Dense Events Grounding in Video

Peijun Bao,[1] **Qian Zheng,**[2] **Yadong Mu**[1*]
[1]Peking University, [2]Nanyang Technological University
{peijunbao, myd}@pku.edu.cn, zhengqian@ntu.edu.sg

Abstract

This paper explores a novel setting of temporal sentence grounding for the first time, dubbed as *dense events grounding*. Given an untrimmed video and a paragraph description, dense events grounding aims to jointly localize temporal moments of multiple events described in the paragraph. Our main motivating fact is that multiple events to be grounded in a video are often semantically related and temporally coordinated according to their order appearing in the paragraph. This fact sheds light on devising more accurate visual grounding model. In this work, we propose Dense Events Propagation Network (DepNet) for this novel task. DepNet first adaptively aggregates temporal and semantic information of dense events into a compact set through a second-order attention pooling, then selectively propagates the aggregated information to each single event with soft attention. Based on such aggregation-and-propagation mechanism, DepNet can effectively exploit both the temporal order and semantic relations of dense events. We conduct comprehensive experiments on large-scale datasets ActivityNet Captions and TACoS. For fair comparisons, our evaluations include both state-of-art single-event grounding methods and their natural extensions to the dense-events grounding setting implemented by us. All experiments clearly show the performance superiority of the proposed DepNet by significant margins.

Introduction

Over the last few years, the computer vision community has witnessed the success of temporal sentence grounding, which aims to localize temporal moment described by a sentence description in a given video. A list of promising methods (Anne Hendricks et al. 2017; Gao et al. 2017; Wang, Ma, and Jiang 2020; Ghosh et al. 2019; Rodriguez et al. 2020; Wang, Huang, and Wang 2019a) have been proposed for temporal sentence grounding. Several recent works (Hendricks et al. 2018; Zhang, Su, and Luo 2019; Liu et al. 2018a; Stroud et al. 2019; Yuan et al. 2019; Zhang et al. 2020) further explore to localize more complicated sentence containing compositional activity like "the woman takes the book across the room to read it on the sofa" .

Most existing methods separately ground an individual event from a video, which we argue is not an optimal option for contextualized description of multiple events. Compared with single-sentence input, a paragraph of multiple sentences that describe video events in time order is seemingly more natural and powerful (Krishna et al. 2017; Li et al. 2018; Zhou et al. 2018). Consider the example in Figure 1, people may be interested in a list of events around the moment "the man with red shorts serves the ball", and they use a paragraph consisting of multiple sentences to describe these events. Furthermore, events like "the man with red shorts serves the ball." appear several times in the video, but people may only be interested in one of them. To avoid ambiguity, they use a paragraph to describe not only the most interested event but also its contextual events.

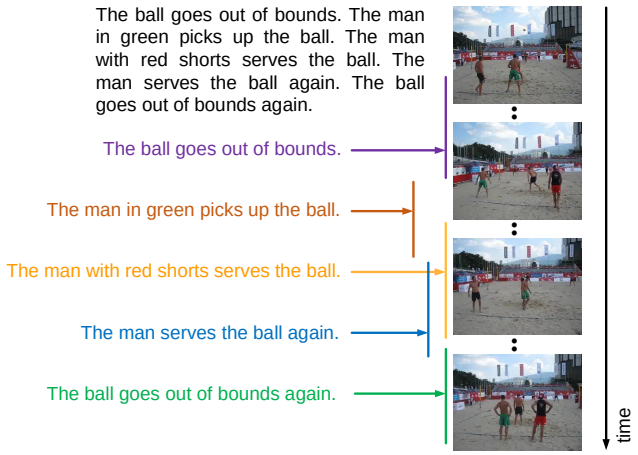

Figure 1: Dense Events Grounding in Video. Given a paragraph description, dense events grounding aims to jointly localize described dense events in untrimmed video.

To localize dense events in a paragraph, one can simply apply some single-event grounding model to each individual sentence in the paragraph. However, the temporal order of dense events occurred in video is often highly correlated with their locations in the descriptive paragraph, as demonstrated in relevant tasks such as dense captioning (Krishna et al. 2017; Regneri et al. 2013). Ignoring the temporal clues in a paragraph tends to lead inferior performance in precisely

*Corresponding author.

finding the temporal boundary of an event. To illustrate it, for a pair of events in a video we investigate the consistency between the visual grounding results and their temporal order in the paragraph. Surprisingly, even state-of-art single event grounding methods (Zhang et al. 2020) have a more than 20% chance to generate visual grounding results that contradict with the temporal order in the corresponding paragraph, which hints a huge space for improvement via contextual grounding.

Moreover, events described in a same paragraph are usually semantically related to one another. As shown in Figure 1, localizing sentence event "the man serves the ball again" requires to understand another sentence event "the man with red shorts serves the ball". Jointly grounding them can utilize their contextual semantic relation, which also contributes to precise time boundary prediction.

To this end, we introduce a novel setting of temporal sentence grounding for the first time in the literature, termed as *dense events grounding*. Given an untrimmed video and a paragraph of sentence descriptions, the goal of dense events grounding is to jointly localize temporal moments described by these sentence descriptions. Rather than grounding each event independently, dense event grounding requires to exploit temporal order and semantic relations of dense events for more accurate localization.

To achieve this, we propose a novel dense events grounding model called Dense Events Propagation Network (DepNet). Our main idea is to adaptively aggregate temporal and semantic information of dense events into a compact set, then selectively propagate the aggregated information to each single event. More specifically, DepNet first generates visual-semantic moment proposals for each single event in the paragraph description. A dense events aggregation module then aggregates these moment proposals into a compact set through a second-order attention pooling. For each moment proposal, a dense events propagation module then selects a desired subset of features from the compact set, and propagates the selected features to the proposal through soft attention. In this way, the moment proposals of each single event can perceive and exploit the temporal order information and context semantic relation from other events in the paragraph description.

Our contributions are summarized as follows:

- We define a new task *dense events grounding* and develop Dense Events Propagation Network (DepNet) as the first attempt of tackling this task. Particularly, DepNet adopts a novel aggregation-and-propagation scheme, which effectively enables context-guided visual grounding.
- Experiments on large-scale datasets ActivityNet Captions and TACoS show that the proposed DepNet outstrips several state-of-the-art single-event grounding methods and their dense-events variants (implemented by us for fair comparisons) by significant margins.

Related Work

Single Event Grounding

Temporal sentence grounding of single event in video is recently introduced by (Anne Hendricks et al. 2017; Gao et al. 2017), which aims to determine the start and end time points of single event given by a query sentence. (Anne Hendricks et al. 2017) proposes a moment context network to jointly model text query and video clips. (Gao et al. 2017) proposes cross-modal localizer to regress action boundary for candidate video clips. (Liu et al. 2018b,c) then advice to apply attention mechanism to highlight the crucial part of visual features or query contents. (Wang, Huang, and Wang 2019b) then develops a semantic matching reinforcement learning framework to reduce the large visual-semantic discrepancy between video and language.

Several recent works (Zhang et al. 2019a, 2020; Wang, Ma, and Jiang 2020) propose to model temporal dependencies within sentence to closely integrate language and video representation. (Zhang et al. 2019a) models temporal dependencies as a structured graph and devises an iterative graph adjustment network for temporal structural reasoning. (Zhang et al. 2020) proposes 2D Temporal Adjacent Networks to model the temporal relations between video moments by a two-dimensional map. (Wang, Ma, and Jiang 2020) uses a lightweight semantic boundaries prediction branch to aggregate contextual information and models the relationship between the referent and its neighbors.

Some recent works (Zhang, Su, and Luo 2019; Stroud et al. 2019; Zhang et al. 2019b) further utilize compositional property of query sentence and decompose sentence as multiple components for better temporal reasoning. (Zhang, Su, and Luo 2019) proposes temporal compositional network where a tree LSTM decomposes a sentence into three description main event, context event. Similarly, (Stroud et al. 2019) first grounds atomic sub-events to short video segments and then establishes the temporal relationships between these segments. (Zhang et al. 2019b) develops a syntactic Graph Convolution Network to leverage the syntactic structure of sentence and a multi-head self-attention module to capture long-range dependencies from video context.

Although compositional activity may be considered, only single event with single sentence description is grounded in the settings of existing works. Unlike this, the proposed dense events grounding aims to jointly localize multiple events described by a paragraph, which requires to the model temporal order and semantic relations of the dense events.

Dense Events Understanding in Video

Understanding dense events in video has been popular in recent years. (Krishna et al. 2017) introduces the task of dense video captioning which aims to both detecting and describing dense events in a video. They use contextual information from past and future events to jointly describe all events. (Li et al. 2018; Zhou et al. 2018) propose a joint and global optimization framework of detection and captioning in an end-to-end manner for dense video captioning. (Wang et al. 2018) develops a hierarchical reinforcement learning algorithm for dense video captioning where a high-level manager module learns to design sub-goals and a low-level worker module recognizes the primitive actions to fulfill the sub-goal. (Duan et al. 2018) further extends dense events captioning in a weakly supervised setting and formulate the problem as dual process of event captioning and sen-

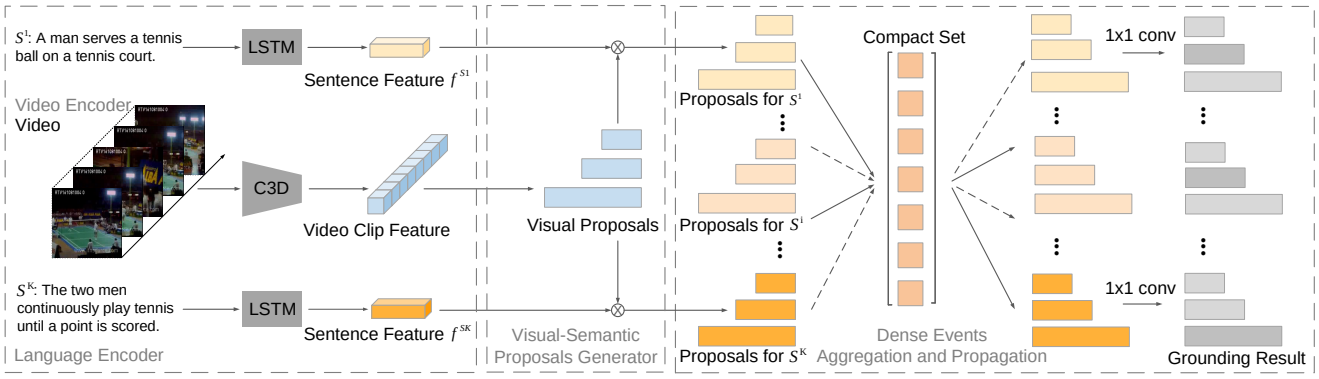

Figure 2: The framework of our proposed Dense Events Propagation Network. It consists of a Language Encoder, a Video Encoder, a Visual-Semantic Proposals Generator, a Dense Events Aggregation and Propagation Module. In the Dense Events Aggregation and Propagation Module, visual-semantic information of dense events is aggregated into a compact set, then selectively propagated to each single event. Only the first and last sentence queries are visualized.

tence localization. The dense captioning task in these papers is to describe dense events in the video with a paragraph. In contrast, our dense events grounding task can be viewed as the inverse problem of dense captioning.

(Bojanowski et al. 2015) proposes to grounding multiple sentences in video with weakly-supervised settings. These sentence events are assumed to not overlap with each other, which does not generalize to most video paragraph descriptions. (Shao et al. 2018) proposes a novel task to retrieve paragraph query in video collections and then localize these events. They propose a find-and-focus framework where the top-level matching narrows the search while the part-level localization refines the results. Our task is different from their events localization step since we jointly localize dense events within the video while they treat each event independently.

Method

Problem Formulation

Given an untrimmed video V and K sentence descriptions $\{S_1, S_2, \cdots, S_K\}$ with temporal order, our goal is to jointly localize temporal moments $\{T_1, T_2, \cdots, T_K\}$ described by these sentences. More specifically, the video is presented as a sequence of frames $V = \{v_i\}_{i=1}^{L_V}$ where v_i is the feature of i-th frame and L_V is the frame number of the video. The k-th sentence description S_k is presented as $S_k = \{s_{ki}\}_{i=1}^{L_k}$ where s_{ki} represents i-th word in the sentence and L_k denotes the total number of words. The k-th temporal moment T_k consists of start and end time point of the event in the video.

Dense Events Propagation Network

As illustrated in Figure 2, our proposed Dense Events Propagation Network (DepNet) consists of four main components: a language encoder, a video encoder, a visual-semantic proposal generator, and a dense-events aggregation-and-propagation module. This section will elaborate on the details of each component.

Language Encoder Given an input of a natural language paragraph query, the goal of language encoder is to encode the sentences in the paragraph such that moments of interest can be effectively retrieved in the video. Our language encoder extracts feature embedding f^{S_k} of each sentence S_k in the paragraph descriptions $\{S_1, S_2, \cdots, S_K\}$ separately.

Instead of encoding each word with a one-hot vector or learning word embeddings from scratch, we rely on word embeddings obtained from a large collection of text documents. In more details, each word s_{ki} in the sentence S_k is first encoded into Glove word embedding (Jeffrey Pennington and Manning 2014) as w_{ki}. Then the sequence of word embedding $\{w_{ki}\}_{i=1}^{L_k}$ is fed to an LSTM (Hochreiter and Schmidhuber 1997). The last hidden state of LSTM is passed to a single fully-connected layer to extract the final sentence feature $f^{S_k} \in \mathbb{R}^{d^S}$. All parameters of the LSTM and fully-connected layer are shared across different sentences in the paragraph.

Video Encoder Video encoder aims to obtain high-level visual representations of video moment proposals from raw input frames. Specifically, the input video is first segmented into small clips where each video clip contains T frames. A fixed-interval sampling is performed over these video clips to obtain N video clips. For each sampled video clip, we extract a sequence of basic C3D features $V = \{v_i\}_{i=1}^{N}$ with a pretrained C3D (Tran et al. 2015) Network.

The visual feature embeddings for moment proposals are constructed from these basic C3D features. For a moment proposal (a, b) with start point at a and end point at b, we apply boundary-matching operation BM (Lin et al. 2019) over all C3D features covered by this proposal to get the feature embedding:

$$\tilde{f}^{V_{ab}} = \mathrm{BM}(\{v_i\}_{i=a}^{b}). \quad (1)$$

The boundary-matching operation can efficiently generate proposal-level feature from basic clip-level feature, through a series of bilinear sampling and convolutional operations. More algorithmic details are omitted here and can be re-

ferred to (Lin et al. 2019). $\tilde{f}^{V_{ab}}$ is passed through a fully-connected layer to obtain the final feature embedding $f^{V_{ab}} \in \mathbb{R}^{d^V}$ for the moment proposal (a, b). Essentially, this extracted feature $f^{V_{ab}}$ summarizes spatial-temporal patterns from raw input frame and thus represents the visual structure of the moment proposal.

Visual-Semantic Proposals Generator Visual-semantic proposals generator constructs visual-semantic features of moment proposals for each sentence query in the paragraph description. Specifically, video moment feature $f^{V_{ab}}$ for all possible moment proposals are computed according to Eq (1) where $1 \leq a \leq b \leq N$.

The features of the visual modality and language modality are then fused to generate visual-semantic for each sentence in the paragraph. To interact the language feature of k-th sentence f^{S_k} with video moment feature $f^{V_{ab}}$, we multiply f^{S_k} with video moment clip feature $f^{V_{ab}}$ and then normalize the fused feature $\hat{M}^k_{ab}$ with its $\mathcal{L}_2$ norm, namely

$$\begin{aligned} \hat{M}^k_{ab} &= f^{V_{ab}} \odot f^{S_k}, \\ M^k_{ab} &= \hat{M}^k_{ab} / ||\hat{M}^k_{ab}||_2, \end{aligned} \tag{2}$$

where $\odot$ denotes Hadamard product.

Inspired by positional encoding of tokens in natural language processing (Vaswani et al. 2017; Devlin et al. 2018), we encode the relative positions of each proposal to better perceive the temporal information. We consider three sorts of the relative positions, *i.e.*, the start point a, end point b and sentence order k. These positions are encoded by sine and cosine functions of different frequencies:

$$\mathrm{PE}_{\mathrm{pos,i}} = \begin{cases} \sin(\mathrm{pos}/10000^{i/d_{\mathrm{pos}}}), & \text{if i is even} \\ \cos(\mathrm{pos}/10000^{i/d_{\mathrm{pos}}}), & \text{otherwise} \end{cases} \tag{3}$$

where pos can be any one of the three relative positions, d is the number of dimensions and i denotes the i-th dimension.

Then these three sorts of positional feature $PE \in \mathbb{R}^{3d_{\mathrm{pos}}}$ are concatenated with M^k_{ab} and transformed by 1×1 convolutional layer. The output of the convolutional layer summarizes visual-semantic patterns and relative positional information of the proposals for each sentence query in the paragraph. For simplicity, we still refer it as M^k_{ab}.

Since the number of moment proposals is large, following (Zhang et al. 2020) we adopt sparse sampling strategy to remove the redundant moment proposals which have large overlaps with the selected one. Such sampling strategy can effectively reduce the computational cost and the number of moment proposals.

Dense Events Aggregation and Propagation Dense event grounding requires to exploit temporal order and semantic relations of sentence events for more accurate localization. Inspired by global feature modelling in image/video classification (Chen et al. 2018), we design dense events aggregation and propagation modules. Specifically, a dense events aggregation module first perceives the entire visual-semantic proposals of sentence queries, and adaptively aggregates global information through a second-order attention pooling. Then a dense events propagation module is cascaded to selectively propagates the aggregated global information to each event proposal with soft attention.

In more details, the dense events aggregation module adaptively aggregates visual-semantic feature $M = \{M^k_{ab}\}(1 \leq a \leq b \leq N, 1 \leq k \leq K)$ of entire moment proposals into a compact set of global features $A = \{\mathbf{a}_1, \cdots, \mathbf{a}_n\}$. To do this, it first transforms M to $G = \{\mathbf{g}_1, \cdots, \mathbf{g}_n\}$ by a convolutional layer g as attention weight to densely attend each moment proposal for global information aggregation in M. Each $\mathbf{a}_i$ is then calculated by second-order attention pooling operation as:

$$\mathbf{a}_i = f(M)\,\mathrm{softmax}(\mathbf{g_i})^\top, \tag{4}$$

where f is another convolutional layer to transform M and $\mathrm{softmax}$ is used to normalized $\mathbf{g_i}$ to a valid attention weight. Above second-order attention pooling adaptively selects informative moment proposals among multiple sentence events and aggregates visual-semantic information of them.

Then the dense events propagation module selectively propagates the global features in the compact set A to each proposal via soft attention. More specifically, a subset of feature vectors are dynamically selected from A for each proposal and propagated to the proposal with soft attention as:

$$\tilde{M}^k_{ab} = \sum_{j=1}^{n} \mathbf{h}^k_{abj} \mathbf{a}_j = A\mathbf{h}_{ab}, \tag{5}$$

where $\mathbf{h}$ is the attention weight and satisfies $\sum_{j=1}^{n} \mathbf{h}^k_{abj} = 1$. Similar to the generation of the attention weight $\mathbf{g}$ in the aggregation module, $\mathbf{h}$ is generated by applying a convolution layer and a followed softmax normalizer on M. In this way, each moment proposal can perceive moment contexts from other sentence events which are complementary to itself.

Such an aggregation and propagation design enables the model to not only perceive the entire moment proposals of the event itself, but also perceive the moment contexts from other events, resulting in learning the distribution of the multi-events.

Finally we pass the output of propagation module $\{\tilde{M}^k_{ab}\}$ to a fully-connected layer and a sigmoid layer to generate a temporal-sentence score map $\{p^k_{ab}\}$. And each value $p^k_{a,b}$ in temporal-sentence score map denotes predicted matching score of the temporal moment (a, b) for k-th sentence. The maximum of k-th score map p^k corresponds to the grounding result for the k-th sentence S_k.

Training Loss

Our training sample consists of an input video, a paragraph query $\{S_1, S_2, \cdots, S_K\}$ and a set of temporal annotations $\{T_1, T_2, \cdots, T_K\}$ associated with the sentences in the paragraph. During training, we need to determine which temporal moment in the temporal-sentence score map corresponds to the annotations and train the model accordingly.

Instead of hard label, we assign each moment proposal with a soft label according to its overlap with the annotations. Specifically, for each moment in the temporal-sentence score map, we compute the IoU score IoU^k_{ab} between its temporal boundary (a, b) and the annotation of the

k-th sentence T_k. Then a soft ground truth label gt_{ab}^k is assigned to it according to IoU_{ab}^k:

$$gt_{ab}^k = \begin{cases} 0 & IoU_{ab}^k \leq \mu_{min}, \\ \frac{IoU_{ab}^k - \mu_{min}}{\mu_{max} - \mu_{min}} & \mu_{min} < IoU_{ab}^k < \mu_{max}, \\ 1 & IoU_{ab}^k \geq \mu_{max}, \end{cases} \quad (6)$$

where μ_{min} and μ_{max} are two thresholds to customize the distribution of soft labels.

For each training sample, the model can be trained in an end-to-end manner with a binary cross entropy loss, which is defined as:

$$\mathcal{L} = -\sum_{(a,b)\in\mathcal{C}} \mathrm{gt}_{ab}^k \log(p_{ab}^k) + (1 - \mathrm{gt}_{ab}^k) \log(1 - p_{ab}^k), \quad (7)$$

where $\mathcal{C} = \{(a,b) | 1 \leq a \leq b \leq N\}$ is the set of all valid moment proposal boundaries.

Experiments

Dataset

ActivityNet Captions ActivityNet Captions (Krishna et al. 2017) consists of 19,209 untrimmed videos. Each video includes multiple sentence descriptions with temporal order and corresponding moment boundary annotations. The contents of video are diverse and open. It is originally built for dense-captioning events (Krishna et al. 2017) and lately introduced for temporal grounding with single sentence setting. For fair comparison, following the experimental setting in single sentence grounding (Zhang et al. 2020; Yuan et al. 2019), we use val_1 as validation set and val_2 as testing set. There are 37,417, 17,505, and 17,031 moment-sentence pairs in the training, validation and testing set, respectively.

TACoS TACOS (Regneri et al. 2013) consists of 127 videos. Each video has several paragraph descriptions and temporal annotations. It is developed on MPII Compositive (Rohrbach et al. 2012). The main video theme is limited to cooking scenes, thus lacking diversity. Compared with ActivityNet Captions, Videos in the TACoS benchmark generally has longer duration and shorter moments. Following the standard data splitting, there are totally 10,146, 4,589 and 4,083 moment-sentence pairs in the training, validation and testing set, respectively.

Evaluation Metrics

The commonly-adopted evaluation metric in single-event grounding (Zhang et al. 2020; Yuan et al. 2019) is known to be "Recall@N,IoU=θ ". For each sentence query in the paragraph, we calculate the Intersection over Union (IoU) between grounded temporal segment and the ground truth. "Recall@N,IoU=θ " represents the percentage of top N grounded temporal segments that have at least one segment with higher IoU than θ. For ease of comparisons, we borrow the identical evaluation metric in our proposed novel multi-event setting. There are specific settings of N and θ for different datasets. To fairly compare with previous single event grounding methods, we follow (Zhang et al. 2020; Yuan et al. 2019) to report the results as $N \in \{1, 5\}$ with $\theta \in \{0.3, 0.5, 0.7\}$ for ActivityNet Captions dataset, and $N \in \{1, 5\}$ with $\theta \in \{0.1, 0.3, 0.5\}$ for TACoS dataset.

Implementation Details

For fair comparison, we use pretrained CNN (Tran et al. 2015) as previous methods to extract C3D video features on both datasets. And we use Glove (Jeffrey Pennington and Manning 2014) word embeddings pretrained on Common Crawl to represent each word in the sentences. A three-layer LSTM is applied to word-embeddings to obtain the sentence representation. The channel numbers of sentence feature and video proposal feature d^S, d^V are all set to 512 . We set the dimension of positional feature d_{pos} to 128 and the size of compact set n to 512. The number of sampled clips N is set to $32, 64$ for ActivityNet Captions and TACoS respectively. For BM operations in the video encoder, we set sampling number of each proposal to $16, 32$ for ActivityNet Captions and TACoS respectively.

During training, We use Adam (Kingma and Ba 2014) with learning rate of 1×10^{-4}, the momentum of 0.9 and batch size of 4 as optimization algorithm. For each training sample, we randomly sample K ($2 \leq K \leq 8$) sentence queries with temporal order and apply padding them with zeros to 8 sentences. Such random sampling strategy reduces model to overfit to moment prior of dense events. We only calculate the values on the valid location and do not take zero-padding into calculation. For binary cross entropy loss, the scaling thresholds μ_{min} and μ_{max} are set to 0.5 and 1.0 for ActivityNet Captions and 0.3 and 0.7 for TACoS.

During inference, we set the whole paragraph as query description if the annotated paragraph has less than 8 sentences. For annotated paragraph queries with more than 8 sentences, we split them to multiple sub-paragraphs to meet the limitation of maximum 8 sentences. We independently choose the moment proposals with the highest confidence score for each sentence as final result. If we need to select multiple moment localizations per sentence (i.e. for R@5), Non Maximum Suppression (NMS) with a threshold of 0.5 is applied to remove redundant candidates.

Competing Methods

In this subsection, we compare the proposed DepNet model with state-of-the-art methods of single event grounding and two competitive baseline models of dense sentence grounding. We refer the proposed model as **DepNet**.

The compared single event grounding methods are listed as followings. **CTRL** (Gao et al. 2017): Cross-model Temporal Regression Localizer. **MCN** (Anne Hendricks et al. 2017): Moment Context Network. **ACRN** (Liu et al. 2018b): Attentive Cross-Model Retrieval Network. **QSPN** (Xu et al. 2019): Multilevel Language and Vision Integration. **ACL-K** (Ge et al. 2019): Activity Concepts based Localizer. **GDP** (Chen et al. 2020): Graph-FPN with Dense Predictions. **SAP** (Chen and Jiang 2019): A two-stage approach based on visual concept mining. **SCDM** (Yuan et al. 2019): Semantic Conditioned Dynamic Modulation. **CBP** (Wang, Ma, and Jiang 2020): Contextual Boundary-aware Prediction. **CMIN** (Zhang et al. 2019b): Cross-Modal Interaction Networks. **2D-TAN** (Zhang et al. 2020): 2D Temporal Adjacent Network.

The current literature of temporal grounding is dominated by single-event methods. For fair comparison, we carefully

Table 1: Performance Evaluation Results on the ActivityNet Captions Dataset ($N \in \{1, 5\}$ and $\theta \in \{0.3, 0.5, 0.7\}$).

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
MCN	39.35	21.36	6.43	68.12	53.23	29.70
CTRL	47.43	29.01	10.34	75.32	59.17	37.54
ACRN	49.70	31.67	11.25	76.50	60.34	38.57
QSPN	52.13	33.26	13.43	77.72	62.39	40.78
GDP	56.17	39.27	–	–	–	–
CBP	54.30	35.76	17.80	77.63	65.89	46.20
SCDM	54.80	36.75	19.86	77.29	64.99	41.53
CMIN	63.61	43.40	23.88	80.54	67.95	50.73
2D-TAN	59.45	44.51	26.54	85.53	77.13	61.96
BS	62.53	46.43	27.12	–	–	–
3D-TPN	67.56	51.49	30.92	87.94	81.53	65.86
DepNet	**72.81**	**55.91**	**33.46**	**90.08**	**83.82**	**68.80**

Table 2: Performance Evaluation Results on the TACoS Dataset ($N \in \{1, 5\}$ and $\theta \in \{0.1, 0.3, 0.5\}$).

Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
MCN	3.11	1.64	1.25	3.11	2.03	1.25
CTRL	24.32	18.32	13.30	48.73	36.69	25.42
ACRN	24.22	19.52	14.62	47.42	34.97	24.88
QSPN	25.31	20.15	15.23	53.21	36.72	25.30
ACL-K	31.64	24.17	20.01	57.85	42.15	30.66
SAP	31.15	–	18.24	53.51	–	28.11
GDP	39.68	24.14	–	–	–	–
CBP	–	27.31	24.79	–	43.64	37.40
SCDM	–	26.11	21.17	–	40.16	32.18
CMIN	32.48	24.64	18.05	62.13	38.46	27.02
2D-TAN	47.59	37.29	25.32	70.31	57.81	45.04
BS	48.46	38.14	25.72	–	–	–
3D-TPN	55.05	40.31	26.54	78.18	63.60	48.23
DepNet	**56.10**	**41.34**	**27.16**	**79.59**	**64.74**	**48.75**

devise two competitive baseline models for dense-events grounding, including **Beam Search (BS)** and **3D Temporal-Paragraph Network (3D-TPN)** (a natural extension of the state-of-the-art single-event model 2D-TAN fully implemented by us). In particular, the BS model first localizes each sentence in the paragraph independently with a base single event grounding model, then applies beam search on the top k grounding results of each event as post processing, such that final dense events grounding results match the temporal order. We set k to 8 in the experiments and choose 2D-TAN as the base model for constructing a strong baseline in dense event grounding. In specific, 3D-TPN model first constructs two-dimensional temporal-sentence feature map as in 2D-TAN for each sentence in paragraph, formulates them as three-dimensional temporal-paragraph map according to the sentence order and then applies a stack of 3D-convolutional layers on the formulated 3D temporal-paragraph map to perceive the temporal order and semantic relations among dense events. Last, a fully-connected layer is applied on the output of 3D-convolution to get the dense events grounding result. Since 3D-TPN is a natural extension of 2D-TAN, more details of its implementation are omitted here and can be referred to (Zhang et al. 2020).

Performance Comparisons

Table 1 and Table 2 show the performance comparisons between DepNet and all above-mentioned baseline methods on ActivityNet Captions and TACoS, respectively. From the tables, all three proposed dense events grounding methods outperform single event grounding methods with a clear margin. This verifies the superiority of the proposed dense events grounding setting. Furthermore, DepNet achieves the best performance among all the methods, which verifies the effectiveness of dense events aggregation and propagation mechanism in DepNet.

In more details, DepNet achieves about 22.3%, 25.6% and 23.4% higher evaluation scores than the best single event grounding method 2D-TAN in term of R1@0.3, R1@0.5 and R1@0.7 on ActivityNet Captions. And DepNet achieves 2D-TAN about 17.9%, 10.9% and 7.3% higher scores than 2D-TAN in term of R1@0.1, R1@0.3 and R1@0.5 on TACoS. The other two dense events grounding models also achieves much better performance than the state-of-the-art single event grounding methods on both datasets. For example, BS and 3D-TPN achieves 3 and 8 points higher than the best single event grounding method 2D-TAN in term of R1@0.3 on ActivityNet Captions, and 1 and 7 points higher than 2D-TAN in term of R1@0.1 on TACoS. This verifies the superiority of the proposed dense events grounding setting, which can use the information temporal order and semantic relations among the dense events compared to existing single event grounding setting.

Moreover, we compare dense events grounding models with intra-sentence context based single event grounding models, including QSPN, GDP, CBP and CMIN. These approaches explicitly model the context moment within the single sentence and achieves better results than the simple sliding window based single events grounding methods, *i.e.*, CTRL, MCN and ACL-K. However, these approaches achieve clearly inferior results than the proposed DepNet. For example, even the best one of them achieves about 9 points lower than DepNet in term of R1@0.3 on ActivityNet Captions, and about 8 points lower than DepNet in term of R1@0.1 on TACoS. The performance gap serves as strong evidence for the necessity of joint multi-event grounding.

Last, we compare DepNet with two dense events grounding baselines BS and 3D-TPN. In term of R1@0.3, DepNet outperforms BS and 3D-TPN more than 10 and 3 points on ActivityNet Captions, and more than 6 and 1 points on TACoS, respectively. In term of R5@0.3, DepNet outperforms 3D-TPN more than 2 points on ActivityNet Captions, and more than 1 points on TACoS. As stated before, 3D-TPN is a natural extension of 2D-TAN. It adopts vanilla stacked temporal convolutions for perceiving information of dense events on the temporal-paragraph map. This represents the standard treatment in the literature of modern video analysis methods. We attribute its inferior performance to the relatively low efficacy in cross-event communication (*i.e.*, proposals with distant locations on the map can only be perceived after several layers of temporal convolutions). In contrast, our proposed aggregation-and-propagation scheme in DepNet treats different proposals equally and leads to better usage of information from dense events.

Ablation Study

In this section, we conduct ablation studies on ActivityNet Captions to analyze the contributions of our proposed Dep-

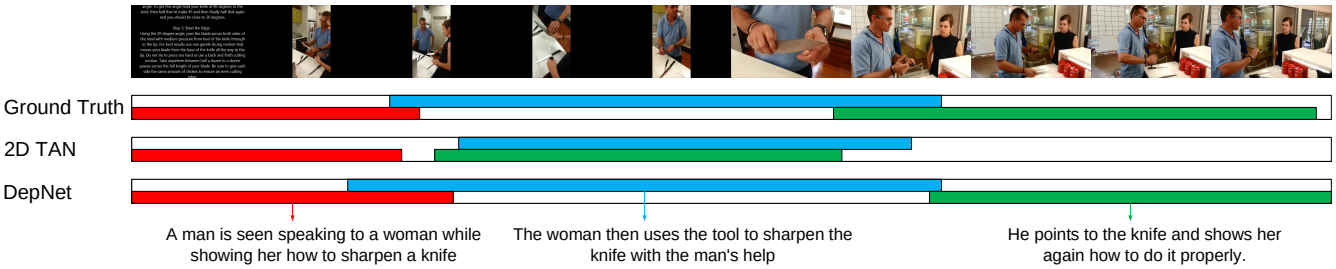

Figure 3: Qualitative prediction examples of our proposed model. The first row shows the ground-truths for the given paragraph queries, and the second and third row shows the grounding results of 2D-TAN and our DepNet.

Table 3: Performance evaluation results of ablation model on the ActivityNet Captions dataset.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
w/o. AP	62.86	46.76	28.08	86.19	78.16	60.52
w/o. PE	58.29	43.63	26.64	85.66	76.58	60.12
w/o. AP, PE	58.35	42.99	25.12	85.41	76.49	59.95
full	72.81	55.91	33.46	90.08	83.82	68.80

Net method. Specifically, we re-train our model with the following settings 1) DepNet (w/o. AP): the dense events aggregation and propagation module is dropped, 2) DepNet (w/o. PE): the positional encoding of moment proposals is dropped, 3) DepNet (w/o. AP,PE): both positional encoding of moment proposals and dense events aggregation and propagation module are dropped.

Table 3 shows the performance comparisons of our proposed full model DepNet (full) with respect to these ablations on the ActivityNet Captions dataset. DepNet (full) outperforms all ablation models on ActivityNet Captions, which demonstrates the positional encoding, dense events aggregation and propagation module is critical to dense events grounding in videos. Without considering the dense events aggregation and propagation module, the performance of the model DepNet (w/o. AP) degenerates dramatically compared to DepNet (full). It shows that dense events aggregation and propagation module is effective to model temporal order and semantic relations of dense events. DepNet (w/o. PE) improves the grounding performance compared to DepNet (w/o. AP,PE), shows that modeling the semantic dependencies between moment proposals can help. However, the improvement is limited compared to DepNet (full). This is mainly because without positional encoding, the temporal information of the moment proposals is dropped by the process of dense events aggregation. This also verifies the importance of modelling the temporal order of dense events.

Moreover, we explore the impact of the number of sentences in the paragraph descriptions as shown in Figure 4. Specifically, we evaluate the same well-trained DepNet on the ActivityNet Captions Dataset with different number of sentences in the paragraph descriptions. Here we select "R@1,IoU=0.3" and "R@1,IoU=0.5" as evaluation metrics. As can be seen, both metrics achieve higher values when the number of sentences increases. This shows that the temporal

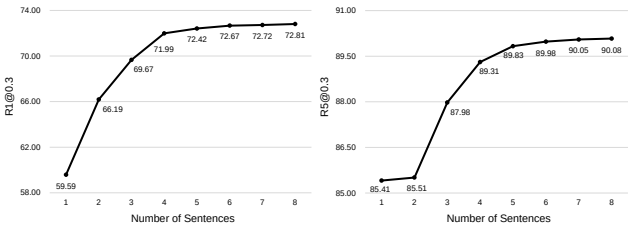

Figure 4: Impact of the number of sentences in the paragraph descriptions on the ActivityNet Captions Dataset.

order and semantic relations of dense events can help each other to obtain more accurate grounding. We further note that when the number of sentences is larger than 5, the improvement is almost saturated. This could be due to that sentences with more intervals are less semantically relevant and provide less information to help the dense events grounding.

Qualitative Analysis

To qualitatively validate the effectiveness of the DepNet method, we display several typical examples of dense events grounding. Figure 3 shows the grounding results of the DepNet method and the single event grounding method 2D-TAN on ActivityNet Captions. DepNet is capable of grounding a diverse set of events including the one requiring strong temporal dependencies with other sentences "he points to the knife and shows her again how to do it properly". DepNet can exploit the temporal order and semantic relations of dense events based on the aggregation and propagation mechanism. This makes it perform better than the 2D-TAN.

Conclusion

This work introduces a novel task dubbed as dense events grounding, a more challenging task than traditional single-event event grounding. To effectively capture the temporal context in the accompanying paragraph of a video, we here propose DepNet that has the unique trait of an aggregation-and-propagation scheme. Evaluations on two large-scale video benchmarks against a large spectrum of baselines (both state-of-the-art single-event models and their natural extensions to the multi-event setting) clearly demonstrate the superiority of DepNet. We strongly believe that our pilot research on this novel task will inspire more works on temporal context modeling in visual grounding.

Acknowledgement: The work is supported by National Natural Science Foundation of China (61772037), Beijing Natural Science Foundation (Z190001) and Tencent AI Lab Rhino-Bird Focused Research Program (JR202021).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*.

Bojanowski, P.; Lajugie, R.; Grave, E.; Bach, F.; Laptev, I.; Ponce, J.; and Schmid, C. 2015. Weakly-supervised alignment of video with text. In *ICCV*.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020. Rethinking the Bottom-Up Framework for Query-Based Video Localization. In *AAAI*.

Chen, S.; and Jiang, Y.-G. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. In *AAAI*.

Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; and Feng, J. 2018. Aˆ 2-nets: Double attention networks. In *NeurIPS*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. In *NeurIPS*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Ge, R.; Gao, J.; Chen, K.; and Nevatia, R. 2019. MAC: Mining Activity Concepts for Language-based Temporal Localization. In *WACV*.

Ghosh, S.; Agarwal, A.; Parekh, Z.; and Hauptmann, A. 2019. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755* .

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2018. Localizing moments in video with temporal language. In *EMNLP*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* .

Jeffrey Pennington, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*.

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *CVPR*.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*.

Liu, B.; Yeung, S.; Chou, E.; Huang, D.-A.; Fei-Fei, L.; and Carlos Niebles, J. 2018a. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*.

Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018b. Attentive moment retrieval in videos. In *SIGIR*.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018c. Cross-modal moment localization in videos. In *ACM MM*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL* .

Rodriguez, C.; Marrese-Taylor, E.; Saleh, F. S.; Li, H.; and Gould, S. 2020. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention. In *WACV*.

Rohrbach, M.; Regneri, M.; Andriluka, M.; Amin, S.; Pinkal, M.; and Schiele, B. 2012. Script data for attribute-based recognition of composite activities. In *ECCV*.

Shao, D.; Xiong, Y.; Zhao, Y.; Huang, Q.; Qiao, Y.; and Lin, D. 2018. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*.

Stroud, J. C.; McCaffrey, R.; Mihalcea, R.; Deng, J.; and Russakovsky, O. 2019. Compositional Temporal Visual Grounding of Natural Language Event Descriptions. *arXiv preprint arXiv:1912.02256* .

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI*.

Wang, W.; Huang, Y.; and Wang, L. 2019a. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*.

Wang, W.; Huang, Y.; and Wang, L. 2019b. Language-driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *CVPR*.

Wang, X.; Chen, W.; Wu, J.; Wang, Y.-F.; and Yang Wang, W. 2018. Video captioning via hierarchical reinforcement learning. In *CVPR*.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *AAAI*.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NeurIPS*.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.

Zhang, S.; Su, J.; and Luo, J. 2019. Exploiting Temporal Relationships in Video Moment Localization with Natural Language. In *ACM MM*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *CVPR*.