Weakly Supervised Video Moment Retrieval From Text Queries

Niluthpol Chowdhury Mithun*, Sujoy Paul*, Amit K. Roy-Chowdhury

Electrical and Computer Engineering, University of California, Riverside

{nmithun, supaul, amitrc}@ece.ucr.edu

Abstract

There have been a few recent methods proposed in text to video moment retrieval using natural language queries, but requiring full supervision during training. However, acquiring a large number of training videos with temporal boundary annotations for each text description is extremely time-consuming and often not scalable. In order to cope with this issue, in this work, we introduce the problem of learning from weak labels for the task of text to video moment retrieval. The weak nature of the supervision is because, during training, we only have access to the video-text pairs rather than the temporal extent of the video to which different text descriptions relate. We propose a joint visual-semantic embedding based framework that learns the notion of relevant segments from video using only video-level sentence descriptions. Specifically, our main idea is to utilize latent alignment between video frames and sentence descriptions using Text-Guided Attention (TGA). TGA is then used during the test phase to retrieve relevant moments. Experiments on two benchmark datasets demonstrate that our method achieves comparable performance to state-of-the-art fully supervised approaches.

1. Introduction

Cross-modal retrieval of visual data using natural language description has attracted intense attention in recent years [10, 43, 15, 12, 40, 41, 24], but remains a very challenging problem [43, 6, 21] due to the differences and ambiguity between different modalities. The identification of the video moment (or segment) is important since it allows the user to focus on the portion of the video that is most relevant to the textual query, and is beneficial when the video has a lot of non-relevant portions. (See Fig. 1). The aforementioned approaches operate in a fully supervised setting, i.e., they have access to text descriptions along with the exact temporal location of the visual data corresponding to the descriptions. However, obtaining such annotations is te-

Figure 1. Illustration of text to video moment retrieval task: given a text query, retrieve and rank videos segments based on how well they depict the text description.

dious and noisy, requiring multiple annotators. The process of developing algorithms which demand a weaker degree of supervision is non-trivial and is yet to be explored by researchers for the problem of video moment retrieval using text queries. In this work, we focus, particularly on this problem.

The text to video moment retrieval task is more challenging than the task of localizing categorical activities in videos, which is a comparatively well-studied field [20, 38, 44, 39, 28, 31]. Although these methods show success on activity localization, unlike text to moment retrieval, they are limited to a pre-defined set of activity classes. In this regard, there has been a recent interest in localizing moments in a video from natural language description [9, 8, 40, 3]. Supervision in terms of text description with their temporal boundaries in a video is used to train these models. However, acquiring such dense annotations of text-temporal boundary tuples are often tedious, as it is difficult to mark the start and end locations of a certain moment, which introduces ambiguity in the training data.

On the contrary, it is often much easier to just describe the moments appearing in a video with a set of natural language sentences, than providing exact temporal boundaries associated with each of the sentences. Moreover, such descriptions can often be obtained easily from captions through some sources on the web. Motivated by this, we pose a question in this paper: *is it possible to develop*
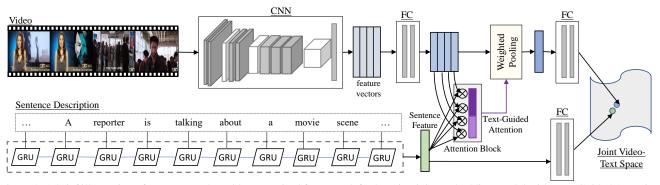
---

*Joint first author

1

Figure 2. A brief illustration of our proposed weakly supervised framework for learning joint embedding model with Text-Guided Attention for text to video moment retrieval. Our framework learns a latent alignment between relevant video frames and text corresponding to the video. This alignment is utilized for attending video features based on relevance and the pooled video feature is used for learning the joint video-text embedding. In the figure, CNN refers to a convolutional neural network, FC refers to a fully-connected neural network, and GRU refers to gated recurrent units. Please see Sec. 3 for details of our approach.

a weakly-supervised framework for video moment localization from the text, leveraging only video-level textual annotation, without their temporal boundaries? Temporal localization of moments using weak description is a much more challenging task than its supervised counterpart. It is extremely relevant to address this question, due to the difficulty and non-scalability of acquiring precise frame-wise information with text descriptions in the fully supervised setting, which require enormous manual labor.

Overview of the Proposed Framework. An illustration of our proposed weakly-supervised framework presented in Fig. 2. Given a video, we first extract frame-wise visual features from pre-trained Convolutional Neural Network (CNN) architectures. We also extract features for text descriptions using Recurrent Neural Network (RNN) based models. Similar to several cross-modal video-text retrieval models [5, 15], we train a joint embedding network to project video features and text features into the same joint space. However, as we have text descriptions for the videos as a whole and not moment-wise descriptions like in a fully supervised setting, the learning procedure for text to video moment retrieval is non-trivial.

Given a certain text description, we obtain its similarity with the video features, which gives an indication of temporal locations which may correspond to the textual description. We call this Text-Guided Attention as it helps to highlight the relevant temporal locations, given a text description. Thereafter, we use this attention to pool the video features along the temporal direction to obtain a single text-dependent feature vector for a video. We then train the network to minimize a loss which reduces the distance between the text-dependent video feature vector and the text vector itself. We hypothesize that along with learning a shared video-text embedding, hidden units will emerge internally to learn the notion of relevance between moments of video and corresponding text description. During the

testing phase, we use TGA for localizing the moments, given a text query, as it highlights the portion of the video corresponding to the query.

Contributions: The main contributions of the proposed approach are as follows.

• We address a novel and practical problem of temporally localizing video moments from text queries without requiring temporal boundary annotations of the text descriptions while training but using only the video-level text descriptions.

• We propose a joint visual-semantic embedding framework, that learns the notion of relevant moments from video using only video-level description. Our joint embedding network utilizes latent alignment between video frames and sentence description as Text-Guided Attention for the videos to learn the embedding.

• Experiments on two benchmark datasets: DiDeMo [9] and Charades-STA [8] show that our weakly-supervised approach performs reasonably well compared to supervised baselines in the task of text to video moment retrieval.

2. Related Works

Image/Video Retrieval using Text Queries. Cross-modal language-vision retrieval methods focus on retrieving relevant images/videos from a database given text descriptions. Most of the recent methods for image-text retrieval task focus on learning joint visual-semantic embedding models [13, 15, 7, 36, 6, 24, 34, 23]. Inspired by the success of these approaches, most video-text retrieval methods also employ a joint subspace model [42, 5, 35, 26, 21, 22]. In this joint space, the similarity of different points reflects the semantic closeness between their corresponding original inputs. These text-based video retrieval approaches focus on retrieving an entire video from dataset given text description. However, we focus on temporally localizing a specific moment relevant to a text query, within a

given video. Similar to the video/image to text retrieval approaches, our proposed framework is also based on learning joint video-text embedding models. However, instead of focusing only on aligning video and text in the joint space as in video-text retrieval, our aim is to learn a latent alignment between video frames and text descriptions, which is used for obtaining the relevant moments corresponding to a given text query.

Activity Localization. The moment retrieval aspect of our work is related to the problem of temporal activity localization in untrimmed videos. From the perspective of our interest, the works in literature pertaining to activity localization can be categorized as either fully supervised or weakly supervised. Works in fully supervised setting include SSN [44], R-C3D [39], TAL-Net [2] among others. Most of these works structure their framework by using temporal action proposals with activity location predictors. However, in the weakly supervised setting, the exact location of each activity is unknown, and only the video-level labels are accessible during training. In order to deal with that, researchers take a Multiple Instance Learning approach [38] with constraints applied for better localization [28, 25]. Our task of video moment retrieval from text description is more challenging than the activity localization task, as our method is not limited to a pre-defined set of categories, but rather sentences in natural language.

Text to Video Moment Retrieval. Most relevant to our work are the methods that focus on identifying relevant portions from text description using fully-supervised annotations: MCN [9], CTRL [8], EFRC [40], ROLE [19], TGN [3]. These methods are severely plagued by the issue of collecting training videos with temporal natural language annotation. Temporal sliding window over videos frames [9], or hard-coded segments containing a fixed number of frames [8] has been used for generating moment candidate corresponding to a text description. Moreover, unlike in images, generating temporal proposals for videos in an unsupervised manner is itself a challenging task. In [40, 39], the authors proposed an end-to-end framework where the activity proposals are generated as one of the initial steps, but for the much easier task of activity localization. Attention mechanism has been used in [19, 40] for the text to video moment retrieval task. Although we also use attention, our usage is significantly different from them. ROLE [19] uses attention over the words using video moment context, which they obtain from the temporal labels. EFRC [40] uses attention in training a temporal proposal network as it has access to temporal boundary annotations of the sentences. We use attention over the temporal dimension of the videos as we do not have access to the temporal boundaries. More importantly, our method is weakly-supervised, which requires only video-level text annotation during training. Hence, the data collection cost for our approach is substan-

tially less, and it is possible to acquire and train using larger video-text captioning datasets.

A weakly supervised setting is considered in [1] for the video-text alignment task, which is to assign temporal boundaries to a set of temporally ordered sentences, whereas our task is to retrieve a portion of the video given a sentence. Moreover, [1] assumes temporal ordering between the sentences as additional supervision. Also, their method would require dense sentence annotations describing all portions of the video including tokens representing background moments (if any). The task considered in this work is a generalization of the task in [1]. We consider that there can be multiple sentences describing different temporal portions of a single video and do not consider any temporal ordering information of the sentences. The Text-Guided Attention mechanism used in our framework allows us to deal with multiple sentence descriptions during training and provide the relevant portions for each of them during testing, even with weak supervision.

3. Approach

In this section, we first describe the network architecture and input feature representation for representing video and text (Sec. 3.1). Then, we present our proposed Text-Guided Attention module (Sec. 3.2). Finally, we describe the framework for learning joint video-text embedding (Sec.3.3).

Problem Definition. In this paper, we consider that the training set consists of videos paired with text descriptions composed of multiple sentences. Each sentence describes different temporal regions of the video. However, we do not have access to the temporal boundaries of the moments referred to by the sentences. At test time, we use a sentence to retrieve relevant portions of the video.

3.1. Network Structure and Features

Network Structure. The joint embedding model is trained using a two-branch deep neural network model, as shown in Fig. 2. The two branches consist of different expert neural networks to extract modality-specific representations from the given input. The expert networks are followed by fully connected embedding layers which focus on transforming the modality-specific representations to joint representations. In this work, we keep the pre-trained image encoder fixed as we have limited training data. The fully-connected embedding layers, the word embedding, the GRU are trained end-to-end. We set the dimensionality (D) of the joint embedding space to 1024.

Text Representation. We use Gated Recurrent Units (GRU) [4] for encoding the sentences. GRU has been very popular for generating a representation for sentences in recent works [6, 15]. The word embeddings are input to the GRU. The dimensionality of the word embeddings is 300.

Video Representation. We utilize pre-trained convolutional neural network models as the expert network for encoding videos. Specifically, following [8] we utilize C3D model [33] for feature extraction from every 16 frames of video for the Charades-STA dataset. A 16 layer VGG model [30] is used for frame-level feature extraction in experiments on DiDeMo dataset following [9]. We extract features from the penultimate fully connected layer. For both the C3D and VGG16 model, the dimension of the representation from the penultimate fully connected layer is 4096.

3.2. Text-Guided Attention

After the feature extraction process, we have a training set $\mathcal{D} = \{\{\boldsymbol{w}_j^i\}_{j=1}^{nw_i}, \{\boldsymbol{v}_k^i\}_{k=1}^{nv_i}\}_{i=1}^{n_d}$, where n_d is the number of training pairs, $\boldsymbol{w}_j^i$ represents the jth sentence feature of ith video, $\boldsymbol{v}_k^i$ represent the video feature at the kth time instant of the ith video, nw_i and nv_i are the number of sentences in the text description and video time instants for the ith video in the dataset. Please note that we do not consider any ordering in the text descriptions.

Each of the sentences provides us information about a certain part of the given video. In a fully supervised setting, where we have access to the temporal boundaries associated with each sentence, we can apply a pooling technique to first pool the relevant portion of the video features and then use a similarity measure to learn a joint video segment-text embedding. However, in our case of weakly supervised moment retrieval, we do not have access to the temporal boundaries associated with the sentences. Thus, we need to first obtain the portions of the video which are relevant to a given sentence query.

If some portion of the video frames corresponds to a particular sentence, we would expect them to have similar features. Thus, the cosine similarity between text and video features should be higher in the temporally relevant portions and low in the irrelevant ones. Moreover, as the sentence described a part of the video rather than individual temporal segments, the video feature obtained after pooling the relevant portions should be very similar to the sentence description feature. We employ this idea to learn the joint video-text embedding via an attention mechanism based on the sentence descriptions, which we name Text-Guided Attention (TGA). Note that during the test phase, we use TGA to obtain the localization.

We first apply a Fully Connected (FC) layer with ReLU [18] and Dropout [32] on the video features at each time instance to transform them into the same dimensional space as the text features. We denote these features as $\bar{\boldsymbol{v}}_k^i$. In order to obtain the sentence specific attention over the temporal dimension, we first obtain the cosine similarity between each temporal feature and sentence descriptions. The similarity between the jth sentence and the kth temporal feature of

the ith training video can be represented as follows,

$$s_{kj}^i = \frac{\boldsymbol{w}_j^{i T} \boldsymbol{v}_k^i}{||\boldsymbol{w}_j^i||_2 ||\boldsymbol{v}_k^i||_2} \tag{1}$$

Once we obtain the similarity values for all the temporal locations, we apply a softmax operation along the temporal dimension to obtain an attention vector for the ith video as follows,

$$a_{kj}^i = \frac{\exp(s_{kj}^i)}{\sum_{k=1}^{nv_i} \exp(s_{kj}^i)} \tag{2}$$

These should have high values at temporal locations which are relevant to the given sentence vector $\boldsymbol{w}_j^i$. We consider this as local similarity because the individual temporal features may correspond to different aspects of a sentence and thus each of the temporal features might be a bit scattered away from the sentence feature. However, the feature obtained after pooling the video temporal features corresponding to the relevant locations should be quite similar to the entire sentence feature. We consider this global similarity. We use the attention in Eqn. 2 to obtain the pooled video feature for the sentence description $\boldsymbol{w}_j^i$ as follows,

$$\boldsymbol{f}_j^i = \sum_{k=1}^{nv_i} a_{kj}^i \boldsymbol{v}_k^i \tag{3}$$

Note that, this feature vector corresponds to the particular sentence description $\boldsymbol{w}_j^i$ only. In a similar procedure, we can extract the text-specific video feature vector corresponding to the other sentences in the text descriptions of the same video and other videos as well. Fig. 3 presents an overview of the sentence-wise video feature extraction procedure using the video temporal features and a set of sentence descriptions for the video. We use these feature vectors to derive the loss function to be optimized to learn the parameters of the network. This is described next.

3.3. Training Joint Embedding

We now describe the loss function we optimize to learn the joint video-text embedding. Many prior approaches have utilized pairwise ranking loss as the objective for learning joint embedding between visual and textual input [15, 45, 37, 13]. Specifically, these approaches minimize a hinge-based triplet ranking loss in order to maximize the similarity between an image embedding and corresponding text embedding and minimize similarity to all other non-matching ones.

For the sake of notational simplicity, we drop the index i, j, k denoting the video number, sentence index and time instant. Given a text-specific video feature vector based on TGA, $\boldsymbol{f}$ ($\in \mathbb{R}^V$) and paired text feature vector $\boldsymbol{w}$ ($\in \mathbb{R}^T$), the projection for the video feature on the joint space can
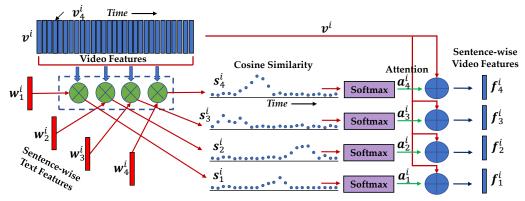
Figure 3. This figure presents the procedure of computing the Text-Guided Attention and using it to generate sentence-wise video features. We first obtain the cosine similarity between the features at every time instant of the video v_i, and its corresponding sentences w_j^i, followed by a softmax layer along the temporal dimension to obtain the sentence-wise temporal attention. Thereafter, we use these attentions to compute a weighted average of the video features to finally obtain the sentence-wise video features.

be derived as $v_p = W^{(v)}f$ ($v_p \in \mathbb{R}^D$). Similarly, the projection of paired text vector in the embedding space can be expressed as $t_p = W^{(t)}w(t_p \in \mathbb{R}^D)$. Here, $W^{(v)} \in \mathbb{R}^{D \times V}$ is the transformation matrix that projects the video content into the joint embedding and D is the dimensionality of the joint space. Similarly, $W^{(t)} \in \mathbb{R}^{D \times T}$ maps input sentence/caption embedding to the joint space.

Using these pairs of feature representation of both videos and corresponding sentence, the goal is to learn a joint embedding such that the positive pairs are closer than the negative pairs in the feature space. Now, the video-text loss function $\mathcal{L}_{VT}$ can be expressed as follows,

$$\mathcal{L}_{VT} = \sum_{(v_p, t_p)} \left\{ \sum_{t_p^-} max\big[0, \Delta - S(v_p, t_p) + S(v_p, t_p^-)\big] \right.$$
$$\left. + \sum_{v_p^-} max\big[0, \Delta - S(t_p, v_p) + S(t_p, v_p^-)\big] \right\} \quad (4)$$

where t_p^- is a non-matching text embedding for video embedding v_p, and t_p is the matching text embedding. This is similar for video embedding v_p and non-matching image embedding v_p^-. Δ is the margin value for the ranking loss. The scoring function $S(v_p, t_p)$ measures the similarity between the image embedding and text embedding in the joint space. We utilize cosine similarity in the representation space to compute similarity. Cosine similarity is widely used in learning joint embedding models in prior works on image-text retrieval [45, 15, 6, 23]. Our approach does not depend on any specific choice of similarity function.

In Eq. (4), the first term attempts to ensure that for each visual input, the matching text inputs should be closer than non-matching text inputs in learning the joint space. However, the second term in Eq. (4) attempts to ensure that for each text input, the matching image input should be closer in the joint space than the non-matching image inputs.

3.4. Batch-wise Training

We train our network using Stochastic Gradient Descent (SGD) by dividing the dataset into batches. For a video with multiple sentences, we create multiple video-sentence pairs, with the same video, but different sentences in the corresponding video's text description. During training, our method learns to automatically identify the relevant portions for each sentence using the Text-Guided Attention. The negative instances v_p^- and t_p^- correspond to all the instances which are not positive in the current batch of data.

4. Experiments

We perform experiments on two benchmark datasets with the goal of comparing the performance of our weakly-supervised approach against different supervised baselines. As we introduce the problem in this paper, to the best of our knowledge, our work is the first to show results on this task. Ideally, any weakly supervised methods would attempt at attaining the performance of the supervised methods, with similar features and setting.

We first describe the details on the datasets and evaluation metric in Sec. 4.1, followed by the training details in Sec. 4.2. Then, we report the results of different methods on DiDeMo and Charades-STA dataset in Sec. 4.3.

4.1. Datasets and Evaluation Metric

We present experiments on two benchmark datasets for sentence description based video moment localization, namely Charades-STA [8] and DiDeMo [9] to evaluate the performance of our proposed framework.

Charades-STA. The Charades-STA dataset for text to video moment retrieval was introduced in [8]. The dataset contains 16,128 sentence-moment pairs with 12,408 in the training set and 3,720 in the testing set. The Charades dataset was originally introduced in [29] which contains

5

Table 1. This table presents the results on the Charades-STA dataset, using the evaluation protocol in [8]. Following [40, 8], we also use C3D feature for a fair comparison. The proposed **weakly-supervised approach** performs significantly better that visual-semantic embedding based baselines: VSA-RNN and VSA-STV. The proposed approach also performs reasonably comparable to state-of-the-art approaches CTRL[8] and EFRC [40], and even achieves a similar performance in some evaluation metrics.

Method	IoU=0.3			IoU=0.5			IoU=0.7		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Random	-	-	-	8.51	37.12	-	3.03	14.06	-
VSA-RNN	-	-	-	10.50	48.43	-	4.32	20.21	-
VSA-STV	-	-	-	16.91	53.89	-	5.81	23.58	-
CTRL	-	-	-	23.63	58.92	-	8.89	29.52	-
EFRC	53.00	94.60	98.50	33.80	77.30	91.60	15.00	43.90	60.90
Proposed	**29.68**	**83.87**	**98.41**	**17.04**	**58.17**	**83.44**	**6.93**	**26.80**	**44.06**

temporal activity annotation and video-level paragraph description for the videos. The authors of [8] enhanced the dataset [29] for evaluating temporal localization of moments in videos given text queries. The video-level descriptions from the original dataset were decomposed into short sentences. Then, these sentences are assigned to segments in videos based on matching keywords for activity categories. The annotations are manually verified at last.

DiDeMo. The Distinct Describable Moments (DiDeMo) dataset [9] is one of the largest and most diverse datasets for the temporal localization of events in videos given natural language descriptions. The videos are collected from Flickr and each video is trimmed to a maximum of 30 seconds. The videos in the dataset are divided into 5-second segments to reduce the complexity of annotation. The dataset is split into training, validation and test sets containing 8,395, 1,065 and 1,004 videos respectively. The dataset contains a total of 26,892 moments and one moment could be associated with descriptions from multiple annotators. The descriptions in DiDeMo dataset are detailed and contain camera movement, temporal transition indicators, and activities. Moreover, the descriptions in DiDeMo are verified so that each description refers to a single moment.

Evaluation Metric. We use the evaluation criteria following prior works in literature [9, 8]. Specifically, we follow [9] for evaluating DiDeMo dataset and [8] for evaluating Charades-STA. We measure rank-based performance $R@K$ (Recall at K) which calculates the percentage of test samples for which the correct result is found in the top-K retrievals to the query sample. We report results for R@1, R@5, and R@10. We also calculate temporal intersection over union (tIoU) for Charades-STA dataset and mean intersection over union (mIoU) for DiDeMo dataset.

4.2. Implementation Details

We used two Telsa K80 GPUs and implemented the network using PyTorch [27]. We start training with a learning rate of 0.001 and keep the learning rate fixed for 15 epochs. The learning rate is lowered by a factor of 10 every 15 epochs. We tried different values for margin α in train-

ing and found $0.1 \leq \Delta \leq 0.2$ works reasonably well. We empirically choose Δ as 0.1 for Charades-STA and 0.2 for DiDeMo in the experiments. We use a batch-size of 128 in all the experiments. ADAM optimizer was used in training the joint embedding networks [14]. The model was evaluated on the validation set on the video-text retrieval task after every epoch. To deal with the over-fitting issue, we choose the best model based on the highest sum of recalls.

4.3. Quantitative Results

We report the experimental results on Charades-STA dataset [8] in Table 1 and DiDeMo dataset [9] in Table 2.

4.3.1 Charades-STA Dataset

The quantitative results on Charades-STA dataset [8] are reported in Table 1. The evaluation setup in Charades-STA dataset [8] considers a set of IoU (Intersection over Union) thresholds. We report for IoU $0.3, 0.5$ and 0.7 in Table 1. For these IoU thresholds, we report the recalls - R@1, R@5, and R@10 in Table 1. Following [8], we use sliding windows of 128 and 256 to obtain the possible temporal segments. The segments are ranked based on the corresponding Text-Guided Attention score.

Compared Methods. We compare our approach with state-of-the-art text to video moment retrieval approaches, CTRL[8], EFRC[40], and baseline approaches, VSA-RNN[12] and VSA-STV[16]. For these methods, we directly cite performances from respective papers when available [8, 40]. We report score for VSA-RNN and VSA-STV from [8]. If the score for multiple models is reported, we select the score of the best performing method in R@1. Here, VSA-RNN (Visual-Semantic Embedding with LSTM) and VSA-STV (Visual-Semantic Embedding with Skip-thought vector) are text-based image/video retrieval baselines. We also report results for "Random" which selects a candidate moment randomly. Similar to these approaches, we also utilize the C3D model for obtaining feature representation of videos for fair comparison. We follow the evaluation criteria utilized in [8, 40].

Analysis of Results. We observe that the proposed approach consistently perform comparably to several fully-supervised approaches in all evaluation metrics. Our weakly-supervised TGA based approach performs significantly better than supervised visual-semantic embedding based approaches VSA-RNN and VSA-STV. We observe that the proposed method achieves a minimum absolute improvement of 6.6% in R@5 and 2.6% in R@1 from VSA-RNN. The maximum relative performance improvement over VSA-STV is 19.3% in R@1 and 13.7% in R@5. We also observe that the proposed approach achieves comparable performance to state-of-the-art method CTRL [8] on R@5 evaluation metrics. The proposed approach also shows reasonable performance compared to the EFRC approach [40].

4.3.2 DiDeMo Datset

Table 2 summarizes the results on the DiDeMo dataset [9]. DiDeMo only has a coarse annotation of moments. As the videos are trimmed at 30 seconds and the videos are divided into 5-second segments, each video has 21 possible moments. We follow the evaluation setup in [9], which is designed for evaluating 21 possible moments from sentence descriptions. Average of Text-Guided Attention scores of corresponding segments is used as the confidence score for the moments and used for ranking. Following previous works [9, 40], the performance in the dataset is evaluated based on R@1, R@5, and mean intersection over union (mIoU) criteria.

Compared Methods. In Table 2, we report results for several baselines to analyze the performance of our proposed approach. We divide the table into 3 rows (2.1-2.3). In row-2.1, we report the results of trivial baselines (i.e., Random and Upper-Bound) following evaluation protocol reported in [9]. In row-2.2, we group the results of LSTM-RGB-Local [9], EFRC [40], and our proposed approach for a fair comparison, as these methods are trained with only the VGG-16 RGB feature. We report the performance of the proposed approach in both validation and test set as LSTM-RGB-local model has been evaluated on validation set [9]. In row-2.3, we report results for state-of-the-art approaches MCN [9] and TGN [3]. We also report results of CCA [17] and natural language object retrieval based baseline Txt-Obj-Retrieval [11] in row-2.3. These methods additionally use optical flow feature along with VGG16 RGB feature. We report the performance of MCN [9], TGN [3] and EFRC [40] from the respective papers. The results of LSTM-RGB-Local, Txt-Obj-Retrieval, Random, and Upper-Bound are reported from [9].

Analysis of Results. Similar to the results on Charades-STA, it is evident from Table 2 that our proposed weakly supervised approach consistently shows comparable performance to several fully-supervised approaches. From row-

Table 2. This table reports results on the DiDeMo dataset, following the evaluation protocol in [9]. Our **proposed approach** performs on par with several competitive fully-supervised approaches

#	Method	R@1	R@5	mIoU
2.1	Upper Bound	74.75	100	96.05
	Random	3.75	22.5	22.64
2.2	LSTM-RGB-Local [9]	13.10	44.82	25.13
	EFRC [40]	13.23	46.98	27.57
	Proposed (Val. Set)	**11.18**	**35.62**	**24.47**
	Proposed (Test Set)	**12.19**	**39.74**	**24.92**
2.3	CCA	18.11	52.11	37.82
	Txt-Obj-Retrieval [11]	16.20	43.94	27.18
	MCN [9]	27.57	79.69	41.70
	TGN [3]	28.23	79.26	42.97

2.2, we observe that our proposed approach achieves similar performance as LSTM-RGB-Local [9] and EFRC [40]. We observe that R@5 accuracy is slightly lower for our approach compared to supervised approaches. However, R@1 accuracy and mIoU is almost similar. Comparing row-2.3, we observe that the performance is comparable to CCA and Txt-Obj-Retrieval baselines. The performance is low compared to MCN [9] and TGN [3]. Both of the approaches use additional optical flow features in their framework. MCN additionally use a moment-context feature. Hence, a performance drop is not unexpected. However, we have already observed from the row-2.2 that the performance of our weakly supervised approach is comparable to the MCN baseline model of LSTM-RGB-Local which uses the same RGB feature in training as our method.

4.4. Qualitative Results

We provide six qualitative examples of moments predicted by the proposed approach from Charades-STA dataset [8] in Fig. 4. In Fig. 4, case 1, 2, and 4 show some examples where our approach was successful in retrieving the ground truth moment with high IoU. Cases 1 and 2 are examples where the same video has been used to retrieve different moments based on two different text descriptions. We see our text-aware attention module was successful in finding the correct segment of the video in both the cases.

While our method retrieves the correct moment from sentence description many cases, it fails to retrieve the correct moment in some cases (e.g., case 3, 5, and 6). Among these three cases, case 3 presents an ambiguous query where the person stands on the doorway but does not enter into the room. The GT moment covers a smaller segment, while our system predicts a longer one. We observe the performance of our system suffers when important visual contents occupy only small portions in frames, e.g., case 5 and 6. In case 6, a sandwich is mentioned in the query which occupies a small portion of frames initially and our framework shifted the start time of the moment to a much later time
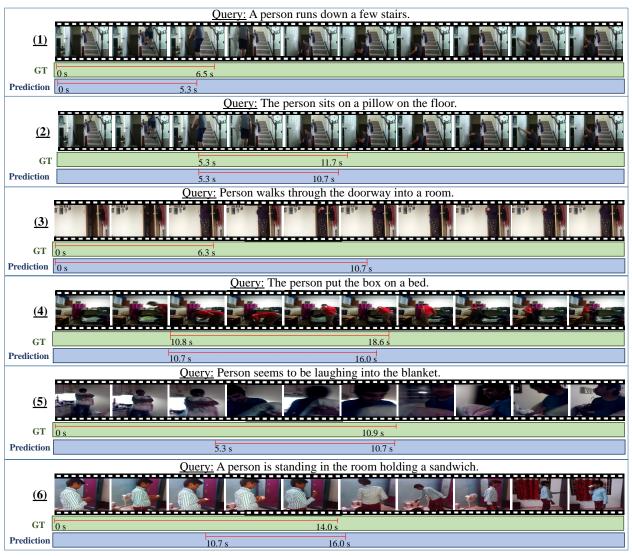
Figure 4. A snapshot of six queries and test videos from Charades-STA dataset with success and failure cases. GT is a ground-truth moment annotation and Prediction is the moment predicted by the proposed weakly-supervised approach. Queries 1, 2, and 4 show success cases where our approach was successful in retrieving the ground truth moment with very high temporal intersection over union (IoU). However, queries 3, 5, and 6 show three cases where our approach was not successful in retrieving the ground truth moment with high IoU.

instant than in the ground truth. Similarly, in case 5, our system was only successful in identifying the person laughing into a blanket after the scene is zoomed in. We believe these are difficult to capture without additional spatial attention modeling or generating region proposals. Moreover, utilizing more cues from videos (e.g., audio, and context) may be helpful in reducing ambiguity in these cases. We leave these as future work.

5. Conclusions

In this paper, we introduce the novel problem of weakly supervised text to video moment retrieval. In the weakly supervised paradigm, as we do not have access to the temporal boundaries associated with a sentence description, we utilize an attention mechanism to learn the same using only video-level sentences. Our formulation of the task makes it more realistic compared to existing methods in the literature which require supervision as temporal boundaries or temporal ordering of the sentences. Moreover, the weak nature of the task allows it to learn from easily available web data, which requires minimal effort to acquire compared to manual annotations. Experiments demonstrate that our method in spite of being weakly supervised performs comparably to several fully supervised methods in the literature.

References

[1] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Weakly-supervised alignment of video with text. In *ICCV*, 2015.

[2] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018.

[3] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018.

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[5] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. In *BMVC*, 2018.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017.

[9] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[10] Christian Andreas Henning and Ralph Ewerth. Estimating the information gap between textual and visual representations. In *ACM ICMR*, pages 14–22, 2017.

[11] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[12] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[13] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897, 2014.

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[16] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015.

[17] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[19] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM Multimedia*, 2018.

[20] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *CVPR*, pages 1942–1950, 2016.

[21] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ACM ICMR*, 2018.

[22] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Joint embeddings with multimodal cues for video-text retrieval. *IJMIR*, pages 1–16, 2019.

[23] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos Papalexakis, and Amit Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *ACM Multimedia*, 2018.

[24] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

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

[26] 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.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Workshop*, 2017.

[28] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, pages 588–607. Springer, 2018.

[29] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *CVPR*, pages 1961–1970, 2016.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[34] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

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

[36] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE T-PAMI*, 41(2):394–407, 2019.

[37] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[38] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017.

[39] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5794–5803, 2017.

[40] Huijuan Xu, Kun He, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Text-to-clip video retrieval with early fusion and re-captioning. *arXiv preprint arXiv:1804.05113v1*, 2018.

[41] 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.

[42] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 5, page 6, 2015.

[43] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. Multi-networks joint learning for large-scale cross-modal retrieval. In *ACM Multimedia*, pages 907–915, 2017.

[44] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017.

[45] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.