Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning

Shaoxiang Chen and Yu-Gang Jiang*
Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University
{sxchen13, ygj}@fudan.edu.cn

Abstract

Dense Event Captioning (DEC) aims to jointly localize and describe multiple events of interest in untrimmed videos, which is an advancement of the conventional video captioning task (generating a single sentence description for a trimmed video). Weakly Supervised Dense Event Captioning (WS-DEC) goes one step further by not relying on human-annotated temporal event boundaries. However, there are few methods trying to tackle this task, and how to connect localization and description remains an open problem. In this paper, we demonstrate that under weak supervision, the event captioning module and localization module should be more closely bridged in order to improve description performance. Different from previous approaches, in our method, the event captioner generates a sentence from a video segment and feeds it to the sentence localizer to reconstruct the segment, and the localizer produces word importance weights as a guidance for the captioner to improve event description. To further bridge the sentence localizer and event captioner, a concept learner is adopted as the basis of the sentence localizer, which can be utilized to construct an induced set of concept features to enhance video features and improve the event captioner. Finally, our proposed method outperforms state-of-the-art WS-DEC methods on the ActivityNet Captions dataset.

1. Introduction

A conventional video captioning task [6, 53] refers to generating a single sentence description for a trimmed video (usually around 10 seconds long), and it has been extensively studied in recent years [45, 58, 34, 17, 46, 61, 67]. Since natural videos contain multiple events, the Dense Event Captioning (DEC) task [20] is later introduced, which aims at generating multiple temporally localized event descriptions for untrimmed videos with the help of temporal boundary annotations for each event. Weakly Supervised Dense Event Captioning (WS-DEC) [11] goes one step further by not relying on the resource-consuming temporal boundary annotations, and this also makes the task more challenging.

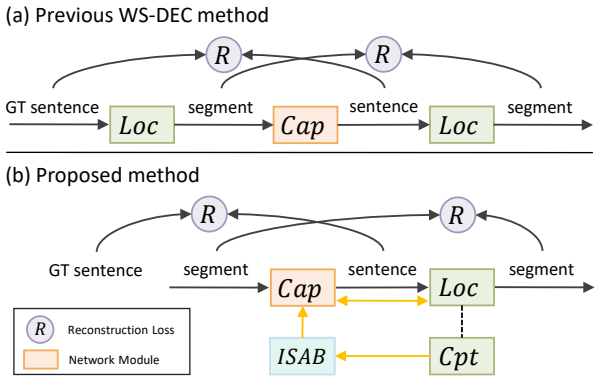

Figure 1. Comparison of the workflows of the previous method [11] (a) and our proposed method (b). 'Cpt' and 'ISAB' denote concept learner and induced set attention block. The yellow connections highlight our advantages, and they bridge the captioner and localizer more closely by allowing richer bidirectional information passing between the two modules.

For strongly supervised dense event captioning, the event localization module can receive strong supervision signals to learn to predict precise temporal boundaries and the captioning module can focus on more accurate video segments during both training and testing. Thus most existing methods [47, 22, 68, 32] design a localization module similar to temporal action detection networks [5, 24] to firstly generate events of interest and then describe them with a captioning module. But for weakly supervised dense event captioning, other than the sentence annotation, the localization and captioning modules can only receive information from each other. As shown in Fig. 1 (a), the first WS-DEC method [11] adopted an iterative approach, in which the event captioner and sentence localizer in turn feed outputs to each other. Its underlying assumption is that by op-

*Corresponding author.

timizing the reconstruction losses, the sentence localizer's output will converge to a point that is optimal for the event captioner. To the best of our knowledge, there are only two published methods [11, 39] of WS-DEC and both adopted the same workflow. In this paper, we propose a new WS-DEC method with its simplified workflow shown in Fig. 1 (b). We argue that directly stacking the event captioner and sentence localizer in a feedforward fashion is not sufficient, since there is no pathway for information communication between these two subtasks other than input-output connections. Thus the goal of this work is introducing additional information communication to more closely bridge these two subtasks. Technically, we base our sentence localizer on a multiple instance concept learner, which can capture localized semantic concepts (although not perfect) for aligning with the sentence features in a frame-to-word level. The localizer can then pass two types of information to the captioner: 1) the word importances learned during video-text alignment to guide the caption generation, 2) the concept features to construct an induced set of concept features to enhance the original video features via an Induced Set Attention Block. In this way, both the language decoder and visual encoder are more closely connected with the sentence localizer and can receive richer information compared to previous methods. Note that we bridge the event captioner and sentence localizer but not deeply couple them, so that it is possible to adopt more sophisticated captioner or localizer models in the future for further improvement without changing the proposed workflow. Our proposed method is abbreviated as EC-SL.

The main contribution of this paper is a new way to integrate the temporal localization and description for events in untrimmed videos under the weakly supervised setting, where temporal boundary annotations are not available. The proposed method is also an exploration of what type of information communication pathway should be built between the two subtasks in order to better bridge and unify them, which is a problem not yet fully investigated in the literature.

2. Related Work

Video Captioning. Most video captioning methods focus on generating a single sentence for a trimmed video, and this can be treated as a sequence-to-sequence translation and modeled by encoder-decoder networks [45]. To enrich video representation, later methods either extend the encoder-decoder network by incorporating temporal [58]/spatial [57, 8, 54] attention mechanism or fusing multimodal feature representations [17, 1, 48, 28, 7]. Recent methods also explore more fine-grained object features [61, 65] to better understand videos.

Dense Event Captioning. Dense event captioning networks can generally be decomposed into two parts: event localization and sentence generation modules, where the latter is usually a RNN-based decoder [20, 47, 22, 32] or a Transformer [68]. Using temporal annotations for the events, these methods design and train their event localization modules like in action detection [12, 5, 24], i.e., event proposals are generated by complex RNNs [20, 47] or temporal convolution networks [22, 68]. However, their event localization modules can not work without temporal annotations (which are resource-consuming), and thus can not be directly adapted for weakly supervised dense event captioning. There are other attempts at generating multiple sentence descriptions for an untrimmed video, such as video paragraph captioning [52, 59]. The major difference is that paragraph captioning focuses on the coherence of the sentences and completely ignores the localization of the events.

Attention-Based Feature Aggregation. Multimodal attention [17, 28, 54], memory models [48, 36, 13, 27], and feature bank [2, 51] can generally be viewed as attention based feature aggregation approaches, which are popular in various video understanding tasks. We draw inspiration from [21], which develops an Induced Set Attention Block to process sets of data. But unlike [21], where the inducing points are trainable parameters, we learn and construct the induced set out of a concept learner.

MIL for Weakly Supervised Localization. Multiple Instance Learning (MIL) is commonly used in weakly supervised localization tasks, such as temporal action localization [26, 33, 49] and object detection [55, 56, 42]. When temporal/spatial annotations are not available, the whole video/image is treated as a bag of instances with bag-level annotations, and the predictions for instances (frames/object proposals) are aggregated as the bag-level prediction. The instances can capture localized features through MIL. In our work, we learn concept features for sentence localization and further construct an induced set of concept features to improve the event captioner.

Temporal Event Localization. Temporally localizing event descriptions in videos is a recently-proposed task [14]. For this task, performing cross-modal interaction between the visual and textual modalities is critical [60, 63, 19, 31, 38, 16], and is usually done by first separately encoding the video and sentence and then jointly processing visual and textual representations. The temporal segments are predicted by regression [60, 31] or ranking proposals [63, 19, 38, 16] based on the fused visual-textual representation. Tackling temporal event localization without temporal annotation is also an open problem, and several methods based on contrastive learning have been recently proposed [30, 29, 64]. We should note that directly stacking captioning and weakly-supervised temporal event localization methods can yield a naïve solution to WS-DEC, but the overall framework will not be end-to-end trainable and there will be no rich information communication be-

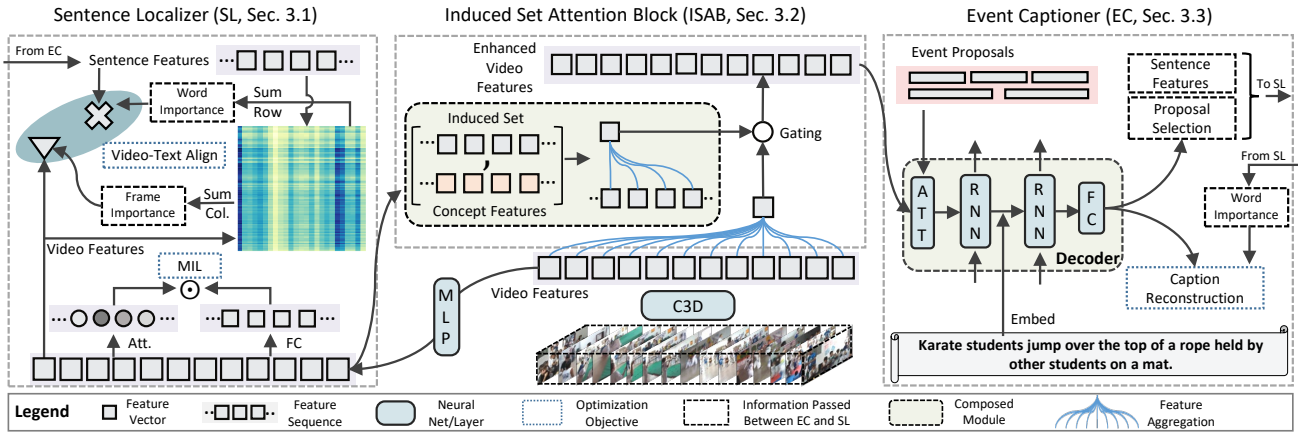

Figure 2. The training workflow of the proposed method. The sentence localizer is built on a MIL (Multiple Instance Learning) concept learner and performs video-sentence interaction and alignment (Sec. 3.1). Based on the concept learner, an Induced Set Attention Block is constructed and used to enhance the video features(Sec. 3.2). The event captioner takes the enhanced features and first generates captions for multiple proposals and selects the most confident one for the sentence localizer to reconstruct (Sec. 3.3). The Induced Set Attention Block acts as a bridge between the sentence localizer and the event captioner.

tween the two subtasks.

Weakly Supervised Dense Event Captioning. To the best of our knowledge, thus far there are only two published WS-DEC methods [11, 39]. The first work [11] proposed an iterative approach. In the training phase, the video and one event description are fed to the sentence localizer to obtain a temporal boundary prediction, and then the temporal segment is fed to the event captioner to generate a sentence, which is again used to relocate the temporal segment. The two modules are learned by alternatively optimizing the caption and temporal segment reconstruction errors across iterations. Following the framework of [11], [39] focuses on utilizing additional audio features to improve event captioning. In existing approaches to WS-DEC, there is no information communication between the event captioner and sentence localizer other than necessary inputs. Besides, the alternative training prevents the whole model from being end-to-end optimized.

3. Proposed Method

As shown in Fig. 2, the model in our proposed method consists of three components, the Sentence Localizer, the Event Captioner, and the Induced Set Attention Block. In the following sections, we describe the details of each component and how the components **benefit from each other**.

3.1. Sentence Localizer

Given a video, the goal of the sentence localizer is to temporally relocate a sentence that the event captioner generated from a video segment (proposal). To achieve this goal, the localizer should analyze the relation between video frames and words, and we tackle this by exploiting semantic concepts in videos.

Learning Semantic Concepts. Inspired by previous work which mine word- or phrase-level semantic concepts [50, 31, 16, 9] from the video to perform sentence localization, in our sentence localizer, we first construct a multiple instance learning subtask to predict concepts with only video-level annotations. The sentence localizer first takes in the video features $\boldsymbol{V} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_L\}, \boldsymbol{V} \in \mathbb{R}^{L \times D_v}$ and projects them into a latent concept space with a position-wise fully connected network, obtaining $\boldsymbol{E} = \{\boldsymbol{e}_l = \Omega(\boldsymbol{v}_l)\}_{l=1}^{L}$, $\boldsymbol{E} \in \mathbb{R}^{L \times D_e}$. Then for each $\boldsymbol{e}_l$ a concept prediction is produced

$$\boldsymbol{p}_l = \texttt{sigmoid}(\Phi(\boldsymbol{e}_l)), \tag{1}$$

where $\boldsymbol{p}_l \in \mathbb{R}^{N_c}$ denotes the probabilities of N_c concepts appearing in the l-th frame (instance), $\Phi(\cdot)$ is a normalized [41] multi-label classifier to alleviate the long-tailed issue of the concept distribution. Since under weak supervision, we only have a single binary label $\hat{\boldsymbol{p}} \in \mathbb{R}^{N_c}$ associated with the entire video (bag), the instance level predictions are first aggregated via attention:

$$\begin{aligned} \bar{\boldsymbol{p}} &= \textstyle\sum_{l=1}^{L} \alpha_l \boldsymbol{p}_l, \quad \text{where} \\ \alpha_l &= \texttt{Softmax}(a_l), \quad a_l = U_{att}(\tanh(U_a \boldsymbol{e}_l + b_a)), \end{aligned} \tag{2}$$

and $U_{att} \in \mathbb{R}^{D_{att} \times 1}$, $U_a \in \mathbb{R}^{D_{att} \times D_e}$, and $b_a \in \mathbb{R}^{D_{att}}$ are learnable parameters. Note that other aggregation functions also work as long as they are permutation-invariant [18]. Then a weighted cross-entropy loss is used to train this MIL network:

$$L_{mil} = -\frac{1}{N_c} \sum_{n=1}^{N_c} \mu_c (\hat{\boldsymbol{p}} \log \bar{\boldsymbol{p}} + (1 - \hat{\boldsymbol{p}}) \log(1 - \bar{\boldsymbol{p}})), \tag{3}$$

where the concept-wise weight μ_c is decided by the occurring frequency of each concept.

Video-Sentence Interaction. Through MIL, the transformed video features $\boldsymbol{E}$ should pickup concept-related information in each frame, and we then perform frame-to-word matching between $\boldsymbol{E}$ and the sentence features $\boldsymbol{S} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_T\}$ produced by the event captioner (Sec. 3.3). The concept and sentence features are projected into a joint space and their matching scores (logits) are computed as

$$\boldsymbol{A}_{lt} = \tau(U_e \boldsymbol{e}_l)\tau(U_s \boldsymbol{s}_t)/\sqrt{D_j}, \tag{4}$$

where D_j is the dimensionality of the joint space, $U_e \in \mathbb{R}^{D_j \times D_e}$ and $U_s \in \mathbb{R}^{D_j \times D_s}$ are transformation matrices, and $\tau(\cdot)$ denotes the ReLU activation. We then sum $\boldsymbol{A} \in \mathbb{R}^{L \times T}$ column-wise and row-wise and normalize them to obtain frame-level and word-level importance weights, respectively.

$$\begin{aligned} \boldsymbol{W}^v &= \texttt{Softmax}(\boldsymbol{A}_{lt}.\texttt{sum}(t)), \\ \boldsymbol{W}^s &= \texttt{Softmax}(\boldsymbol{A}_{lt}.\texttt{sum}(l)), \end{aligned} \tag{5}$$

where $\texttt{sum}(\cdot)$ denotes tensor summation along a specified axis. The normalized $\boldsymbol{W}^v \in \mathbb{R}^L$ and $\boldsymbol{W}^s \in \mathbb{R}^T$ can be viewed as the importance of frames and words, respectively. Based on these weights, the global video and sentence representations can be obtained by weighted aggregation:

$$\bar{\boldsymbol{V}} = \textstyle\sum_{l=1}^{L} \boldsymbol{W}^v \boldsymbol{e}_l, \quad \bar{\boldsymbol{S}} = \textstyle\sum_{t=1}^{T} \boldsymbol{W}^s \boldsymbol{s}_t. \tag{6}$$

Since $\bar{\boldsymbol{V}}$ and $\bar{\boldsymbol{S}}$ are in the joint space, we use a contrastive loss to learn the representations by constructing positive and negative video-sentence pairs.

$$\begin{aligned} L_{cst} = [\triangle - \texttt{sim}(\bar{\boldsymbol{V}}^+, \bar{\boldsymbol{S}}^+ + \texttt{sim}(\bar{\boldsymbol{V}}^+, \bar{\boldsymbol{S}}^-)]_+ \\ + [\triangle - \texttt{sim}(\bar{\boldsymbol{V}}^+, \bar{\boldsymbol{S}}^+) + \texttt{sim}(\bar{\boldsymbol{V}}^-, \bar{\boldsymbol{S}}^+)]_+, \end{aligned} \tag{7}$$

where $[\cdot]_+$ denotes truncating negative values to zero, $(\bar{\boldsymbol{V}}^+, \bar{\boldsymbol{S}}^+)$ is the positive pair as annotated, and the negative pairs $(\bar{\boldsymbol{V}}^+, \bar{\boldsymbol{S}}^-)$ and $(\bar{\boldsymbol{V}}^-, \bar{\boldsymbol{S}}^+)$ can be constructed by replacing the sentence or video with an arbitrary sentence or video in the mini-batch, this is because the ActivityNet Captions dataset is diverse enough. $\texttt{sim}(\cdot)$ is the cosine similarity function and $\triangle$ is the margin and is empirically set to 0.3 in all our experiments. Because the sentence fed to the localizer is generated from an event proposal, $\hat{G} = \{\hat{s}, \hat{e}\}$, we use a cross-entropy loss to ensure the frame-level importance scores $\boldsymbol{W}^v$ can relocate the proposal.

$$\begin{aligned} L_{loc} = -\frac{1}{L} \textstyle\sum_{l=1}^{L} \big(M(l, \hat{G}) \log \boldsymbol{W}^v + \\ (1 - M(l, \hat{G})) \log(1 - \boldsymbol{W}^v)\big), \end{aligned} \tag{8}$$

where

$$M(l, \hat{G}) = \texttt{sigmoid}(l - \hat{s}) - \texttt{sigmoid}(l - \hat{e}) \tag{9}$$

is a soft mask construction function. Segment predictions can then be generated by applying the Temporal Actionness Grouping [66] strategy to the the predicted $\boldsymbol{W}^v$. During inference, there are multiple proposals and the event captioner will help rank them (Sec. 3.3).

3.2. Induced Set Attention Block

Preliminary. We begin this section by defining some common attention operations that will be used throughout this section. Given two sets of features $\boldsymbol{X} \in \mathbb{R}^{N \times D_x}, \boldsymbol{Y} \in \mathbb{R}^{N \times D_y}$, the Multihead Attention Block (MAB): $\mathbb{R}^{N \times D_x} \times \mathbb{R}^{N \times D_y} \rightarrow \mathbb{R}^{N \times D_x}$, is the same as defined in previous methods [43, 21].

$$\begin{aligned} &\text{MAB}(\boldsymbol{X}, \boldsymbol{Y}) = \texttt{LN}(\boldsymbol{H}, \texttt{FFN}(\boldsymbol{X})), \\ &\text{where} \quad \boldsymbol{H} = \texttt{LN}(\boldsymbol{X} + \texttt{MultiHead}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Y})), \end{aligned} \tag{10}$$

$\texttt{LN}(\cdot)$ is Layer Normalization [3], $\texttt{FFN}(\cdot)$ is position-wise feedforward layer, and $\texttt{MultiHead}(\cdot)$ is the same as the original [43]. Then using MAB we can conveniently define Self-Attention Block (SAB) and Vector Attention Block (VAB):

$$\begin{aligned} \text{SAB}(\boldsymbol{X}) &:= \text{MAB}(\boldsymbol{X}, \boldsymbol{X}), \\ \text{VAB}(\boldsymbol{x}, \boldsymbol{Y}) &:= \text{MAB}(\boldsymbol{x}.\texttt{unsqueeze}(0), \boldsymbol{Y}), \end{aligned} \tag{11}$$

where the $\texttt{unsqueeze}(\cdot)$ operator expands a vector to a matrix with one row. In [21], the Induced Set Attention Block is defined as

$$\begin{aligned} &\text{ISAB}(\boldsymbol{X}) = \text{MAB}(\boldsymbol{X}, \boldsymbol{H}), \\ &\text{where} \quad \boldsymbol{H} = \text{MAB}(\boldsymbol{I}, \boldsymbol{X}). \end{aligned} \tag{12}$$

The inducing points $\boldsymbol{I}$ are randomly initialized trainable parameters, and are expected to implicitly encode some global structure of the inputs $\boldsymbol{X}$.

Induced Set Attention. We construct the inducing points out of the sentence localizer's learned concept representations $\boldsymbol{E} = \Omega(\boldsymbol{V})$. Through MIL, the position-wise concept predictions $\boldsymbol{p}_l$ in Eq. (1) can capture the semantic concepts (although not perfect) in each short video segment represented by $\boldsymbol{v}_l$, and have associated concept features $\boldsymbol{e}_l$. We select K top concepts $\{c_1, ..., c_K\}$ from the video-level prediction $\bar{\boldsymbol{p}}$. For each concept c_k, since $\bar{\boldsymbol{p}}[c_k] = \sum_l \alpha_l \boldsymbol{p}_l[c_k]$ (Eq. (2)), we obtain the most contributing feature $\boldsymbol{e}_{l_k}$, where

$$l_k = \underset{l}{\text{argmax}}\, \alpha_l \boldsymbol{p}_l[c_k], \tag{13}$$

and $[\cdot]$ is the vector indexing operator. Then the induced set of concept features are $\boldsymbol{F} = \{\boldsymbol{e}_{l_1}, ..., \boldsymbol{e}_{l_K}\}, \boldsymbol{F} \in \mathbb{R}^{K \times D_e}$. We define our Induced Set Attention Block as

$$\begin{aligned} &\text{ISAB}_{cf}(\boldsymbol{V}) = \text{MAB}(\boldsymbol{V}, \boldsymbol{H}), \\ &\text{where} \quad \boldsymbol{H} = \text{MAB}(\boldsymbol{F}, \Omega(\boldsymbol{V})). \end{aligned} \tag{14}$$

Here each $\boldsymbol{H}_k = \text{VAB}(\boldsymbol{F}_k, \boldsymbol{E})$ also has a clear meaning, which is attentively selecting relevant concept features for each $\boldsymbol{e}_{l_k}$. Since the predictions $\boldsymbol{p}_l$ can be noisy, this can help construct robust concept features to enhance the original video features.

Gating Enhanced Features. We denote the enhanced

features and original features as $\boldsymbol{V}^E = \text{ISAB}_{cf}(\boldsymbol{V})$ and $\boldsymbol{V}^O = \text{SAB}(\boldsymbol{V})$, respectively. To combine them, we devise a gating mechanism

$$\begin{aligned} \boldsymbol{V}_l^C &= \boldsymbol{g}_l \boldsymbol{V}_l^E + (1-\boldsymbol{g}_l)\boldsymbol{V}_l^O, \\ \text{where} \quad \boldsymbol{g}_l &= \texttt{sigmoid}(U_{g_E}\boldsymbol{V}_l^E + U_{g_O}\boldsymbol{V}_l^O + b_g), \end{aligned} \tag{15}$$

and $U_{g_E} \in \mathbb{R}^{1\times D_e}$, $U_{g_O} \in \mathbb{R}^{1\times D_v}$, and $b_g \in \mathbb{R}$ are learnable parameters. The combined features $\boldsymbol{V}^C$ are the inputs to the event captioner. The gating is to allow the captioner to dynamically choose to rely on the enhanced feature or the original feature at each frame, and this can stabilize training at the early stage.

What is the functionality of ISAB? Since video is a collection of semantic concepts and their interactions, the information captured by the ISAB can be seen as a highly abstracted global structure of the video. This information is better captured by the concept learner inside the sentence localizer since it is location-sensitive, whereas the event captioner (even with attention) tends to ignore such structure when it can easily draw rich information in an untrimmed video for generating a sentence. Instead, we expect the event captioner to benefit from such information.

3.3. Event Captioner

Before feeding the enhanced video features to the event captioner, we note that the induced set attention is permutation-invariant. However, the video's temporal structure should be modeled in order to understand an event. So we use a single-layer bidirectional LSTM to sequentially process the enhanced video features, and the hidden states of both directions are concatenated at each time step as the encoded feature. The encoded features are denoted as $\widetilde{\boldsymbol{V}}$.

Since the sentence's corresponding temporal boundaries are unknown, we first generate a set of event proposals to describe and choose the most confident one to feed to the sentence localizer to reconstruct the proposal. The event proposals are denoted as $\boldsymbol{G} = \{G_i\}_{i=1}^M$, where $G_i = (s_i, e_i)$ is a segment of the video, and a mask $M(:, G_i) \in [0,1]^L$ can be generated as in Eq. (9). The mask $M(:, G_i)$ and the encoded video features $\widetilde{\boldsymbol{V}}$ are fed to the event captioner, which is a two-layer LSTM network to generate a sentence. During training, the annotated caption is given as $\boldsymbol{C} = \{\hat{w}_1, ..., \hat{w}_T\}$.

Then the LSTM network runs for T steps to generate a sentence and each step is formulated as

$$\begin{aligned} \boldsymbol{h}_t^{(1)} &= \texttt{LSTM}^{(1)}(\texttt{att}(\widetilde{\boldsymbol{V}}, M(:,G_i), \boldsymbol{h}_{t-1}^{(1)}), \boldsymbol{h}_{t-1}^{(1)}), \\ \boldsymbol{h}_t^{(2)} &= \texttt{LSTM}^{(2)}([\texttt{embed}(w_{t-1}), \boldsymbol{h}_t^{(1)}], \boldsymbol{h}_{t-1}^{(2)}), \end{aligned} \tag{16}$$

where $[\cdot]$ means feature concatenation, $\boldsymbol{h}_t^{(k)}$ is the state of the k-th LSTM at the t-th timestep, $\texttt{embed}(\cdot)$ is a word-embedding function, and $\texttt{att}(\cdot)$[1] is the temporal attention [58] mechanism on the video features with the attention weights masked by $M(:, G_i)$. Each $\boldsymbol{h}_t^{(2)}$ is fed to a fully-connected layer with Softmax activation to produce a probability prediction over the vocabulary, denoted by $\boldsymbol{y}_t^i \in [0,1]^{N_w}$, where N_w is the vocabulary size. Given the annotated caption $\boldsymbol{C}$, the confidence of the prediction for proposal G_i is

$$\text{Conf}_i = \frac{1}{T}\log(\prod_{t=1}^{T} \boldsymbol{y}_t^i[\hat{w}_t]) = \frac{1}{T}\sum_{t=1}^{T}\log(\boldsymbol{y}_t^i[\hat{w}_t]). \tag{17}$$

We select the proposal with the highest confidence $\hat{G} = \text{argmax}_{G_i}\,\text{Conf}_i$ and the features of its generated sentence as inputs to the sentence localizer for training. While during inference, we compute the confidences of each proposal based on their predicted words to rerank them. The hidden states $\boldsymbol{h}_t^{(2)}$ for the words are used as the sentence features (*i.e.* $\boldsymbol{s}_t$ in Eq. (4)) for the localizer. The word-level importance weights $\boldsymbol{W}^s$ from the sentence localizer are incorporated (controlled by hyperparameter γ) into the captioning loss to recalibrate the word importances

$$L_{cap} = -\frac{1}{T}\textstyle\sum_{t=1}^{T}(1+\gamma \boldsymbol{W}^s)\log(\boldsymbol{y}_t^i[\hat{w}_t]). \tag{18}$$

3.4. Overall Loss

The sentence localizer and the event captioner can be jointly trained with the overall loss

$$L_{all} = L_{cap} + \lambda_l L_{loc} + \lambda_c L_{cst} + \lambda_m L_{mil}, \tag{19}$$

where λ_l, λ_c, and λ_m are tunable hyperparamaters used to balance these losses during training.

3.5. Discussion

Although we have noted in both Fig. 2 and the subsections about what information is passed between the captioner and localizer, we would like to further discuss and emphasize about this.

From localizer to captioner. Since the core of the localizer is modeling the relation between frames and words, its learned features contain richer semantic concept information than the raw CNN features, and this is helpful for the captioner. We further devise the ISAB to condense and stabilize the concept features. In addition, the word importance learned in the localizer is utilized in the training of the captioner.

From captioner to localizer. The captioner is mainly responsible for the sentence-related aspects. It produces feature vectors of each word for the localizer, and most importantly, the confidence scores of generated sentences that can be used to rerank the corresponding video segments.

This bidirectional information communication is the main difference between our method and existing ones.

[1] More details are placed in the Supplementary Materials.

Method	WS	M	C	R	B@1	B@2	B@3	B@4
DCE [20]	✗	4.82	17.29	-	17.95	7.69	3.86	2.20
DVC [22]	✗	6.93	12.61	-	12.22	5.72	2.27	0.73
Bi-SST [47]	✗	9.60	12.68	19.10	18.99	8.84	4.41	2.30
Masked-Transformer [68]	✗	9.56	-	-	-	-	4.76	2.23
SDVC [32]	✗	8.82	30.68	-	17.92	7.99	2.94	0.93
RUC AI·M^3 [40]	✗	11.28	14.03	-	16.59	9.65	5.32	2.91
Multi-modal MUTAN fusion [39]	✓	4.93	13.79	10.39	10.00	4.20	1.85	0.90
WS-DEC [11]	✓	6.30	18.77	12.55	12.41	5.50	2.62	1.27
EC-SL (Ours)	✓	**7.49**	**21.21**	**13.02**	**13.36**	**5.96**	**2.78**	**1.33**
EC-SL (less data)	✓	7.03	19.53	12.79	12.13	5.68	2.59	1.23

Table 1. Dense event captioning performances (%) of our proposed method and state-of-the-art methods on the ActivityNet Captions validation set. "WS" denotes "weakly supervised".

Method	WS	IoU=0.1	IoU=0.3	IoU=0.5
CTRL [14]	✗	-	47.43	29.01
ABLR [60]	✗	73.30	55.67	36.79
2D-TAN [62]	✗	-	58.75	44.05
FIAN [38]	✗	-	64.10	47.90
WSLLN [15]	✓	75.40	42.80	22.70
LCGB [10]	✓	74.20	44.30	23.60
SCN [25]	✓	71.48	47.23	29.22
WS-DEC [11]	✓	62.71	41.98	23.34
EC-SL (Ours)	✓	68.48	44.29	24.16

Table 2. Temporal event localization performances (%) of our proposed method and state-of-the-art methods on the ActivityNet Captions dataset. We measure the top-1 recall under different temporal IoU thresholds between the predicted temporal segment and the ground-truth. "WS" denotes "weakly supervised".

4. Experiments

4.1. Experimental Setting

Dataset. The ActivityNet Captions dataset originally contains 10,009, 4,917, and 5,044 videos for training, validation, and testing, respectively. The video duration is about 150 seconds on average. Each video contains 3.65 sentence annotations on average, and each sentence has an average length of 13.48 words. Since the testing set is not publicly available, we test our method on the original validation set (both val_1 and val_2) as previous methods [11, 39]. We also randomly split a small validation set with 577 videos from the training set for hyperparameter tuning and model selection.

Training and Inference. In the first training epoch, we generate non-overlapping sliding window proposals with $\{1/4, 1/3, 1/2\}$ of the video length. As described in Sec. 3, the captioner chooses one proposal and the localizer predicts frame importance weights. We use the Temporal Actionness Grouping (TAG) [66] to generate proposals based on the predictions and use them for the next epoch. For inference, the initial proposals are sliding windows with $\{1/3, 1/2, 1\}$ of the video length. The proposals generated by TAG are ranked by their confidences (Eq. (17)), and we keep the top 4 proposals after reranking.

Evaluation. Like previous methods, we use the official evaluation script[2] released in [20], which uses METEOR [4], CIDEr [44], ROUGE [23], and BLEU [35] metrics to evaluate event caption quality. Specifically, the above metrics are computed for the proposals if they overlap with the ground-truth segments with a temporal IoU larger than a given threshold, and otherwise the scores are set to 0. All the metrics are computed with IoU thresholds $\{0.3, 0.5, 0.7, 0.9\}$ and finally averaged.

Implementation Details. For the video features, we use the officially released C3D features as [11] and uniformly sample 128 features for each video for efficient computing. The sentences are trimmed to have a maximum sentence length of 25 words, and the vocabulary constructed on the training set contains 10,300 words. For fair comparison, we do not use pretrained word embeddings and all the word embedding vectors are randomly initialized and jointly trained with the whole model. To build the concept vocabulary, we select all the noun and verb words from the sentences of training videos and lemmatize them, and the words that occur more than 5 times are kept, which results in a concept vocabulary of 2,186 words. Each sentence is then assigned a binary label according to whether its words are in the vocabulary. The λ_l, λ_c, and λ_m are set to 1, 1, and 10, respectively. The batch size is set to 32 in all our experiments, and we use the Adam optimizer with a learning rate of 1e-4. We implement all our code in PyTorch 1.6.0.

4.2. Performance Comparison

Table 1 shows the dense event captioning performances of our method and state-of-the-art methods, note that there are only two published weakly supervised methods [11, 39]. [39] used 80% of the training data due to YouTube video downloading issues (8,026 training video), so we also performed an experiment on randomly sampled 8,026 videos (less data) for fair comparison. As can be observed, our method outperforms weakly supervised competitors by a clear margin on most metrics (1.19 on METEOR and 2.44 on CIDEr), which proves that our proposed method of bridging the event captioner and sentence localizer can re-

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

#	Method	M	C	R	B@1	B@2	B@3	B@4
1	EC-SL base	6.85	19.09	11.92	12.14	5.22	2.41	1.15
2	EC-SL base + L_{mil}	7.03	20.46	12.24	12.66	5.59	2.55	1.18
3	EC-SL base + L_{mil} + ISAB	7.18	20.98	12.79	13.09	5.63	2.56	1.23
4	EC-SL base + L_{mil} + ISAB + gating	7.39	21.44	12.92	13.36	5.96	2.69	1.28
5	EC-SL base + L_{mil} + ISAB + gating + $\boldsymbol{W}^s$ in L_{cap}	7.49	21.21	13.02	13.36	6.05	2.78	1.33

Table 3. Ablation Experiments. The 'base' model is our final model with L_{mil}, ISAB, gating, and $\boldsymbol{W}^s$ in L_{cap} removed.

Group	Component	Value	M	C	R	B@1	B@2	B@3	B@4
1	Induced Set Size	8	7.26	21.05	12.74	13.11	5.89	2.72	1.25
		<u>16</u>	7.49	21.21	13.02	13.36	5.96	2.78	1.33
		32	7.45	21.44	12.92	13.37	6.02	2.74	1.27
		64	7.48	21.45	12.90	13.36	5.99	2.73	1.26
2	Choosing Inducing Points	Random	7.14	20.67	12.75	13.10	5.65	2.56	1.20
		<u>Max</u>	7.49	21.21	13.02	13.36	5.96	2.78	1.33
3	Word Representations	Logits	6.57	18.41	11.76	12.06	5.19	2.35	0.98
		Embed	7.37	21.36	12.81	13.28	5.95	2.62	1.14
		<u>Hidden</u>	7.49	21.21	13.02	13.36	5.96	2.78	1.33

Table 4. More Ablation Experiments. We vary the size of the induced set of concept features, and study different ways to construct the induced set and different types of sentence features. Underlined are the default values for each component in our final model.

ally boost event captioning performance. It should also be noted that comparing our method with strongly supervised methods, there is still a significant performance gap. The advantage of supervised method is that (1) their captioning models can get accurate video segments during training, which critically affects description quality, and (2) their temporal localization models are strongly supervised to accurately find video segments which may contain events. Also, using reinforcement learning to optimize for the evaluation metrics are commonly seen [22, 32, 40].

To compare with [11], we evaluate the single sentence localization performance of our method and the event captioner is completely not involved, which is also the setting of the temporal event localization task [14]. Table 2 presents the temporal event localization performances of our method and compared methods. Note that this evaluation is not intended for WS-DEC methods because: 1) WS-DEC methods are not trained to localize the ground-truth sentences, and 2) pretrained word embeddings (such as GloVe [37]) are not used. Nevertheless, our method achieves results comparable to state-of-the-art weakly supervised event localization methods and outperforms the previous WS-DEC method [11]. This can prove the effectiveness of our sentence localizer. From Table 2 we can also see that strongly supervised methods have a clear advantage over weakly supervised ones regarding top-1 recalls. This is mainly because they generate a large number of segment proposals by dense sliding windows [38] or even enumerating all possible temporal segments [62].

4.3. Ablation Study

In this section, we present ablation experiment results to verify the effectiveness of the important components and design choices of our proposed model. The results are presented in Table 3 and 4. Due to the space limit, more ablation results are placed in the Supplementary Materials.

Effects of MIL. From Table 3 (#1 and #2) we can first confirm that multiple instance concept learning is important for performing a frame-to-word matching between the video and sentence in the localizer, which may also affect event captioning through $\boldsymbol{W}^s$. The multi-label concept prediction has a top-20 precision and recall of 22.52% and 37.93%, respectively. Note that it is possible to use an external concept detector (like in [50]), but that would require additional annotations and is somehow against the intention of weak supervision. Also, there is a gap between external datasets and the ActivityNet regarding the concepts.

Effects of the Induced Set Attention Block. Comparing experiments #2 and #3, the enhanced features from the ISAB can clearly improve captioning performances without the gating mechanism ($\boldsymbol{V}^E$ and $\boldsymbol{V}^O$ are summed). This proves that the ISAB really passes useful information to the captioner. As indicated by experiments #3 and #4, the gating in Eq. (15) helps the captioner utilize concept features more properly and achieve a further performance boost.

Effects of caption loss weighting. The caption loss weighting in Eq. (18) is an important connection between the sentence localizer and the event captioner in our method. The weights provided by the sentence localizer represents the importance of the words for localizing the corresponding event in video, thus as shown by comparing experiments #4 and #5, guiding the captioner to focus on more important words can improve captioning performances.

In Table 4, we study different configurations of the model components.

Different constructions of the induced set. We can observe that as the size of the induced set increases (Group 1), the performance first improves and then saturates. This indicates that the inducing points really extracts crucial information from the concept features and a small number of

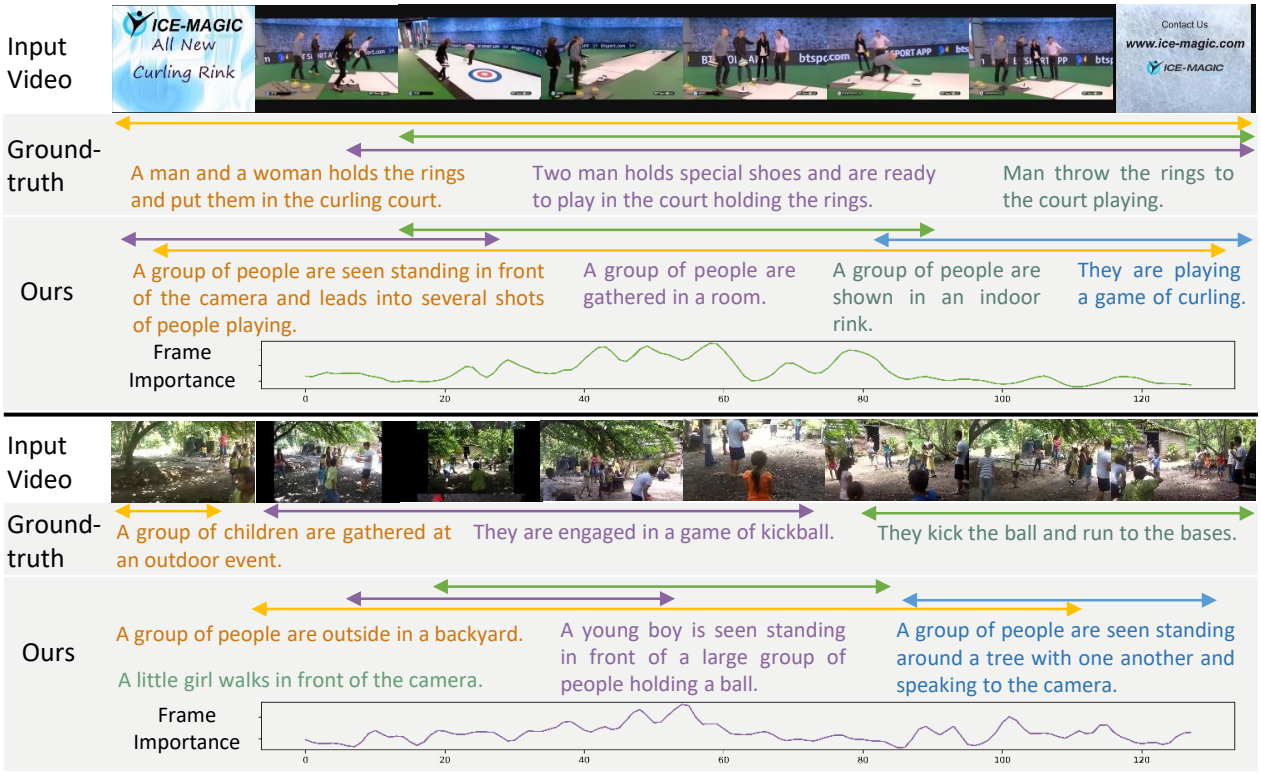

Figure 3. Qualitative Results (Best viewed in color). The ground-truth temporal segments and captions as well as ours are shown. The segments and captions are matched using different colors. We also show the frame importance weights for one of our predicted segments.

them is sufficient. As pointed out by [21], this also helps reduce the time complexity of the attention block compared to using all the concept features. We also validate the way of choosing inducing points in our method by comparing against randomly selecting the concept features (Group 2).

Different sentence representations. Since we do not use pretrained word embeddings, the word representations can only be provided by the event captioner. We study three types of representations (Group 3): the randomly-initialized word embedding, the LSTM hidden state after encoding the word embedding, and the logits of word prediction. Among these choices, the logits performs the worst, this is because it is high-dimensional and leads to severe over-fitting for the localizer. The LSTM hidden states perform better than plain word embeddings, we conjecture the reason is that the hidden state encodes not only the current word, but also the previous words, giving it richer context about the sentence.

4.4. Qualitative Results

Fig. 3 presents the qualitative results of two videos. We can see that our model is able to generate localized and accurate event descriptions despite the lack of temporal annotations. Compared to the ground-truth, some captions are comparable in terms of capturing the overall environment (e.g., indoor rink, outside, and backyard), however, it is difficult to recognize small objects like the rings, shoes and ball. This could be the limitation of our concept learner, which is trained on long-tailed data and is based on the C3D network that crops the video to 112×112 pixels.

5. Conclusion

In this paper, we have studied the task of weakly supervised dense video captioning, which is both challenging and rarely investigated in the literature. To more closely bridge the event captioner and sentence localizer, which is overlooked by existing methods, we have proposed an Induced Set Attention Block built on top of the multiple instance concept learner to act as the pathway for information communication from the sentence localizer to the event captioner. The experimental results on the ActivityNet Captions dataset have shown that our proposed method outperforms state-of-the-art methods when adopting the same video features. The limitation of our method is that without external training data, the concept learner can not accurately detect concepts that are visually small and still suffers from the long-tailed issue. Addressing these limitations and exploring ways to pass information from the event captioner to the sentence localizer are the focuses of our future work.

Acknowledgement. This work was supported in part by National Natural Science Foundation of China (#62032006).

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, pages 12487–12496, 2019. 2

[2] Tim Althoff, Hyun Oh Song, and Trevor Darrell. Detection bank: an object detection based video representation for multimedia event recognition. In *ACM MM*, pages 1065–1068, 2012. 2

[3] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv: 1607.06450*, 2016. 4

[4] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72, 2005. 6

[5] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. SST: single-stream temporal action proposals. In *CVPR*, pages 6373–6382, 2017. 1, 2

[6] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011. 1

[7] Shizhe Chen, Jia Chen, Qin Jin, and Alexander G. Hauptmann. Video captioning with guidance of multimodal latent topics. In *ACM MM*, pages 1838–1846, 2017. 2

[8] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *AAAI*, pages 8191–8198, 2019. 2

[9] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, pages 8199–8206, 2019. 3

[10] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K. Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv: 2001.09308*, 2020. 6

[11] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *NeurIPS*, pages 3063–3073, 2018. 1, 2, 3, 6, 7

[12] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *ECCV*, pages 768–784, 2016. 2

[13] Kuan Fang, Alexander Toshev, Fei-Fei Li, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *CVPR*, pages 538–547, 2019. 2

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017. 2, 6, 7

[15] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. WSLLN: weakly supervised natural language localization networks. In *EMNLP*, pages 1481–1487, 2019. 6

[16] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. MAC: mining activity concepts for language-based temporal localization. In *WACV*, pages 245–253, 2019. 2, 3

[17] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *ICCV*, pages 4203–4212, 2017. 1, 2

[18] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2132–2141, 2018. 3

[19] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *ICMR*, pages 217–225, 2019. 2

[20] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 1, 2, 6

[21] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019. 2, 4, 8

[22] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500, 2018. 1, 2, 6, 7

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[24] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, pages 988–996, 2017. 1, 2

[25] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, pages 11539–11546, 2020. 6

[26] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, pages 1298–1307, 2019. 2

[27] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 33, 2020. 2

[28] Xiang Long, Chuang Gan, and Gerard de Melo. Video captioning with multi-faceted attention. *TACL*, 6:173–184, 2018. 2

[29] 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*, 2020. 2

[30] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601, 2019. 2

[31] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10807–10816, 2020. 2, 3

[32] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, pages 6588–6597, 2019. 1, 2, 6, 7

[33] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, pages 6752–6761, 2018. 2

[34] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016. 1

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 6

[36] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *CVPR*, pages 8347–8356, 2019. 2

[37] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 7

[38] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *ACM MM*, pages 4280–4288, 2020. 2, 6, 7

[39] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, pages 8907–8916, 2019. 2, 3, 6

[40] Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. Team ruc_aim3 technical report at activitynet 2020 task 2: Exploring sequential events detection for dense video captioning. *arXiv preprint arXiv: 2006.07896*, 2020. 6, 7

[41] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[42] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, pages 3059–3067, 2017. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 4

[44] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[45] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *ICCV*, pages 4534–4542, 2015. 1, 2

[46] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, pages 7622–7631, 2018. 1

[47] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, pages 7190–7198, 2018. 1, 2, 6

[48] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: multimodal memory modelling for video captioning. In *CVPR*, pages 7512–7520, 2018. 2

[49] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 6402–6411, 2017. 2

[50] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019. 3, 7

[51] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 2

[52] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, pages 489–505, 2018. 2

[53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1

[54] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. STAT: spatial-temporal attention mechanism for video captioning. *IEEE TMM*, 22(1):229–241, 2020. 2

[55] Gao Yan, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-MIDN: coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *ICCV*, pages 9833–9842, 2019. 2

[56] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *ICCV*, pages 8371–8380, 2019. 2

[57] Ziwei Yang, Yahong Han, and Zheng Wang. Catching the temporal regions-of-interest for video captioning. In *ACM MM*, pages 146–153, 2017. 2

[58] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015. 1, 2, 5

[59] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593, 2016. 2

[60] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, pages 9159–9166, 2019. 2, 6

[61] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *CVPR*, pages 8327–8336, 2019. 1, 2

[62] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, pages 12870–12877, 2020. 6, 7

[63] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, pages 655–664, 2019. 2

[64] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106, 2020. 2

[65] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, pages 13275–13285, 2020. 2

[66] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2933–2942, 2017. 4, 6

[67] Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-aware action targeting for video captioning. In *CVPR*, pages 13093–13102, 2020. 1

[68] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018. 1, 2, 6