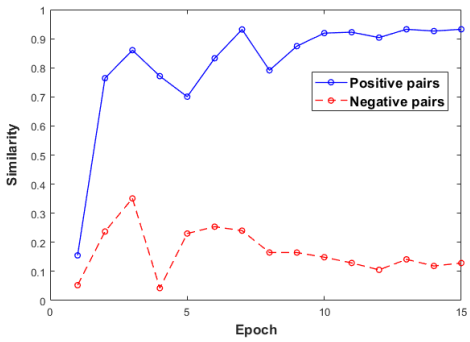


Fig. 4: The multi-modal similarity prediction by VLANet on the positive and negative pairs while training. The similarity gap increases as epoch increases.

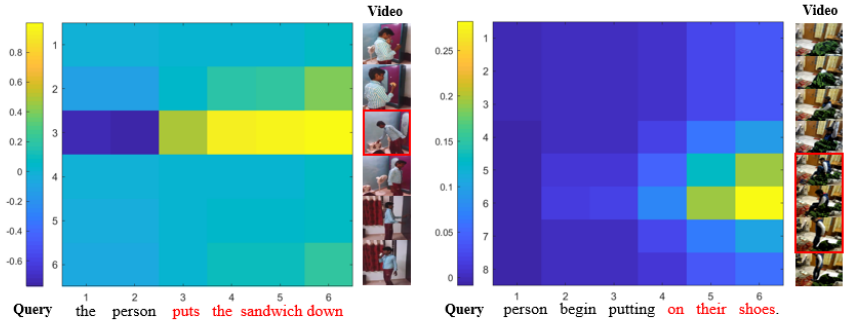


Fig. 5: Visualization of Cascaded Cross-modal Attention. The attention map is calculated by the outer-product of video and query features that are obtained after the Cascaded Cross-modal Attention module and before the pooling layer.

4.4 Analysis of multi-modal similarity

Figure 4 shows similarity predicted by the network on the whole test set of Charades-STA while training. The x-axis indicates the epoch of training, and the y-axis indicates the similarity. It is observed that the similarity scores of the positive pairs (blue) increase and reach a high plateau of about 0.9, while those of the negative pairs (red) keep a low value of about 0.15. These demonstrate that contrastive learning was successfully conducted.

4.5 Visualization of attention map

Figure 5 visualizes the attention map of the proposed Cascaded Cross-modal Attention. The x-axis indicates the words in the query and the y-axis indicates the time. In the left example, the attention weight of the "put the sandwich

Fig. 6: At inference time, VLANet successfully retrieves the moment described by the query. Due to the limited space, only some proposals are visualized. The color indicates the attention strength. The top-2 predicted moments are visualized with the temporal boundaries.

down" is high when the person is putting the sandwich down. Similarly in the right example, important words such as action or object have high attention weight with the related moment of the video. The high attention weights are biased on the right side in Figure 5 as the final GRU feature has the context information about the whole sentence. The above example demonstrates that VLANet can learn the latent multi-modal alignment.

4.6 Visualization of inference

Figure 6 provides a visualization of the inference of VLANet. Only a subset of total proposals were depicted whose color indicates the attention strength. In the first example, both top-1 and top-2 predictions by VLANet have high overlaps with the ground truth moment. In the second example, the network localizes the moment when the person actually *opens* the refrigerator. Similarly in the

third example, the network localizes the moment when person *puts* the pillow. This shows that the network successfully captures the moment when a certain action is taken or an event occurs. The inference visualization demonstrates the moment retrieval ability of VLANet and suggests its applicability to real-world scenarios.

5 Conclusions

This paper considers Video-Language Alignment Network (VLANet) for weakly-supervised video moment retrieval. VLANet is able to select appropriate candidate proposals using a more detailed query representation that include intermediate hidden features of the GRU. The Surrogate Proposal Selection module reduces the number of candidate proposals based on the similarity between each proposal and the query. The ablation study reveals that it has the largest influence on performance. The Cascaded Cross-modal Attention module performs a modified self-attention followed by a cascade of cross-attention based on the Dense Attention defined. It also has a significant influence on performance. VLANet is trained in an end-to-end manner using contrastive loss which enforces semantically similar videos and queries to cluster in the joint embedding space. The experiments shows that VLANet achieves state-of-the-art performance on Charades-STA and DiDeMo datasets.

Acknowledgement

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding) and partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data)

References

1. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Proceedings of SSST@EdMNLP 2014. pp. 103–111. Association for Computational Linguistics (2014)
2. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018. pp. 3063–3073 (2018)
3. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 5277–5285. IEEE Computer Society (2017)
4. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. In: British Machine Vision Conference 2017, BMVC 2017. BMVA Press (2017)
5. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C.H., Wang, X., Li, H.: Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. pp. 6639–6648. Computer Vision Foundation / IEEE (2019)
6. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with natural language. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 5804–5813. IEEE Computer Society (2017)
7. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4555–4564 (2016)
8. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014. pp. 1725–1732. IEEE Computer Society (2014)
9. Kim, J., Ma, M., Pham, T., Kim, K., Yoo, C.D.: Modality shifting attention network for multi-modal video question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
10. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4437–4446 (2015)
11. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: BSN: boundary sensitive network for temporal action proposal generation. In: Computer Vision - ECCV 2018 - 15th European Conference. vol. 11208, pp. 3–21. Springer (2018)
12. Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. CoRR **abs/1911.08199** (2019), http://arxiv.org/abs/1911.08199
13. Liu, Z., Wang, L., Zhang, Q., Gao, Z., Niu, Z., Zheng, N., Hua, G.: Weakly supervised temporal action localization through contrast based evaluation networks. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
14. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. pp. 11592–11601. Computer Vision Foundation / IEEE (2019)

15. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. pp. 6752–6761. IEEE Computer Society (2018)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014. pp. 1532–1543. ACL (2014)
17. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany. vol. 11220, pp. 162–179. Springer (2018)
18. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision - ECCV 2016 - 14th European Conference. vol. 9905, pp. 510–526. Springer (2016)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015 (2015)
20. Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. p. 1057–1063. MIT Press (1999)
21. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015. pp. 4489–4497. IEEE Computer Society (2015)
22. Wang, L., Xiong, Y., Lin, D., Gool, L.V.: Untrimmednets for weakly supervised action recognition and detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 6402–6411. IEEE Computer Society (2017)
23. Xu, H., He, K., Sigal, L., Sclaroff, S., Saenko, K.: Text-to-clip video retrieval with early fusion and re-captioning. ArXiv **abs/1804.05113** (2018)
24. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. pp. 6281–6290. Computer Vision Foundation / IEEE (2019)
25. Zhang, D., Dai, X., Wang, X., Wang, Y., Davis, L.S.: MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. pp. 1247–1257. Computer Vision Foundation / IEEE (2019)
26. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. International Journal of Computer Vision **128**(1), 74–95 (2020)