Weakly Supervised Dense Video Captioning via Jointly Usage of Knowledge Distillation and Cross-modal Matching

Bofeng Wu[1,2], **Guocheng Niu**[2], **Jun Yu**[1*], **Xinyan Xiao**[2], **Jian Zhang**[3] and **Hua Wu**[2]
[1]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
[2]Baidu Inc., Beijing, China
[3]Zhejiang International Studies University, Hangzhou, China
{wubofeng, yujun}@hdu.edu.cn, {niuguocheng, xiaoxinyan,wu_hua}@baidu.com,
jayzhang@outlook.com

Abstract

This paper proposes an approach to Dense Video Captioning (DVC) without pairwise event-sentence annotation. First, we adopt the knowledge distilled from relevant and well solved tasks to generate high-quality event proposals. Then we incorporate contrastive loss and cycle-consistency loss typically applied to cross-modal retrieval tasks to build semantic matching between the proposals and sentences, which are eventually used to train the caption generation module. In addition, the parameters of matching module are initialized via pretraining based on annotated images to improve the matching performance. Extensive experiments on ActivityNet-Caption dataset reveal the significance of distillation-based event proposal generation and cross-modal retrieval-based semantic matching to weakly supervised DVC, and demonstrate the superiority of our method to existing state-of-the-art methods.

1 Introduction

Dense Video Captioning (DVC) [Krishna *et al.*, 2017] refers to detecting and describing multiple events from a given video. Each event is represented by a video segmentation reflecting distinct semantic with starting and ending timestamps in the video and the semantic should be expressed by a sentence constructed by lexical words. Being able to deliver fine-grained semantic of a long video sequence, DVC boosts the applications of techniques like content-based video retrieval, personalized recommendation and intelligent video surveillance. Most of the state-of-the-art approaches to DVC adopt a fully supervised learning pipeline [Krishna *et al.*, 2017; Zhou *et al.*, 2018; Xiong *et al.*, 2018; Mun *et al.*, 2019], which means in addition to the videos, the lexical description of each involved event with its starting and ending timestamps should also be known before training. However, locating events in each video and tagging each event with appropriate sentence is extremely labor-intensive, given mass amount of videos.

In the past several years, weakly supervised DVC methods [Shen *et al.*, 2017; Duan *et al.*, 2018] emerged to achieve event localization and sentence captioning for each event, which only needs videos as well as lexical descriptions of the whole video sequence as training data. Therefore, weakly supervised DVC is much more challenging than supervised DVC due to the lack of event-level annotation. Shen et al. [Shen *et al.*, 2017] mapped sentence lexical words to frame regions and constructed region-based sequences, then associated lexical words to sequences. Region-based sequences describe more object-level than event-level semantic, so this method is insufficient to generate proposals of events. Duan et al. [Duan *et al.*, 2018] firstly extracted video and caption features using RNNs, then associated these features by cross-attention multi-modal fusion process for event clip generation, the training of captioning module simultaneously exploits event clips and captions. Their method needs captions to localize clips, but the captions are unavailable during the testing phase, so their clips only rely on random generation, which is inaccurate. So far as we know, there is still no satisfactory solution to event-caption generation in current weakly supervised DVC.

This paper proposes a novel weakly supervised DVC pipeline towards high-quality event proposal generation and accurate caption generation. This pipeline consists of three modules: the distillation-based proposal generation, the proposal-caption matching by cross-modal retrieval and the event caption generation. The proposal generation module detects a group of candidate proposals for event depiction, and is optimized with soft label constraints and intermediate feature constraints, which are distilled from several teacher networks designed for similar tasks. The proposal-caption matching module selects one proposal for each sentence to formulate pairwise data. The matching module individually extracts video and sentence features through the encoder of a transformer followed by an attention block, and performs semantic alignment between visual and lexical features using contrastive loss and cycle-consistency loss. Apart from this, images with captions are used to construct pseudo dense video with captions to pretrain the matching module to obtain good initialization parameters. The event caption generation module adopts Attention Long Short-Term Memory (Attention-LSTM) [Anderson *et al.*, 2018] to achieve the sentence generation for selected proposals. Compared with end-

*Jun Yu is the corresponding author

to-end methods, each module in our pipeline is trained independently, which enables performance promotion of the pipeline via individual optimization of each module. In the meanwhile, positive interactions between different modules also improve the performance of the pipeline. For example, high-quality proposal-sentence matching improves the proposal and caption generation results, while good proposal generation benefits the matching module.

The contributions of this paper are three-fold:

- We propose an efficient pipeline to tackle the weakly supervised DVC task. In particular, our pipeline innovatively introduces a cross-modal matching module to achieve proposal-caption matching, which is currently ignored by other methods. With the help of the proposed matching module, the remaining two modules (i.e. the proposal generation module and the caption generation module) are effectively integrated to better deal with this task. We conduct comprehensive experiments to empirically analyze our method on Activity-Caption dataset. The experimental results demonstrate the superiority of our proposed approach to existing state-of-the-art methods.
- We adopt knowledge distillation techniques to conduct event proposal generation. By distilling the knowledge from other well trained teacher networks and using the well designed learning strategy, the proposal generation module can make full use of proposal data from other fields and it is quiet flexible. Experiments show that our multi-teachers knowledge distillation learning method achieves outstanding results.
- We introduce a novel cross-modal retrieval mechanism into the proposal-caption matching. Specifically, we exploit contrastive loss and cycle-consistency loss to optimize cross-modal matching. It is worthy noting that we also improve matching performance via pretraining the network using preprocessed images with ground truth captions.

2 Related Works

2.1 Proposal Generation

The event proposal generation plays an important role in the DVC tasks. [Krishna *et al.*, 2017] adopted Deep Action Proposals (DAP) model [Escorcia *et al.*, 2016], which generated variable-length proposals based on the clustered ground truth annotation. [Xiong *et al.*, 2018] exploited Structured Segment Network (SSN) [Zhao *et al.*, 2017], which boosted the accuracy of proposals by modeling their temporal structure via a structured temporal pyramid. Similarly, [Zhou *et al.*, 2018] used multi-layer transformer encoder to generate proposals from visual features. Single-Stream Temporal (SST) [Buch *et al.*, 2017] adopted in [Mun *et al.*, 2019] constructed candidate proposals using GRU hidden states with various lengths and individual confidence score for proposal refinement. *However, these methods usually cannot generate proposals with precise starting and ending boundaries.* Bound Sensitive Network (BSN) proposed in [Lin *et al.*, 2019] defined candidate proposal boundaries as frames with high probability and adopted confidence score to achieve proposal generation. The state-of-the-art Boundary-Matching Network (BMN) [Lin *et al.*, 2019] further developed BSN into an end-to-end framework.

In weakly supervised DVC scenario, the lacking of event timestamps annotation turns proposal generation into an unsupervised problem, therefore the above mentioned methods are invalid under this challenging situation. Integrating the knowledge supplied from other similar tasks can mitigate this problem, which can be regarded as a knowledge distillation strategy [Hinton *et al.*, 2015].

2.2 Proposal Matching

Given generated proposals as well as sentence-level captions obtained from annotation, the proposal caption generation is still unavailable because there is no correspondence between the proposals and captions. To this end, a cross-modal matching process is necessary, which often finds its usage in cross-modal retrieval task.

[Peng *et al.*, 2019] proposed to project the data of different modalities into one common space, in which the projection embraces the geometric consistency (GC) and the cluster assignment consistency (CAC). [Song and Soleymani, 2019] extracted modality-specific features through self-attention and residual learning. The model was optimized by minimizing the distance between relevant video and text representations. [Gabeur *et al.*, 2020] adopted multi-modal transformer and BERT to learn video and text features respectively, and used bi-directional loss to rank the cross-modal similarity. [Ging *et al.*, 2020] employed transformer and attention for modality-specific feature extraction, and performed multi-level cross-modal semantic alignment reinforced by a cross-modal cycle-consistency loss, which enabled this approach to yield state-of-the-art results. [Yu *et al.*, 2020] proposed a memory attention mechanism to identify the critical visual representations related to the language representations, which also similar to multi-modal matching.

3 The Proposed Method

Weakly supervised DVC detects N video clips $\{c_i\}_{i=1}^{N}$ (each can be described as a proposal with both visual features and starting/ending timestamps) corresponding to certain events from a whole video sequence V, and describes c_i with a lexical sentence $s_i = \{w_i^j\}_{j=1}^{M_i}$ containing M_i words w_j. The training data in weakly supervised DVC task only include full video sequences and their lexical descriptions called paragraphs. In the rest part of the paper, we use V and c to represent visual features (V denotes video-level features and c denotes clip-level features), and adopt s to describe lexical features. The proposed method is presented in Figure 1 which consists of three modules: a) knowledge distillation based proposal generation (KDPG, Section 3.1) generates candidate event proposals about certain events in the video; b) proposal-caption matching(PCM, Section 3.2) selects a corresponding event proposal from candidates for each sentence; and c) event caption generation (ECG, Section 3.3) generates caption for each event proposal.

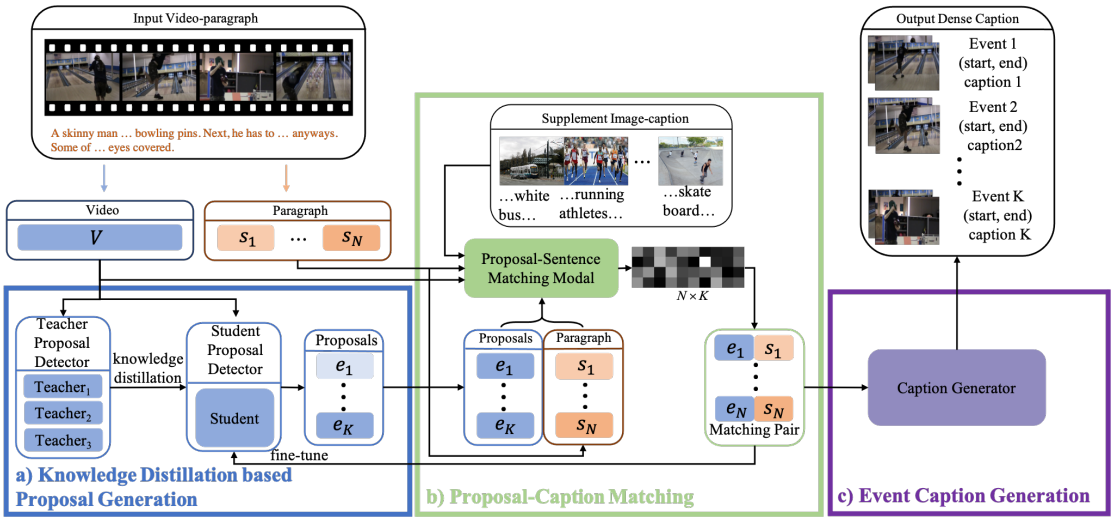

Figure 1: Overview of the proposed pipeline: a) Proposal Generation detects valid event proposals based on Knowledge Distillation in an untrimmed video; b) Proposal-caption Matching selects one event proposal for each sentence by matching between all event proposals and sentences; c) Event Caption Generation module generates captions for each proposal.

3.1 Knowledge Distillation based Proposal Generation

Generating accurate event proposal is very difficult in weakly supervised scenario. This enlightens us to introduce knowledge learned from relevant problems that have been well solved. Specifically, we resort to network distillation. We train several teacher networks on other proposal detection tasks (in this paper, i.e. activity, action, and video highlight) in fully supervised way, and use soft labels as well as intermediate features produced by teacher networks to train our student network.

Teacher Network

modifies the BMN model proposed in [Lin *et al.*, 2019] via changing the CNN into Temporal Convolutional Network (TCN) [Lea *et al.*, 2017] to improve the temporal information extraction ability. We train teacher networks on the aforementioned three tasks. Given a video, teacher networks output soft-labels $\{O_i\}_{i=1}^{n}$ (n is the teacher number) comprising probability scores that each frame being a starting/ending point of a proposal, and the confidence scores that each pair of the potential starting and ending frames being a proposal. These proposals are selected according to the ranking of their scores computed based on probability scores and confidence scores. Also, the teacher networks extracts visual encoding features $\{V_i\}_{i=1}^{n}$.

Student Network

adopts the same network structure as teachers. In training, student network learns visual features V' and labels O' under the guidance of soft labels $\{O_i\}_{i=1}^{n}$ and intermediate features $\{V_i\}_{i=1}^{n}$ produced by teachers. We compute a weight $\{g_i\}_{i=1}^{n}$ for each teacher to leverage its contribution to the student:

$$\begin{cases} g_i = e^{q_i} \Big/ \sum_{j=1}^{n} e^{q_j} \\ q_i = \mathrm{FC}(\sigma(\mathrm{MaxPool}(\sigma(V_i \cdot V')))) \end{cases}, \quad (1)$$

where FC represents fully connected layer and $\sigma(x) = \mathrm{ReLU}(\mathrm{FC}(x))$. Then, we train the student network via L_p

$$L_p = f(O', \sum_{i=1}^{n} g_i O_i) + \sum_{i}^{n} (g_i \cdot \mathrm{MSE}(V', V_i)), \quad (2)$$

where f is loss function used in [Lin *et al.*, 2019] and MSE is mean squared error loss function, we detail f in supplementary material. The student network follows the same proposal score computation method adopted by teacher networks.

In order to prevent the student network from generating a too high weight for one teacher network, we add suppression weights $\{\bar{g}_i = \frac{1}{n \cdot g_i}\}_{i=1}^{n}$ inspired by weighted binary cross entropy loss function to compute suppression loss $\bar{L}_p$:

$$\bar{L}_p = f(O', \sum_{i=1}^{n} \bar{g}_i O_i) + \sum_{i}^{n} (\bar{g}_i \cdot \mathrm{MSE}(V', V_i)), \quad (3)$$

and the final loss function of the event proposal generation is calculated as:

$$L_p^{final} = \gamma \cdot L_p + \eta \cdot \bar{L}_p, \quad (4)$$

where γ and η are hyperparameters and we set as 0.8, 0.2 after several experiments on the premise of avoiding the excessive role of the suppression loss.

We compute scores of all candidate proposals that exist in a video, and select top K proposals as valid event proposals $E = \{e_i\}_{i=1}^{K}$, the proposal score computation is detailed in supplementary material.

3.2 Proposal-Caption Matching

We adopt COOT [Ging *et al.*, 2020] for weakly supervised proposal-caption matching. In original model, given ground truth event-level annotation, COOT uses contrastive loss L_{con} to enforce the relevant (positive) visual-lexical pair are close while irrelevant (negative) pair are far from

each other. Through COOT, visual modal generates three-level(video, event and contextual) representation, and lexical modal also generates three-level(paragraph, sentence and contextual) representation corresponding to visual modal. Our training data only contains video-level annotation so we can not construct event and contextual level representation in visual modal. Owing to the absent of event-level annotations, we evenly divide a video into multiple clips such that each clip corresponds to a sentence in the paragraph. Thus, given a sentence, the corresponding clip can be the sentence's positive clip and the rest are the negative clips. We create the positive and negative sentences in the same way.

Also, to enhance the matching performance, COOT exploits cycle-consistency loss L_{cycle} to guarantee the matched event to a sentence can be used to locate the same sentence, and vice versa. Similarly, we use evenly divided events to substitute the true event for training. Note that matching aims to generate sentence-proposal(i.e. pseudo event-level annotation) pairwise data that will be used to train the subsequent caption generation module and refine the proposal generation module (detailed in supplementary material). In the testing phase that we showed in Fig 2, we can neglect the matching module and feed the generated event proposals to caption generation module directly because the two generation modules have been already well-trained on pairwise data. Below are the two losses adopted in our training:

Contrastive Loss

COOT adopts contrastive loss on all three-level visual-lexical(video-paragraph, event-sentence, contextual-contextual) representation and losses share the same form, we only explain the event-sentence level loss. Given a positive pair (c^+, s^+) and two negative pairs $(c^-, s^+), (c^+, s^-)$, L_{con} is defined as:

$$\begin{aligned} L_{con} = &\sum L((c^+, s^+), \{(c^-, s^+), (c^+, s^-)\}, h) \\ &+ \sum L((1,1), \{(c^+, c^-), (s^+, s^-)\}, h) \end{aligned}, \quad (5)$$

where h is a margin hyperparameter, the $(1, 1)$ pair denotes that positive samples are not changed, and

$$\begin{aligned} &L((c^+, s^+), \{(c^-, s^+), (c^+, s^-)\}, h) \\ &\quad = \max(0, h + D(c^+, s^+) - D(c^-, s^+)), \\ &\quad + \max(0, h + D(c^+, s^+) - D(c^+, s^-)) \end{aligned} \quad (6)$$

where $D(c^+, s^+) = 1 - (c^+)^T s^+ / (||c^+|| ||s^+||)$ is the cosine distance between two vectors.

Cycle Consistency Loss

Given a sentence s_i, we first compute its clip counterpart $\bar{c}_{s_i}$, and then cycle back to the sentence sequence $\{s_i\}_{i=1}^N$ and calculate the soft location u:

$$\begin{cases} \bar{c}_{s_i} = \sum_{j=1}^N \alpha_j c_j & \text{where } \alpha_j = \frac{e^{-||s_i - c_j||^2}}{\sum_{k=1}^N e^{-||s_i - c_k||^2}} \\ u = \sum_{j=1}^N \beta_j j & \text{where } \beta_j = \frac{e^{-||\bar{c}_{s_i} - s_j||^2}}{\sum_{k=1}^N e^{-||\bar{c}_{s_i} - s_k||^2}} \end{cases}, \quad (7)$$

α_j is the similarity of clip c_j to sentence s_i, and β_j is the similarity of s_j to $\bar{c}_{s_i}$. The object of L_{cycle}^{sent} is to reduce the distance between the source location i and the soft location u:

$$L_{cycle}^{sent} = ||i - u||^2. \quad (8)$$

Similarly, we conduct the cycle consistency evaluation given a clip representation c_i by L_{cycle}^{clip}. Then, the cycle-consistency loss is defined as $L_{cycle} = L_{cycle}^{sent} + L_{cycle}^{clip}$.

The final matching loss L_m is defined as follows:

$$L_m = L_{con} + L_{cycle}. \quad (9)$$

Pretraining based on Annotated Images

We further propose to enhance the matching module using images with ground truth captions. We use annotated images because image is the basic unit of video and the annotation is widely available. We believe annotated images can provide additional static information for building the match between video clips and sentences.

Specifically, given images $\{x_i\}_{i=1}^n$ with captions $\{y_i\}_{i=1}^n$, we simulate pseudo video by duplicating these images and adding Gaussian noise. The pseudo paragraph is constructed by concatenating corresponding captions. We use pseudo data to pretrain COOT to obtain good initialization parameters before training with true videos and paragraphs.

Method	AR@100(%)	AUC(%)
Self-Attn [Zhou *et al.*, 2018]	52.95	-
$\text{TN}_{\text{activity}}$	67.87	68.13
$\text{TN}_{\text{action}}$	63.65	68.18
$\text{TN}_{\text{highlight}}$	53.26	59.99
KDPG	**69.38**	**69.86**

Table 1: Comparison of event proposal generation performance between proposed KDPG and other methods on ActivityNet-Caption validation set.

3.3 Event Caption Generation

The contribution of this paper is on how to use knowledge distillation and cross-modal matching to design a pipeline for weakly supervised DVC. In order to highlight the advantages of our pipeline rather than the advanced sub-model, we use the widely used Attention-LSTM network, then train it on the generated pairwise data to obtain the sentence description for each generated proposal. Attention-LSTM includes two LSTM layers and one attention layer to encode input proposal features and decode to a sentence. At each time step, Attention-LSTM uses previous hidden state and generated word to generate a word probability vector, which we detail in supplementary material. We apply cross-entropy loss L_c as follows to minimize the distance between the one-hot vector of ground-truth caption $s = \{w_i\}_{i=1}^M$ and our prediction $\bar{s} = \{\bar{w}_i\}_{i=1}^M$:

$$L_c = -\sum_{t=1}^{M} w_t \cdot log(\bar{w}_t | w_1 : w_{t-1}). \quad (10)$$

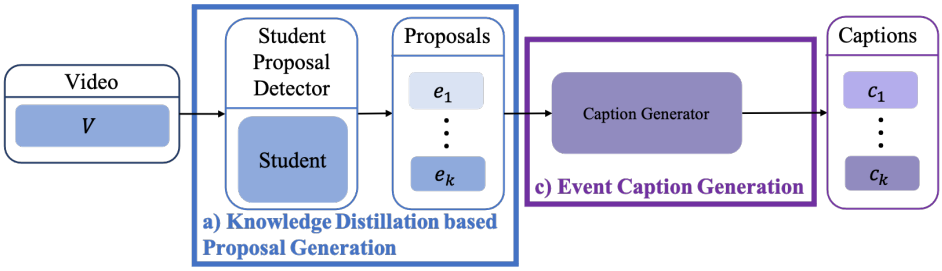

Figure 2: In testing phase we utilize Knowledge Distillation based Proposal Generation and Event Caption Generation to generate dense captions.

Method	M	C	B@1	B@2	B@3	B@4
DCEV(fully supervised)	4.82	17.29	17.95	7.69	3.86	2.20
SDVC(fully supervised)	8.82	30.68	17.92	7.99	2.94	0.93
WSDECV(weakly supervised)	6.30	18.77	12.41	5.50	2.62	1.27
ECG(weakly supervised)	7.06	14.25	11.85	5.64	2.71	1.33

Table 2: Caption generation performance comparison of different full supervised and weakly supervised DVC method with our pipeline.

4 Experiments

To demonstrate the effectiveness of the proposed pipeline, we conduct experiments on the dataset ActivityNet-Caption [Krishna *et al.*, 2017]. For knowledge-distillation, we adopt multiple external datasets used in the task of Temporal Proposal Detection (TPD), i.e, THUMOS-14 [Jiang *et al.*, 2014], ActivityNet-Action [Krishna *et al.*, 2017], BROAD-Video Highlights[1]. For pretraining COOT, we utilize a typical Image-Captioning dataset MS-COCO [Lin *et al.*, 2014], then we use ActivityNet-Caption to continue training COOT. Supplementary material summarizes these datasets.

We compare our method with several representative methods, such as classical DVC model proposed in [Krishna *et al.*, 2017], streamlined DVC (SDVC) method [Mun *et al.*, 2019] and weakly supervised DVC (WSDVC) method [Duan *et al.*, 2018]. We also compare the event proposal generation module with Self-Attn [Zhou *et al.*, 2018] to demonstrate the effectiveness of knowledge distillation strategy.

4.1 Datasets and Processing

ActivityNet-Caption dataset has 20k untrimmed videos devided into training, validation and testing subsets by ratio 2:1:1, and each video on average contains 3.6 event clips with captions. THUMOS-14 dataset has 0.4k untrimmed videos containing multiple action clips with timestamps. We split it into training and validation subsets by ratio 4:1. ActivityNet-Action dataset shares same untrimmed videos as ActivityNet-Caption but has another type of labeled information for temporal action detection task. BROAD-Video Highlights dataset has 1.4k untrimmed entertainment long video with average 12.2 highlights with timestamps. We split it into training and validation subsets by ratio 4:1. MS-COCO dataset aims to solve image-captioning task, and we sampled 50k images with randomly selected one corresponding caption as our supplement fine-grained dataset.

[1]http://ai.baidu.com/broad/

In KDPG module, we extract video features using pretrained C3D [Tran *et al.*, 2015] network on each TPD dataset and reduce the dimensionality of output from 4096 to 500 using PCA. In PCM module, to guarantee the consistency of video and image features, we adopt ResNet to scan each frame in video and each image into a 2048-D vector on ActivityNet-Caption and MS-COCO. In ECG module, we extract video and lexical features on ActivityNet-Caption using video feature extraction method in [Ging *et al.*, 2020] and pretrained BERT [Devlin *et al.*, 2018].

4.2 Comparison Baseline and Setup

For event proposal generation, we follow [Lin *et al.*, 2019] to rescale the length of each feature sequence to 100 using linear interpolation. In the training of teacher networks, we use SGD optimizer and set learning rate to 0.001, batch size to 16. In the training of student network, we set learning rate, batch size, γ and η to 0.0005, 32, 0.8 and 0.2 respectively. In the caption generation of testing phase we use beam search strategy and set the beam size to 5. We mainly compare with Self-Attention (Self-Attn) [Zhou *et al.*, 2018], which uses video encoder with proposal decoder to generate event proposal.

For event caption generation, we use Adam optimizer and set learning rate to 0.02 and decays every 32 epochs with decay rate 0.1, batch size to 16. We compare our method with the following two fully supervised methods and one weakly supervised methods:

1) Dense-Caption Events in Video (DCEV) proposed in [Krishna *et al.*, 2017] used DAPs to generate event proposals and context-based caption generator to generate captions.
2) Streamlined Dense Video Captioning (SDVC) used SST [Buch *et al.*, 2017]+ESGN to generate event proposals and used a hierarchical episode-event RNN to generate caption.
3) Weakly Supervised Dense Event Captioning in Video

(WSDECV) [Duan *et al.*, 2018] used pretrained caption generator and sentence localizer to accomplish end-to-end generation.

4.3 Experimental Results

Table 1 shows the proposal generation performance comparison of our KDPG module with other methods and every single teacher network (TN) we used. To evaluate the quality of a given proposal, we use the proposal to retrieve true event clips and compute the Average Recall (AR) under multiple IoU thresholds [0.5:0.05:0.95]. Then we calculate the mean AR about the top 100 generated proposals (AR@100). In addition, we calculate the Area under the AR vs. AN (Average Number of proposals) curve (AUC). These results are calculated from the validation set in Activity-Caption dataset. The reported results in Table 1 demonstrate the outstanding proposal generation capability of KDPG. Due to the superior global-context extraction ability of BMN+TCN and the appropriate distillation strategy we used, KDPG outperforms the Self-Attn and three teacher networks.

Table 2 shows the caption generation performance comparison of DVC metrics among the DCEV, SDVC, WSDECV and our ECG. To measure the performance of the captioning results, we use commonly-adopted evaluation metrics, i.e. METEOR(M), CIDEr(C) and BLEU@N. We generate event proposals based on different IoU thresholds of [0.3,0.5,0.7,0.9], and compute average metrics over the threshold using official codes[2]. The presented results in Table 2 illustrate that the performance of ECG has comprehensively exceeded the weakly supervised method WSDECV. In some metrics, ECG even exceeds fully supervised methods. We believe the reason is two-fold: First, the distillation strategy produces much better proposals than other methods; Second, our cross-modal matching module builds good correspondences between proposals and sentence captions, which is unavailable in other methods.

Figure 3 shows the qualitative results of our method KDPG-DVC. It is worth mentioning that different events in the same video may have different descriptions, such as "martial arts moves" and "dance", the reason is that each video segment input into the captioner is equivalent to a separate short video. The generated caption is also independent and the context is not strongly related. Although this issue is not our focus, we will consider an appropriate solution in our future work.

[2]https://github.com/ranjaykrishna/densevid_eval

Method	AR@100(%)	AUC(%)
Fully supervised BMN	70.43	71.42
Averaged weights	67.89	68.13
KDPG	69.38	69.86
KDPG + pairwise	**70.77**	**72.05**

Table 3: Ablation results of event proposal generation module when using different strategies.

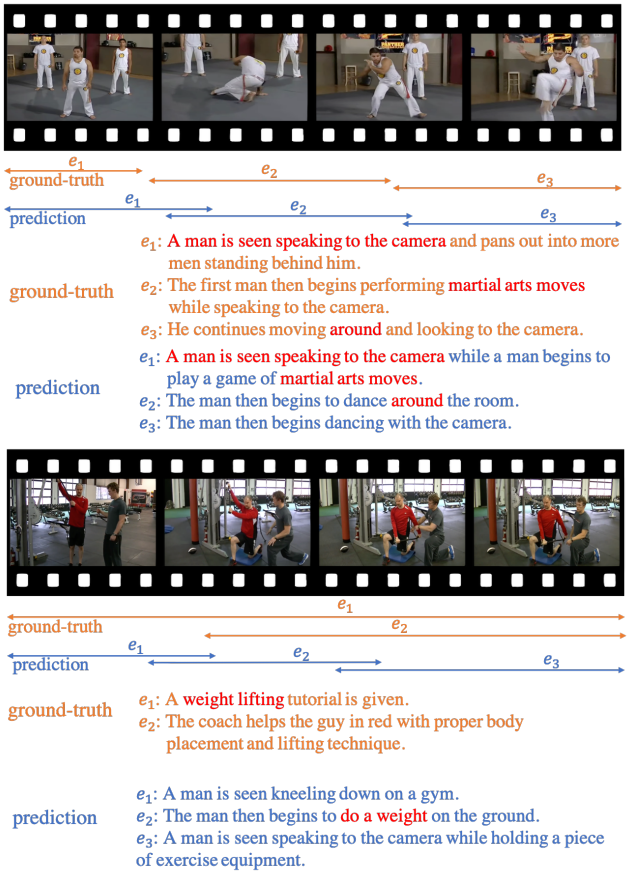

Figure 3: Qualitative results on ActivityNet Captions.

4.4 Ablation Study

In the ablation study, we alter the training pipeline of our model by overlying different training steps to justify the effects of different modules.

Table 3 demonstrates the effectiveness of our novel multi-teachers knowledge distillation learning for event proposal generation. In detail, fully supervised BMN method means BMN trained on fully supervised data. The averaged weights method stands for averaging the weights of all teacher networks. The method of KDPG plus pairwise data denotes that we simultaneously use event annotation of pairwise proposal-sentence data generated by PCM module as hard labels. Experimental results show that self-adaptive weight learning in KDPG is much better than the averaged weights method, and this is because adaptive weight computation transfers appropriate knowledge from all kinds of teachers to the student. In addition, we find the PCM module to be a great help to event proposal generation. With the supervision of pairwise data generated by PCM module, KDPG has been significantly improved and even outperforms the fully supervised method. This demonstrated that our pairwise data supplements more event information that cannot be completely learned from the event annotated data. With our knowledge distillation method, the video clip information of different types is greatly learned, which can help the event extraction.

Method	M	C	B@1	B@2	B@3	B@4
Vanilla	6.51	12.83	11.31	5.31	2.49	1.20
ECG w/o	6.88	13.92	11.71	5.52	2.68	**1.39**
ECG	**7.06**	**14.25**	**11.85**	**5.64**	**2.71**	1.33

Table 4: Comparison of event captioning results with different training corpus.

Method	R@1(%)	MR
PCM w/o	52.6	1.8
PCM	**53.8**	1.8

Table 5: The matching performance comparison of whether using image-captioning data to pretrain the PCM module.

From experience, we believe that iterating multiple times this training strategt can improve the performance of the model, but this is not an innovation of our method, we only iterated once.

Besides, we conduct another experiment to confirm that our PCM module greatly benefits the final event caption generation as shown in Table 4. The method of vanilla indicates that the event caption generator is trained on the existing video-paragraph data corpus, which is the only data we can use without our PCM module. While the ECG w/o means the training data is automatically generated by PCM module. We believe the PCM module provides more precise and general visual-language semantic information via cross-modal matching, which indeed helps a lot. Our final ECG method takes full advantage of both. Specifically, we get event proposals by KDPG+pairwise model and generate pairwise annotated data by PCM module next. Then We train ECG with video-paragraph data firstly and refine it using pairwise annotated data.

To prove that pretraining with image-captioning data promotes the matching capability of PCM module, we report the commonly used retrieval metrics (i.e. R@1 and Median Rank (MR)) calculated on the validation set of ActivityNet-Caption in Table 5. Specifically, the method of PCM w/o is the basic version of our PCM module. It can only depend on limited number of coarse-grained training corpus, including the video-paragraph level and the constructed event-sentences level data. While in the approach of PCM, more accessible and fine-grained image-captioning data is utilized by the proposal-caption matching network for pretraining. Experiments show that the R@1 of PCM outperforms PCM w/o by 1.2%, which demonstrates the effectiveness of the pretraining phase.

5 Future Work

In our future work, we consider conducting comparative analysis research about the strength and weaknesses of pipeline and end-to-end architectures to decide whether we convert our model architecture from pipeline to end-to-end.

Limited by the number of knowledge datasets, our knowledge distillation module has not reached the optimal state, we will continue to investigate related datasets to enhance the performance. And we believe use other caption generation models also can help our pipeline a lot, especially in enhancing the context relation between different events in the same video.

It is also important to note that the usage of labeled images is simple, and we will consider how to better integrate the information of static images as a supplement into the video encoding and decoding phase.

6 Conclusion

In this paper, we present an efficient pipeline which contains three modules i.e., distillation learning based proposal generation, proposal-caption matching and event caption generation addressing the weakly supervised DVC task. Knowledge distillation learning is used to solve the unsupervised proposal generation task and cross-modal matching is used to generate precise proposal-sentence pairs from video-paragraph. Joint usage of the above-mentioned methods solves the proposal generation and event-caption generation challenges of weakly supervised DVC. This pipeline architecture can provide promotions to the full pipeline by improving every single module, and the positive interaction between every module also promotes the full pipeline.

Experimental results on the dataset of ActivityNet-Caption demonstrate the significance of distillation-based event proposal generation and cross-modal retrieval-based semantic matching to weakly supervised DVC.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 61836002.

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[Buch *et al.*, 2017] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Duan *et al.*, 2018] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou

Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3059–3069, 2018.

[Escorcia *et al.*, 2016] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016.

[Gabeur *et al.*, 2020] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, volume 5. Springer, 2020.

[Ging *et al.*, 2020] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *arXiv preprint arXiv:2011.00597*, 2020.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Jiang *et al.*, 2014] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[Krishna *et al.*, 2017] 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*, pages 706–715, 2017.

[Lea *et al.*, 2017] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2019] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3889–3898, 2019.

[Mun *et al.*, 2019] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6588–6597, 2019.

[Peng *et al.*, 2019] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International Conference on Machine Learning*, pages 5092–5101, 2019.

[Shen *et al.*, 2017] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1924, 2017.

[Song and Soleymani, 2019] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[Tran *et al.*, 2015] Du Tran, Lubomir 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*, pages 4489–4497, 2015.

[Xiong *et al.*, 2018] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483, 2018.

[Yu *et al.*, 2020] Ting Yu, Jun Yu, Zhou Yu, Qingming Huang, and Qi Tian. Long-term video question answering via multimodal hierarchical memory attentive networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[Zhao *et al.*, 2017] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[Zhou *et al.*, 2018] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.