Weakly-Supervised Video Moment Retrieval via Semantic Completion Network

Zhijie Lin,[1] **Zhou Zhao,**[1]* **Zhu Zhang,**[1] **Qi Wang,**[2] **Huasheng Liu**[2]
[1]College of Computer Science, Zhejiang University, Hangzhou, China,
[2]Alibaba Inc., China
{linzhijie, zhaozhou, zhangzhu}@zju.edu.cn, {wq140362, fangkong.lhs}@alibaba-inc.com

Abstract

Video moment retrieval is to search the moment that is most relevant to the given natural language query. Existing methods are mostly trained in a fully-supervised setting, which requires the full annotations of temporal boundary for each query. However, manually labeling the annotations is actually time-consuming and expensive. In this paper, we propose a novel weakly-supervised moment retrieval framework requiring only coarse video-level annotations for training. Specifically, we devise a proposal generation module that aggregates the context information to generate and score all candidate proposals in one single pass. We then devise an algorithm that considers both exploitation and exploration to select top-K proposals. Next, we build a semantic completion module to measure the semantic similarity between the selected proposals and query, compute reward and provide feedbacks to the proposal generation module for scoring refinement. Experiments on the ActivityCaptions and Charades-STA demonstrate the effectiveness of our proposed method.

Introduction

Video moment retrieval, a key topic in information retrieval and computer vision, has attracted more and more interests in recent years (Gao et al. 2017; Hendricks et al. 2017). As two examples in Figure 1 show, according to a given natural language query, moment retrieval aims to locate the temporal boundary of the most related moment in the video, which can help us quickly filter out useless contents in the video. More accurate moment retrieval requires sufficient understanding of both the video and the query, which makes it a challenging task. Although recent works (Chen et al. 2018; Zhang et al. 2019b; 2019a) has achieved good results, they are mostly trained in a fully-supervised setting, which requires the full annotations of temporal boundary for each video. However, manually labeling the ground truth temporal boundaries is time-consuming and expensive, requiring a large amount of human labor. Moreover, considering an untrimmed video contains multiple consecutive temporal activities, it can be difficult to mark the boundaries accurately, which produces ambiguity and noise in training data.

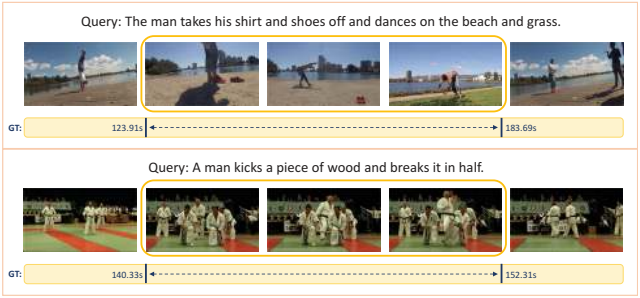


Figure 1: Examples of video moment retrieval: search the temporal boundary of the most relevant moment in video according to the given natural language query.

Relatively, it is much easier to obtain coarse descriptions of a video without marking the temporal boundaries, such as the captions of videos in YouTube. This motivates us to develop a weakly-supervised method for moment retrieval that needs only coarse video-level annotations for training. Existing weakly-supervised method in (Mithun, Paul, and Roy-Chowdhury 2019) proposes to learn a joint visual-text embedding, and utilizes the latent alignment produced by intermediate Text-Guided Attention (TGA) to localize the target moment. However, the latent attention weights without extra supervision usually focus on the most discriminative but small regions (Singh and Lee 2017) instead of covering complete regions. To deal with these issues, in this paper, we devise a novel weakly-supervised Semantic Completion Network (SCN) including proposal generation and selection, semantic completion for semantic similarity estimation and scoring refinement.

Firstly, rather than localizing the most relevant moment relying on the ambiguous attention weights, we extract the semantically important proposals through a proposal generation module. Further than method in (Gao et al. 2017) that treats each candidate proposal separately, we leverage the cross-modal fusion representations of video and query to score all the candidate proposals sampled at different scales in a single pass, which makes full use of context information for scoring other proposals.

*Corresponding author.

With a large set of densely sampled proposals, we then devise an algorithm that considers both exploitation and exploration to select top-K proposals. Concretely, we first rank all candidate proposals based on their corresponding confidence scores. Further than just selecting the proposals with high confidence score based on Non Maximum Suppression (NMS), to encourage full exploration, we select next proposal randomly with a decay possibility, which is helpful for finding potentially good proposals and giving more accurate confidence scores for proposals. At the beginning of training, we tend to select next proposal randomly for exploration. As the model converges gradually, proposals with high confidence score are chosen more often for exploitation.

To explicitly model the scoring of proposal generation module rather than rely on attention weights without extra supervision (Mithun, Paul, and Roy-Chowdhury 2019), we are supposed to further measure the semantic similarity between the selected proposals and query for scoring refinement. Inspired by the success of recent works about masked language model (Devlin et al. 2019; Song et al. 2019; Wang, Li, and Smola 2019), we design a novel semantic completion module that predicts the important words (e.g. noun, verb) that are masked according to the given visual context. In detail, by masking the important words in the decoder side, SCN forces the decoder rely on the visual context to reconstruct the query and the most semantically matching proposal can provide enough information to predict the key words. Then with the evaluation results given by semantic completion module, we compute reward for each proposal based on the reconstruction loss and formulate a rank loss to encourage the proposal generation module to give higher confidence score for those proposals with greater rewards.

In total, the main contributions of our work are listed as follows:

- We propose a novel weakly-supervised moment retrieval framework requiring only coarse annotations for training, and experiments on two datasets: ActivityCaptions (Caba Heilbron et al. 2015) and Charades-STA (Gao et al. 2017) demonstrate the effectiveness of our method.
- We build a proposal generation module to score all candidate proposals in a single pass and formulate a rank loss for scoring refinement.
- We devise an algorithm for top-K proposals selection that encourages both exploitation and exploration.
- We design a novel semantic completion module that predicts the important words that are masked according to the given visual context for semantic similarity estimation.

Related Work

In this section, we briefly review some related works on image/video retrieval, temporal action detection and video moment retrieval.

Image/Video Retrieval: Image/Video Retrieval aims to select image/video that is most relevant to the queries from a set of candidate images/videos. The methods in (Karpathy and Fei-Fei 2015; Escorcia et al. 2016; Xu et al. 2015; Otani et al. 2016) all propose to learn a joint visual-semantic space for cross-modal representations. In such space, the similarity of cross-modal representations reflects the closeness between their original inputs. In moment retrieval, however, we focus on retrieving a target moment in video based on the given query, rather than simply selecting a target image/video from pre-defined candidate sets.

Temporal Action Detection: Temporal Action Detection aims at identifying the temporal boundary as well as the category for each action instance in untrimmed videos. The approaches of action detection can be also summarized into supervised settings and weakly-supervised settings. These methods in (Shou, Wang, and Chang 2016; Escorcia et al. 2016; Buch et al. 2017; Shou et al. 2017; Zhao et al. 2017) are trained in two-stage supervised learning manner, which first generate temporal action proposals through a proposal network, and then predict the action category for each proposal through a classification network. In the weakly-supervised settings, however, only the coarse video-level labels instead of the exact temporal boundary is available. The UntrimmedNet in (Wang et al. 2017) make use of the principle of multiple instance learning and the generated attention weights to select proposals that most probably contain action instances. The method presented in (Nguyen et al. 2018) combines temporal class activation maps and class agnostic attentions for localizing the boundary of action instances. Further than action detection that is limited to a pre-defined set of categories, moment retrieval according to natural language query is much more challenging but general.

Video Moment Retrieval: Video Moment Retrieval is to address the target moment that is semantically aligned with the given natural language query. Prior works (Gao et al. 2017; Hendricks et al. 2017; 2018; Liu et al. 2018; Chen et al. 2018; Xu et al. 2019; Zhang et al. 2019b; 2019a; Wang, Huang, and Wang 2019) mainly focus on localizing the most relevant moment in a fully-supervised settings. Among them, methods proposed in (Gao et al. 2017; Hendricks et al. 2017; 2018) sample candidate moments by sliding windows with various length, and perform coarse fusion to estimate the correlation between the queries and moments in a multi-modal space. Further, the Temporal GroundNet (TGN) (Chen et al. 2018) proposes an interactor to exploit the evolving fine-grained frame-by-word interactions and simultaneously score a set of candidate moments in one single pass. The Cross-Modal Interaction Network (CMIN) (Zhang et al. 2019b) advises a multi-head self-attention mechanism to capture the long-range dependencies in videos and a syntactic GCN to obtain the fine-grained queries representations. The Semantic Matching Reinforcement Learning (SM-RL) (Wang, Huang, and Wang 2019) proposes a recurrent neural network based reinforcement learning model and introduce mid-level semantic concepts to bridge the semantic gap between visual and semantic information.

Though those methods achieve good performance, they still suffer from collecting a large amount of manually labelled temporal annotations. Some works (Bojanowski et al. 2015; Duan et al. 2018; Mithun, Paul, and Roy-Chowdhury

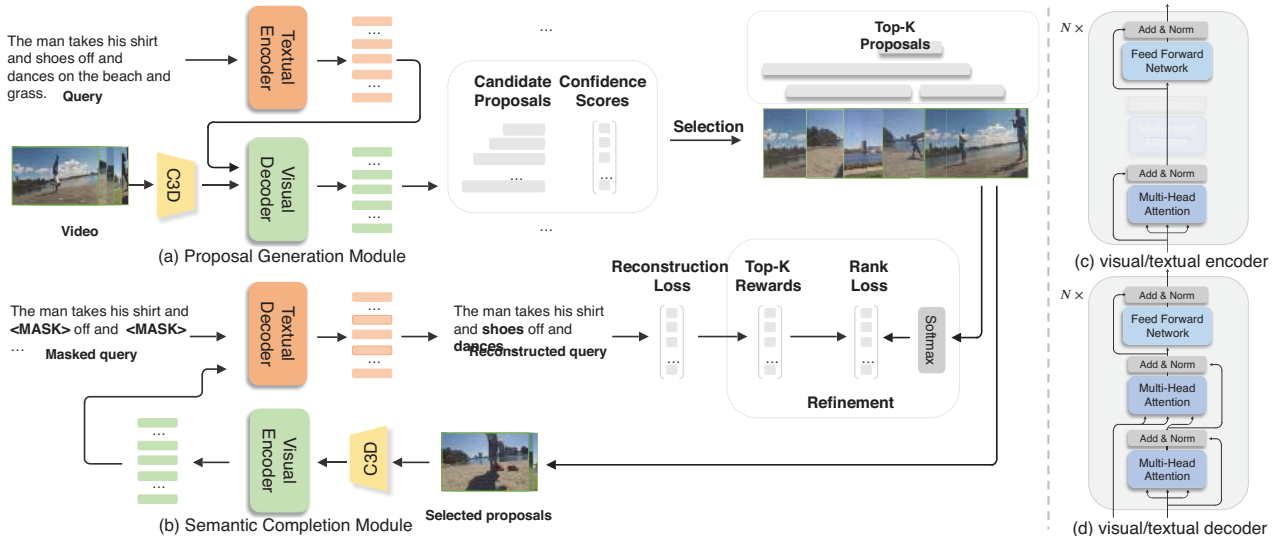


Figure 2: The Framework of our Semantic Completion Network for Video Moment Retrieval. (a) The proposal generation module leverages the cross-modal fusion representations of video and query to score all the candidate proposals at each time step, and then select the top-K proposals considering both exploitation and exploration. (b) The semantic completion module reconstruct the query in which the important words are masked according to the visual representations of proposal, compute the rewards based on the reconstruction loss, and provide feedbacks to the proposal generation module for scoring refinement.

2019) also study this task in a weakly-supervised setting. The method proposed in (Bojanowski et al. 2015) consider the task of aligning a video with a set of temporal ordered sentences, in which temporal ordering can be seen as additional constraint and supervision. The method proposed in (Duan et al. 2018) decomposes the problem of weakly-supervised dense event captioning in videos (WS-DEC) into a cycle of dual problems: caption generation and moment retrieval and explores the one-to-one correspondence between the temporal segment and event caption, and has a complex training pipeline such as pre-training and alternating training. The Text-Guided Attention (TGA) (Mithun, Paul, and Roy-Chowdhury 2019) proposes to learn a joint visual-semantic representations and utilizes the attention score as the alignment between video frames and query.

Approach

Problem Formulation

In this paper, we consider the task of Video Moment Retrieval in a weakly-supervised setting. Given an untrimmed video $\mathbf{v} = \{\mathbf{v}_i\}_{i=1}^{n_v}$ where n_v is the frame number of the video and $\mathbf{v}_i$ is the i-th feature vector, and a corresponding query $\mathbf{q} = \{\mathbf{q}_i\}_{i=1}^{n_q}$ where n_q is the word number of the query and $\mathbf{q}_i$ is the i-th feature vector, aims to localize the most relevant moment $\hat{\tau} = (\hat{s}, \hat{e})$ during inference, where $\hat{s}$, $\hat{e}$ are the indices of start frame and end frame respectively.

Proposal Generation Module

In this section, we introduce the proposal generation module. As mentioned above, the attention weights usually focus on the most discriminative but small regions, and thus fails to cover the entire temporal extent of target moment. As Figure 2(a) shows, instead, this module scores the candidate proposals according to the cross-modal representations of video and query. Moreover, further than these methods in (Gao et al. 2017; Hendricks et al. 2018) that handle different proposals separately in a sliding window fashion, our method scores all the candidate moments in a single pass, which makes full use of the context information.

In detail, the feature vector $\mathbf{q}_i$ of each word can be extracted using a pre-trained word2vec embedding. Then we develop a textual encoder $\mathbf{Enc}_q$ to obtain the textual representations for the query $\mathbf{q}$. After that, we input the textual representations and the video features $\mathbf{v}^i$ to the visual decoder $\mathbf{Dec}_v$ to obtain the final cross-modal representations $\mathbf{c} = \{\mathbf{c}_i\}_{i=1}^{n_v}$ of video and query, given by

$$\mathbf{c} = \mathbf{Dec}_v(\mathbf{v},\ \mathbf{Enc}_q(\mathbf{q})), \tag{1}$$

To generate confidence score in a single pass, we first predefine a set of candidate proposals at each time step, denoted by $C_t = \{(t - r_k * n_v,\ t)\}_{k=1}^{n_k}$, where $t - r_k * n_v$, t are the start and end boundaries of the k-th candidate proposal at the t-th time step, $r_k \in (0, 1)$ is the k-th ratio and n_k is the number of candidate proposals. Note that r_k is a fix ratio for each time step. Then based on the cross-modal representations $\mathbf{c}$, we can simultaneously give the confidence scores for these proposals at all time steps by a fully connected layer with sigmoid nonlinearity, denoted by

$$SC_t = \sigma(\mathbf{W}_s\mathbf{c}_i + \mathbf{b}_s), \tag{2}$$

where $SC_t \in \mathcal{R}^{n_k}$ represents the vector of confidence scores for the n_k candidate proposals at the t-th time step.

Given the candidate proposals $\{C_t\}_{t=1}^{n_v}$, we apply the selection algorithm that considers both exploitation and exploration to select the top-K proposals $G = \{G^k\}_{k=1}^{K}$ and give

the corresponding confidence scores $S = \{S^k\}_{k=1}^K$, where $G^k = (s_k,\ e_k)$ represents the k-th proposal in top-K proposals and S^k is its confidence score. Concretely, we rank the proposals according to their corresponding confidence scores. At each step, we choose a proposal randomly with a possibility of p or choose the proposal with the highest score with a possibility of $1-p$, and use Non Maximum Suppression (NMS) to remove those proposals that have high overlap with the chosen one. We define the sampling possibility p with a decay function dependent on the times of parameter updates n_{update}, given by

$$p = \lambda_1 * \exp(-n_{update}/\lambda_2), \tag{3}$$

where λ_1, λ_2 are the hyper-parameters to control the decay rate. As the training proceeds, the possibility of choosing next proposal randomly decreases gradually.

Semantic Completion Module

In this section, we introduce the semantic completion module to measure the semantic similarity between proposals and query, compute rewards and provide feedbacks to previous module for scoring refinement. As shown in Figure 2(b), the important words (e.g. noun, verb) are masked and predicted according to the given visual context. The most semantically matching proposal can provide enough useful information to predict the key words and also contains less noise.

First, we extract video features for the k-th proposal $G^k = (s_k,\ e_k)$, denoted by $\hat{\mathbf{v}}^k = \{\mathbf{v}_i\}_{i=s_k}^{e_k}$, and obtain the visual representations through the visual encoder $\mathbf{Enc}_v$. We denote the original words sequence as $\mathbf{w} = \{\mathbf{w}_i\}_{i=1}^{n_q}$, where $\mathbf{w}_i$ is the i-th word of the query. Then given the words sequence $\mathbf{w}$ and a set of masked position $\mathcal{X}$, we denote $\hat{\mathbf{w}}$ as a modified version of $\mathbf{w}$ where those words $\mathbf{w}_i, i \in \mathcal{X}$ are replaced by a special symbol. We can extract word features for $\hat{\mathbf{w}}$, denoted as $\hat{\mathbf{q}} = \{\hat{\mathbf{q}}_i\}_{i=1}^{n_q}$. Next, through a bi-directional textual decoder $\mathbf{Dec}_q$, we can obtain the final cross-modal semantic representations $\mathbf{f}^k = \{\mathbf{f}_i^k\}_{i=1}^{n_q}$ for the proposal G^k, given by

$$\mathbf{f}^k = \mathbf{Dec}_q(\hat{\mathbf{q}},\ \mathbf{Enc}_v(\hat{\mathbf{v}}^k)), \tag{4}$$

To predict the masked words, we can compute the energy distribution $\mathbf{e}^k = \{\mathbf{e}_i^k\}_{i=1}^{n_q}$ on the vocabulary by a fully connected layer, denoted by

$$\mathbf{e}_i^k = \mathbf{W}_v \mathbf{f}_i^k + \mathbf{b}_v, \tag{5}$$

where $\mathbf{e}_i^k \in \mathcal{R}^{n_w}$ is the energy distribution at the i-th time step, n_w is the number of words in the vocabulary.

Training of Semantic Completion Network

In this section, we describe the loss function we optimize to train the Semantic Completion Network.

Reconstruction Loss. With the energy distribution $\mathbf{e}^k$ for the proposal G^k, we first adopt a reconstruction loss to train the semantic completion module and make it able to extract key information from the visual context to predict the masked words. Formally, we then compute the negative log-likehood of each masked word and add them up, denoted by

$$\mathcal{L}_{rec}^k = -\sum_{i=1}^{n_q-1} \log p(\mathbf{w}_{i+1}|\hat{\mathbf{w}}_{1:i}, \hat{\mathbf{v}}^k) \tag{6}$$

$$= -\sum_{i=1}^{n_q-1} \log p(\mathbf{w}_{i+1}|\mathbf{e}_i^k), \tag{7}$$

where $\mathcal{L}_{rec}^k$ represents the reconstruction loss based on the visual context of the proposal G^k.

Rank Loss. As Figure 2 shows, in order to correct the confidence scores given by the proposal generation module, we further apply a rank loss to train this module. Note that we correct the confidence scores based on reward rather than one-hot label. Specifically, we define the reward R^k for the proposal G^k with a reward function to encourage proposals with lower reconstruction loss. The reward is reduced from one to zero in steps of $1\ /\ (K-1)$.

Then the strategy of policy gradient is used to correct the scores. Note that the confidence scores are normalized by a $softmax$ layer, which is an extremely important operation to highlight the semantically matching proposals and weaken the mismatched ones. The rank loss $\mathcal{L}_{ran}^k$ for the proposal G^k is computed by

$$\mathcal{L}_{ran}^k = -R^k \log\left(\frac{\exp(S^k)}{\sum_{i=1}^K \exp(S^i)}\right), \tag{8}$$

Multi-Task Loss. With the reconstruction loss and the rank loss for each proposal, we average losses over all proposals and compute a multi-task loss to train the semantic complete network in an end-to-end manner, denoted by

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}(\mathcal{L}_{rec}^k + \beta\mathcal{L}_{ran}^k), \tag{9}$$

where β is a hyper-parameter to control the balance of two losses.

Network Design

In this section, we introduce the details of the semantic completion network, including the components of visual/textual encoder and visual/textual decoder.

Encoder and Decoder. It has been indicated in (Tang et al. 2018) that Transformer (Vaswani et al. 2017) is a strong feature extractor. In this paper, we build our visual/textual encoder and visual/textual decoder based on the bi-directional Transformer, as Figure 2(c)(d) shows. The encoder/decoder is composed of a stack of layer that contains the multi-head attention sub-layer and the fully connected feed-forward network.

Parameter Sharing. We share the parameters between the visual/textual encoder and visual/textual decoder. As Figure 2(c)(d) shows, an encoder can be also regarded as a decoder without computing attention from another input of different modality. Parameter sharing greatly reduces the number of parameters and save memory. This is also a kind of model-level dual learning (Xia et al. 2018) sharing parameters across tasks, which promotes knowledge sharing.

Experiments

Datasets

We perform experiments on two public datasets for video moment retrieval to evaluate the effectiveness of our SCN method.

ActivityCaptions. The ActivityCaptions (Caba Heilbron et al. 2015) dataset is originally developed for human activity understanding. This dataset contains 20,000 various untrimmed videos and each video includes multiple natural language descriptions with temporal annotations. The released ActivityCaptions dataset comprise 17,031 description-moment pairs for training. Since the caption annotations of test data of ActivityCaptions are not publically available, we take the val_1 as the validation set and val_2 as test data. The average length of the description, also regarded as query in moment retrieval, is 13.16 words, and the average duration of the video is 117.74 seconds.

Charades-STA. The Charades-STA dataset is released in (Gao et al. 2017) for moment retrieval, and comprises 12,408 description-moment pairs for training, and 3,720 for testing. The average length of the query is 8.6 words, and the average duration of the video is 29.8 seconds. The Charades dataset, originally introduced in (Sigurdsson et al. 2016), contains only temporal activity annotation and multiple video-level descriptions for each video. The authors of (Gao et al. 2017) design a semi-automatic way to generate sentence temporal annotations. First, the video-level descriptions from the original dataset were split into sub-sentences. Then, by matching keywords for activity categories, these sub-sentences are aligned with moments in videos. The rule-based annotations are ultimately verified by humans.

Evaluation Metric

To evaluate the performance of our SCN method and baselines, we adopt the evaluation metric proposed by (Gao et al. 2017) to compute "R@n, IoU=m". Specifically, we compute the percentage of at least one of the top-n predicted moments having Intersection over Union (IoU) larger than m, denoted by $R(n,m) = \frac{1}{n_t}\sum_{i=1}^{n_t} r(n,m,q_i)$, where q_i is the i-th query, n_t is the number of testing query, $r(n,m,q_i)$ is 1 only if the top-n returned moments about q_i contains at least one that has a temporal IoU $>$ m and $R(n,m)$ is the overall performance.

Implementation Details

Data Preprocessing. For each video, we pre-extract visual frame-based features by a publicly available pre-trained 3D-ConvNet model which has a temporal resolution of 16 frames. This network was not fine-tuned on our data. We reduce the dimensionality of the activations from the second fully-connected layer (fc7) of the network from 4096 to 500 dimensions using PCA. The C3D features were extracted every 8 frames. The maximum number of frame is set to 200.

For each description, we split it into words using NLTK and extract word embeddings using the pretrained Glove (Pennington, Socher, and Manning 2014) word2vec for each word token. The maximum description length is set to 20. We also keep the most common n_w words in training set, resulting in a vocabulary size of 8,000 for ActivityCaptions and 1,111 for Charades-STA.

Model Settings. At each time step of video, we score n_k candidate proposals of multiple scales. We set n_k to 6 with ratios of [0.167, 0.333, 0.500, 0.667, 0.834, 1.0] for ActivityCaptions, and to 4 with ratios of [0.167, 0.250, 0.333, 0.500] for Charades-STA. We then set the decay hyper-parameter λ_1 to 0.5, λ_2 to 2000, the number of selected proposals K to 4, the balance hyper-parameter β to 0.1. Also, we mask one-third of words in a sentence and replace with a special token for semantic completion. Note that noun and verb are more likely to be masked. Moreover, for TransformerEncoder as well as TransformerDecoder, the dimension of hidden state is set to 256 and the number of layers is set to 3. During training, we adopt the Adam optimizer with learning rate 0.0002 to minimize the multi-task loss. The learning rate increases linearly to the maximum with a warm-up step of 400 and then decreases itself based on the number of updates (Vaswani et al. 2017).

Compared Methods

Random. We simply select a candidate moment randomly.

VSA-RNN and VSA-STV. (Gao et al. 2017) This two methods both simply project the visual feature of all candidate proposals and the textual feature of the query into a common space, and computes the confidence scores based on cosine similarity.

CTRL. (Gao et al. 2017) The CTRL method introduces a cross-modal temporal localizer to estimate the alignment scores and uses clip location regression to further adjust the the boundary.

QSPN. (Xu et al. 2019) The QSPN method devises a multilevel approach for integrating vision and language features using attention mechanisms, and also leverages video captioning as an auxiliary task.

WS-DEC. (Duan et al. 2018) The WS-DEC method decomposes the problem of weakly-supervised dense event captioning in videos into a cycle of dual problems: caption generation and moment retrieval, and explores the one-to-one correspondence between the temporal segment and event caption.

TGA. (Mithun, Paul, and Roy-Chowdhury 2019) The TGA method proposes a weakly-supervised joint visual-semantic embedding framework for moment retrieval, and utilizes the latent alignment for localization during inference.

Quantitative Results and Analysis

The overall performance results of our SCN and baselines on ActivityCaptions and Charades-STA datasets are presented in Table 1 and Table 2 respectively. We consider the evaluation metric "R@n, IoU=m", where $n \in \{1,5\}, m \in \{0.1,0.3,0.5\}$ for ActivityCaptions, and $n \in \{1,5\}, m \in \{0.3,0.5,0.7\}$ for Charades-STA. By observing the evaluation results, we can discover some facts:

- Compared with Random method, the overall performance results of SCN have a huge improvements on both two

Table 1: Performance Evaluation Results on the ActivityCaptions Dataset ($n \in \{1, 5\}$ and $m \in \{0.1, 0.3, 0.5\}$).

Method	R@1			R@5		
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
Random	38.23	18.64	7.63	75.74	52.78	29.49
VSA-RNN	-	39.28	23.43	-	70.84	55.52
VSA-STV	-	41.71	24.01	-	71.05	56.62
CTRL	-	47.43	29.01	-	75.32	59.17
QSPN	-	52.12	33.26	-	77.72	62.39
WS-DEC	62.71	41.98	23.34	-	-	-
SCN	**71.48**	**47.23**	**29.22**	**90.88**	**71.45**	**55.69**

Table 2: Performance Evaluation Results on the Charades-STA Dataset ($n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$).

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
Random	20.12	8.61	3.39	68.42	37.57	14.98
VSA-RNN	-	10.50	4.32	-	48.43	20.21
VSA-STV	-	16.91	5.81	-	53.89	23.58
CTRL	-	23.63	8.89	-	58.92	29.52
QSPN	54.70	35.60	15.80	95.60	79.40	45.40
TGA	32.14	19.94	8.84	86.58	65.52	33.51
SCN	**42.96**	**23.58**	**9.97**	**95.56**	**71.80**	**38.87**

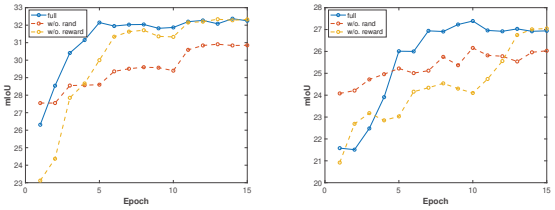
datasets, which demonstrates that optimizing the multi-task loss instead of explicitly optimizing the localization loss can reach the goal of predicting the target moment and also indicates the feasibility of our SCN method.

- As the results show, the proposed SCN method outperforms the supervised visual-embedding approaches VSA-RNN and VSA-STV significantly, and obtains results comparable to the other fully-supervised methods on two datasets, indicating that even without the full annotations of temporal boundary, our SCN method can still effectively exploit the alignment relationship between video and query and find the most semantically relevant moment.
- The coarse methods VSA-RNN and VSA-STV achieve the worst performance on two datasets, even compared with the weakly-supervised SCN method, demonstrating the key role of visual and textual modeling in moment retrieval and indicating the limitation of learning a common visual-semantic space in high-quality retrieval.
- Also, compared with the weakly-supervised methods WS-DEC and TGA, our method achieves tremendous improvements on both ActivityCaptions and Charades-STA datasets. These results verify the effectiveness of the proposal generation module, the semantic completion module, the algorithm of proposals selection and the multi-task loss.

Ablation Study

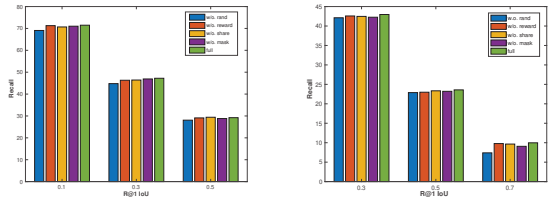
To prove the validity of different parts of our method, we simplify the algorithm to generate different ablation models as follows:

- **SCN(w/o. rand).** During proposals selection, we assign the sample possibility p to zero, which means we select next proposal completely based on the confidence scores without random selection at each step.



(a) ActivityCaptions (b) Charades-STA

Figure 3: Training Process of Different Models on ActivityCaptions and Charades-STA Datasets



(a) ActivityCaptions (b) Charades-STA

Figure 4: Evaluation Results of Different Models on ActivityCaptions and Charades-STA Datasets

- **SCN(w/o. reward).** With feedbacks given by the semantic completion module, we modify the rank loss by using one-hot label instead of computing rewards for scoring refinement. Concretely, we simply assign a reward of one to the best proposal, and zero to the other ones. The rank loss is equivalent to the cross entropy loss.
- **SCN(w/o. mask).** To validate the effectiveness of the semantic completion module and the reconstruction loss, we replace this module with a ordinary captioning generator (Duan et al. 2018) without masking words.
- **SCN(w/o. share).** Instead of parameter sharing between the proposal generation module and the semantic completion module, we use two separate sets of parameters for this two modules.

The training process of different models on ActivityCaptions and Charades-STA is presented in Figure 3. By analyzing the results, we can find some interesting points:

- The simplified models SCN(w/o. rand) and SCN(w/o. reward) still achieve results comparable to the fully-supervised methods and outperform the existing weakly-supervised methods, which further demonstrates the effectiveness of our framework including proposal generation and selection, semantic completion for semantic similarity estimation and scoring refinement.
- The SCN(full) achieves better results than the SCN(w/o. rand) and the evaluation results of SCN(w/o. rand) grows more gently as the model converges, which proves the ability of the random selection to find potentially good proposals during proposals selection. When the model has

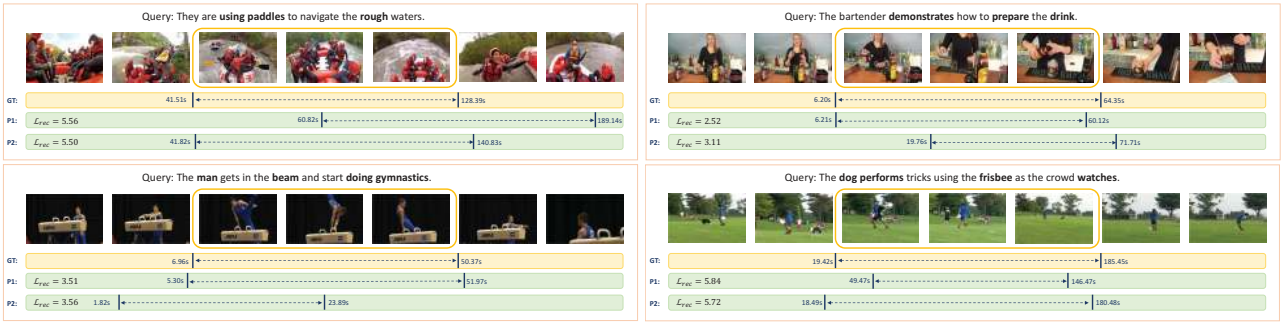


Figure 5: Qualitative Examples on the ActivityCaptions dataset

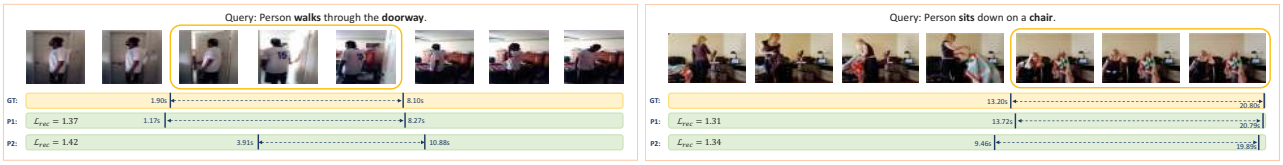


Figure 6: Qualitative Examples on the Charades-STA dataset

not converged, selecting proposals randomly provide opportunities for those potentially good proposals and speed up training.

- The SCN(full) achieves the best results faster than SCN(w/o. reward), which demonstrates the effectiveness of employing rewards as feedbacks to train the proposal generation module. In the early of training stage, the semantic module can't provide accurate feedbacks but the one-hot label force the previous module to accept only one proposal and reject other proposals that are actually reasonable.
- As shown in 4, the SCN(full) achieves better results than the SCN(w/o. mask), which indicating the effectiveness of the semantic completion module and the masking operation. By masking the important words of the query, we forces the decoder to absorb the cross-modal visual information in the decoder side.
- The SCN(full) also perform a bit better than the SCN(w/o. share), demonstrating the effectiveness of parameter sharing to promote knowledge sharing. Also, The amount of parameters is greatly reduced by parameter sharing.

Qualitative Results

To qualitatively validate the performance of our SCN method, several examples of video moment retrieval from ActivityCaptions and Charades-STA are provided in Figure 5 and Figure 6 respectively. Each example provide the ground truth of temporal boundaries, the first two proposals with the highest confidence score given by the proposal generation module. The bold words in the sentence are considered as the important words associated with the video context, and are masked for semantic completion. The corresponding reconstruction loss $\mathcal{L}_{rec}$ is also computed and presented in each example.

It can be observed that both the first two proposals with the highest confidence score cover the most discriminative video contents relevant to the query, which qualitatively verify that the proposal generation module can locate those semantically important proposals, and the rank loss is helpful for scoring refinement during training. Additionally, the proposal with higher IoU has lower reconstruction loss, also indicating the proposal that is more semantically matching with the query can be recognized by the semantic completion module. Therefore, due to the effectiveness of two submodules and the training algorithm, our method is successful in localizing the moment that has high IoU with the target moment.

Conclusion

In this paper, we study the task of video moment retrieval from the perspective of weak-supervised learning without manually-labelled temporal boundaries of start time and end time, which makes this task more realistic but more challenging. We propose a novel semantic completion network (SCN) including the proposal generation module to score all candidate proposals in a single pass, an efficient algorithm for proposals selection considering both exploitation and exploration, the semantic completion module for semantic similarity estimation and a multi-task loss for training. The experiments on the ActivityCaptions and Charades-STA datasets also demonstrate the effectiveness of our method to exploit the alignment relationship between video and query, and the efficiency of the proposal selection algorithm and the rank loss.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.61602405, No.U1611461, No.61751209 and No.61836002, China

Knowledge Centre for Engineering Sciences and Technology, and Alibaba Innovative Research.

References

Bojanowski, P.; Lajugie, R.; Grave, E.; Bach, F.; Laptev, I.; Ponce, J.; and Schmid, C. 2015. Weakly-supervised alignment of video with text. In *IEEE CVPR*, 4462–4470.

Buch, S.; Escorcia, V.; Shen, C.; Ghanem, B.; and Niebles, J. C. 2017. Sst: Single-stream temporal action proposals. In *IEEE CVPR*, 6373–6382.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE CVPR*, 961–970.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *EMNLP*, 162–171.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. In *NIPS*, 3059–3069.

Escorcia, V.; Heilbron, F. C.; Niebles, J. C.; and Ghanem, B. 2016. Daps: Deep action proposals for action understanding. In *ECCV*, 768–784. Springer.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *IEEE CVPR*, 5267–5275.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *IEEE ICCV*, 5803–5812.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2018. Localizing moments in video with temporal language. In *EMNLP*, 1380–1390. ACL.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE CVPR*, 3128–3137.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *MM*, 843–851. ACM.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *IEEE CVPR*, 11592–11601.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *IEEE CVPR*, 6752–6761.

Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; and Yokoya, N. 2016. Learning joint representations of videos and sentences with web image search. In *ECCV*, 651–667. Springer.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Shou, Z.; Chan, J.; Zareian, A.; Miyazawa, K.; and Chang, S.-F. 2017. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *IEEE CVPR*, 1417–1426.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *IEEE CVPR*, 1049–1058.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding.

Singh, K. K., and Lee, Y. J. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE ICCV*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Tang, G.; Müller, M.; Rios, A.; and Sennrich, R. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In *EMNLP*, 4263–4272.

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 *NIPS*, 5998–6008.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. Untrimmednets for weakly supervised action recognition and detection. In *IEEE CVPR*.

Wang, W.; Huang, Y.; and Wang, L. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *IEEE CVPR*, 334–343.

Wang, C.; Li, M.; and Smola, A. J. 2019. Language models with transformers. *CoRR* abs/1904.09408.

Xia, Y.; Tan, X.; Tian, F.; Qin, T.; Yu, N.; and Liu, T.-Y. 2018. Model-level dual learning. In *ICML*, 5383–5392.

Xu, R.; Xiong, C.; Chen, W.; and Corso, J. J. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 5, 6.

Xu, H.; He, K.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 2, 7.

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 *IEEE CVPR*, 1247–1257.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *ACM SIGIR*, 655–664.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *IEEE ICCV*.