Text-to-Clip Video Retrieval with Early Fusion and Re-Captioning

Huijuan Xu[1], Kun He[1], Leonid Sigal[2], Stan Sclaroff[1], and Kate Saenko[1]

[1]Boston University, [2]University of British Columbia
[1]{hxu, hekun, sclaroff, saenko}@bu.edu, [2]lsigal@cs.ubc.ca

Abstract. We propose a novel method capable of retrieving clips from untrimmed videos based on natural language queries. This cross-modal retrieval task plays a key role in visual-semantic understanding, and requires localizing clips in time and computing their similarity to the query sentence. Current methods generate sentence and video embeddings and then compare them using a late fusion approach, but this ignores the word order in queries and prevents more fine-grained comparisons. Motivated by the need for fine-grained multi-modal feature fusion, we propose a novel early fusion embedding approach that combines video and language information at the word level. Furthermore, we use the inverse task of dense video captioning as a side-task to improve the learned embedding. Our full model combines these components with an efficient proposal pipeline that performs accurate localization of potential video clips. We present a comprehensive experimental validation on two large-scale text-to-clip datasets (Charades-STA and DiDeMo) and attain state-of-the-art retrieval results with our model.

1 Introduction

Temporal localization of events or activities of interest is a key problem in computer vision, and recently there has been increased interest in specifying the queries directly using natural language. In this paper, we focus on solving the task of retrieving temporal segments in untrimmed video through natural language queries, or simply, "text-to-clip." A commonly adopted pipeline in existing solutions first generates candidate clips from videos and then retrieves nearest neighbors of the sentence query in those candidates, using a learned similarity metric. This similarity metric is what we focus on improving in this paper.

A general recipe for solving cross-modal retrieval tasks, such as text-to-clip, is to learn a common vector embedding space, project objects in different modalities (*e.g.* sentences and video clips) separately into this space, and compute standard similarity metrics. We refer to this as a *late fusion* approach, since information is not shared in the embedding processes. Although late fusion approaches are quite successful in many cross-modal tasks, we argue that for the fine-grained text-to-clip task, there is valuable sentence structure that does not get preserved by this approach. Specifically, the sentence embedding is usually generated by pooling the hidden states of a recurrent neural network, such

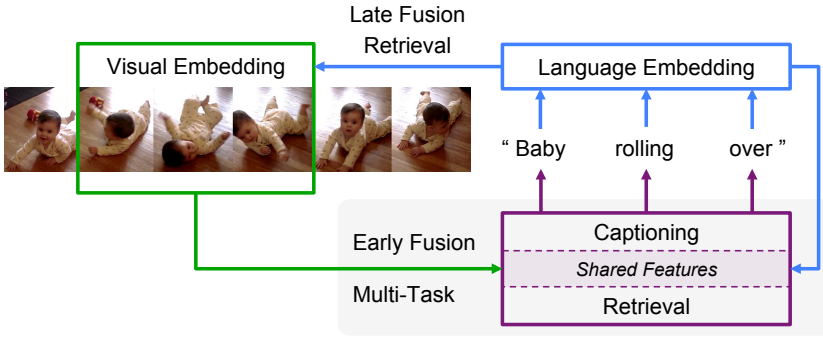


Fig. 1. We are interested in using natural language descriptions to retrieve events in untrimmed video. This problem is usually solved by a *late fusion* approach by learning a common vector embedding space. Instead, we propose an *early fusion* model that better preserves detailed sentence structure. Our model additionally benefits from a multi-task formulation that adds video captioning as a auxiliary task, using the retrieved video clip to re-generate the sentence query.

as a Long Short-Term Memory (LSTM), which is used to model the sentence. This essentially gives a representation that is averaged over the time dimension, which is not likely to capture fine-grained sentence structure. Even with attention mechanisms that weight the contributions of each word differently, without access to the visual content, it would be difficult for the attention mechanism to “anticipate” the visual content and adjust the weights accordingly.

We propose a novel *early fusion* approach for text-to-clip. Instead of embedding sentences and video clips separately to vectors, our learned similarity metric allows for more structured inference in the language modality. Specifically, we learn an integrated LSTM model that recurrently processes the query sentence, conditioned on the visual feature embedding, and produces a nonlinear similarity score in the end. Importantly, this model can potentially learn to associate each word in the sentence query with different portions of the visual features within a video, which is not possible in a late fusion model. Experimentally, early fusion significantly improves over late fusion approaches. Current approaches adopt token level attention, thus each word is able to be linked with corresponding visual feature.

We also improve the learned similarity metric through a novel multi-task formulation. This is inspired by the fact that the inverse task of text-to-clip, video dense captioning [19], is also a valuable task that shares a demand for cross-modal feature fusion. Therefore, we conjecture that learning shared feature representations in one task is likely to help the other. We thus add video captioning as an auxiliary task alongside text-to-clip, and demonstrate further improved retrieval performance.

To summarize our contributions, in this paper we:

- take an early fusion approach to tackle the text-to-clip retrieval task, modeling fine-grained structure in the query,
- leverage the captioning task to learn better shared feature representations and improve retrieval performance.

Besides a good similarity metric, solving the text-to-clip task also requires a temporal localization component in the pipeline, for initially proposing candidate clips. For this component, differently from existing work that employ computationally-expensive sliding windows or handcrafted heuristics, we adopt an accurate temporal segment proposal network from the R-C3D model [50], originally designed for activity detection. Our full model achieves state-of-the-art retrieval performance on two challenging benchmarks: Charades-STA [9] and DiDeMo [11]. Code will be released for public use.

2 Related Work

Activity Detection and Temporal Proposals: Fine-grained video understanding often requires localizing activities of interest in time. The problem of activity localization, or activity detection, is to predict the start and end times of the activities within untrimmed videos. Early approaches [36,46] use sliding windows to generate segments and subsequently classify them, which is computationally inefficient and constrains the granularity of detection. More recent approaches have bypassed exhaustive sliding window search to detect activities with arbitrary lengths. In [25,38,33] temporal localization is obtained by modeling the evolution of activities using Recurrent Neural Networks (RNNs) and predicting activity labels or activity segments at each time step. CDC [35] and SSN [60] propose bottom-up activity detection by first predicting at the frame-level/snippet-level and then fusing them. Temporal action proposals are studied in [4,7,59]. R-C3D [50] adapts the proposal and classification pipeline from object detection [28] to perform activity detection using 3D convolutions [41] and 3D Region of Interest pooling, and SSAD [20] performs single-shot temporal activity detection following the one-stage object detection method SSD [23]. In this paper, we use a proposal-based pipeline to solve the video language localization task, and adopt the proposal generation technique of R-C3D.

Another thread of activity detection research is spatio-temporal detection, which involves localizing the activities in "action tubes." For example, [10,31,48], [57,61] temporally track bounding boxes corresponding to activities in each frame. Other recent models [13,17,30] propose to first detect small tubelets spanning multiple frames, and connect them into final detection tubes using heuristics. [3,27] produce spatiotemporal saliency maps aimed at explaining generated captions or activity classifications, with the side effect of spatiotemporal localization of salient activities.

A limitation of existing activity localization methods is that they treat activities as distinct classes, and therefore require a discrete and fixed vocabulary of class labels. Instead, we solve the task of temporally localizing free-form language queries in videos, and our approach can be potentially extended to spatiotemporal localization.

Vision and Language: The main problem that we solve in this paper is a typical vision-language task: cross-modal retrieval of visual events that match a query sentence. There are two main types of approaches to solve such cross-modal

retrieval tasks: early fusion and late fusion. The late fusion approach embeds different modalities into a common embedding space, and then measures the similarity between the feature embeddings using a standard inner product or cosine similarity. In fact, such approaches are not restricted to vision and language, and can be applied across modalities such as image, video, text, and sound [1,2,42]. The early fusion approach combines the features from each modality at an earlier stage [24,47,58] and predicts similarity scores directly based on the fused feature representation. [5] argues against the dominant late-fusion pipeline where linguistic inputs are mostly processed independently, and shows that modulating visual representations with language at earlier levels improves visual question answering. For the text-to-clip task considered in this paper, existing models [9,11] perform late fusion at the sentence level: they embed the query sentence into a single vector and only then combine it with the video feature vector. However, this removes information about word order in the feature fusion, which may be important for computing the score. In this paper, we propose a text-to-clip retrieval model that performs early fusion of the video and query features, combining them at the word level, and we compare this early fusion model with the late fusion and sentence-level fusion approaches.

Other typical vision-language tasks include image/video captioning [6,43,44], [45,53,54,56] and visual question answering (VQA) [39,49,52,55]. We note that these tasks are rarely isolated and often influence each other. For example, image captioning can be solved as a retrieval task [8]. Also, there is recent research that suggests that VQA can be leveraged to benefit the image-caption retrieval task [21]. Our proposed multi-task formulation, which uses captioning as an auxiliary task, is partly motivated by these observations.

Localization-based Cross-modal Tasks: Several vision-language tasks also share the need for a localization component. Hu, et al. [14] propose the task of natural language object retrieval, which localizes objects in images given language queries. Rohrbach, et al.[29] propose models for grounding textual phrases in images by reconstruction with different levels of supervision. In the dense captioning task, models need to localize interesting events in images [16] or videos [19,34,51] and provide textual descriptions. Recently, the task of grounding text in images has been extended into videos, which introduces the task of retrieving video segments using language queries [11,9]. We note that the localization mechanisms in [11,9] are either inefficient (sliding-window based) or inflexible (hard-coded). In contrast with these approaches, we adopt segment proposals as the first step in our multi-modal retrieval pipeline.

3 Approach

We propose a novel approach for temporal activity localization and retrieval based on input language queries, or the text-to-clip task. This is posed as a cross-modal retrieval problem. Our key idea is to integrate language and vision more closely before computing a match, using an early fusion scheme and a multi-task formulation that re-generates the caption.

We first define the cross-modal retrieval problem we are solving. Given an untrimmed video V and a sentence query S, the goal is to retrieve a temporal segment (clip) R in V that best corresponds to S. In other words, we learn a mapping $\mathcal{F}_{\text{RET}} : (V, S) \mapsto R$. At training time, we are given a set of annotated videos $\{V_1, V_2, \ldots, V_N\}$. For each video V_i, its annotation is a set of matching sentence-clip pairs $A_i = \{(S_{ij}, R_{ij})\}_{j=1}^{n_i}$, where S_{ij} is a sentence, and clip $R_{ij} = (t_{ij}^0, t_{ij}^1)$ is represented as a pair of timestamps that define its start and end. We tackle the retrieval problem through learning a similarity score $\sigma(S, R) \in \mathbb{R}$ that measures how well S and R match each other. At test time, given V and S, the retrieval problem is formulated as

$$R^* = \arg\max_{R \in V} \sigma(R, S). \tag{1}$$

On the other hand, the video dense captioning task involves generating sentence descriptions for densely generated temporal segments in video. It can be formulated as an inverse task: $\mathcal{F}_{\text{CAP}} : (V, R) \mapsto S$, assuming a mechanism for generating the temporal segments R is available. A typical solution is to train a recurrent neural network that predicts each word in the sentence sequentially, conditioned on the visual features extracted from R in V.

We will link these two tasks in our proposed model. Unlike current clip retrieval models, video captioning models integrate visual features with language at the word level. This inspires our early fusion architecture, as well as the addition of captioning as an auxiliary loss. But first, we describe the shared component, the Segment Proposal Network, used to generate the set of temporal segments R. In the remainder of this section, we first introduce the localization component, the segment proposal network, in Sec. 3.1. We then describe our early fusion model in Sec. 3.2, and contrast it with late fusion. Next, Sec. 3.3 introduces a multi-task formulation that adds captioning as an auxiliary task. Finally, implementation details are discussed in Sec. 3.4.

3.1 Segment Proposal Network

For unconstrained localization in videos, it is important to generate variable-length candidate temporal segments for further processing. However, generating exhaustive multiscale sliding windows in videos is computationally expensive, and we need a selective strategy. We employ a segment proposal network (SPN), similar to the one used in R-C3D [50] for action localization.

Figure 2 (left) depicts the segment proposal network. Given input video V, the segment proposal network first encodes all input frames in V using a 3D convolutional network (C3D). Then, variable-length segment proposals are obtained by predicting a relative offset to the center location and the length of a set of predefined anchor segments. To compute a visual representation of each proposal R, we encode predicted proposals into features $f(R)$ by 3D Region of Interest Pooling, and the fc6 layer of the C3D network [41].

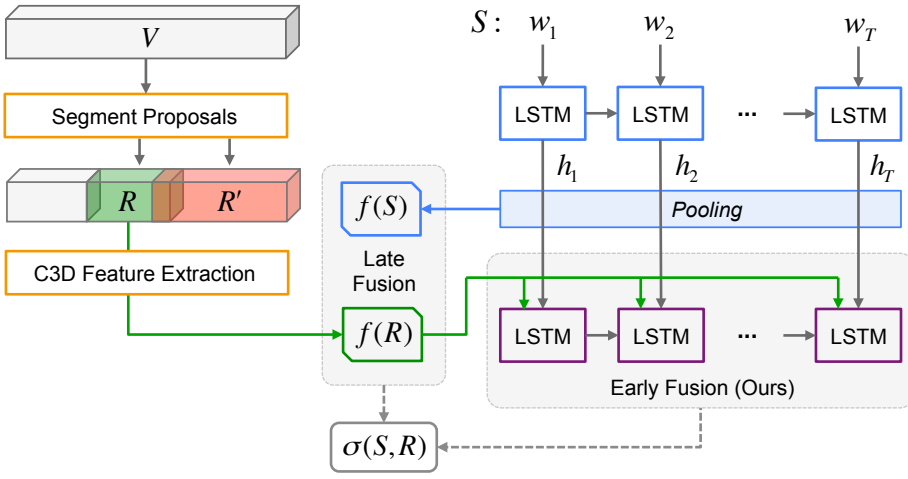


Fig. 2. Our goal is to retrieve the clip R in video V that best corresponds to query sentence S. **Left**: We use the segment proposal network in R-C3D [50] to generate candidate clips and extract visual features $f(R)$. **Right**: In the baseline late fusion model, the sentence feature $f(S)$ is formed by pooling the hidden states from a sentence embedding LSTM, and similarity is computed between embedding vectors $f(S)$ and $f(R)$. Our early fusion model uses an additional LSTM layer, conditioned on $f(R)$ at each step, to directly predict the similarity score $\sigma(S, R)$.

3.2 Early Fusion Retrieval Model

In this subsection, we introduce our retrieval model using early fusion, utilizing the proposals from the segment proposal network. Before that, we first describe a retrieval model that uses late fusion, which will serve as a baseline to our model later.

In the baseline late-fusion model, illustrated in Fig. 2 (right), proposal video segments and query sentences are embedded into a common vector embedding space, where similarity between vectors can be measured. To compute the sentence feature $f(S)$, a common strategy is to take the word embeddings $\{w_t\}_{t=1}^{T}$ of each word in S, and feed them into a sentence embedding LSTM. Then, $f(S)$ is pooled from the hidden states of the embedding LSTM, which can simply be the last hidden state, or more generally a weighted average. Next, a retrieval loss is applied to enforce ranking constraints on the similarity measure, such that ground truth sentence-clip pairs always score the highest.

The drawback of the late fusion model is that the sentence is represented in a holistic manner. As a result, fine-grained word sequence information is lost by the time the video and language features are fused together to compute similarity. We now introduce our early fusion model that mitigates this problem.

As also shown in Fig. 2, our early fusion model takes the form of a two-layer LSTM, where the first layer is the previous sentence embedding LSTM. In the second layer, the visual feature embedding is used as input at each step, along

with hidden states from the sentence embedding LSTM. The final hidden state is passed through additional layers to predict a scalar similarity value. We note that this is not simply an increase in the number of learnable parameters in the model, but brings additional structure into the similarity metric: since each word in the sentence now can interact with the visual feature, the model can learn to associate each word with a different part of visual feature. We do not explicitly use attention mechanisms to enforce such behavior, but instead let the LSTM learn in a data-driven manner.

In this work, we use a triplet-based retrieval loss (also called pairwise ranking loss [15]), which has shown good performance in metric learning tasks [12,32]. Specifically, we take triplets of the form (S, R, R') where (S, R) is a matching sentence-clip pair, and R' is some clip sampled from a negative set $\mathcal{N}(S)$ that does not match S. Note that R' can either come from the same video as R with a low overlap, or a different video. The loss encourages the similarity score between the matching pair, $\sigma(S, R)$, to be greater than $\sigma(S, R')$ by some margin $\eta > 0$:

$$L_{\text{RET}} = \sum_{(S,R)} \sum_{R' \in \mathcal{N}(S)} \max\{0, \eta + \sigma(S, R') - \sigma(S, R)\}. \tag{2}$$

For the late fusion model, $\sigma(S, R)$ is computed as the cosine similarity between embedding vectors $f(S)$ and $f(R)$, *i.e.* $\sigma(S, R) = \frac{\langle f(S), f(R) \rangle}{\|f(S)\| \|f(R)\|}$. In our early fusion model, $\sigma(S, R)$ is directly predicted by the LSTM.

3.3 Captioning as Auxiliary Task

After defining the retrieval model, we now seek to gain additional benefit from multi-tasking, specifically, by adding a captioning loss.

A motivation for the multi-task formulation is that captioning serves as verification for retrieval: if a separate model is able to re-generate the query sentence from the retrieved video clip, then it verifies the correctness of retrieval, in the sense that all necessary semantic meaning is retained in the visual representations. Moreover, it is observed in the captioning literature that captioning models can implicitly learn features and attention mechanisms to associate spatial/spatiotemporal regions to words in the captions [27]. Conversely, we also expect such mechanisms to benefit retrieval, since a model would be able to look for features/regions associated with words in the input query.

With the reasoning above, we now add a captioning loss into the training of the early-fusion retrieval model. Note that the paired sentence-clip annotation format in the text-to-clip task allows us to easily add captioning capabilities to our LSTM model. Specifically, we require the top-layer LSTM to re-generate the input query sentence, conditioned on the proposal's visual features $f(R)$ at each step. When generating word w_t at step t, the hidden state from the previous step in the sentence embedding LSTM, $h^{(1)}_{t-1}$, is used as input. We use a standard captioning loss that maximizes the normalized log likelihood of the

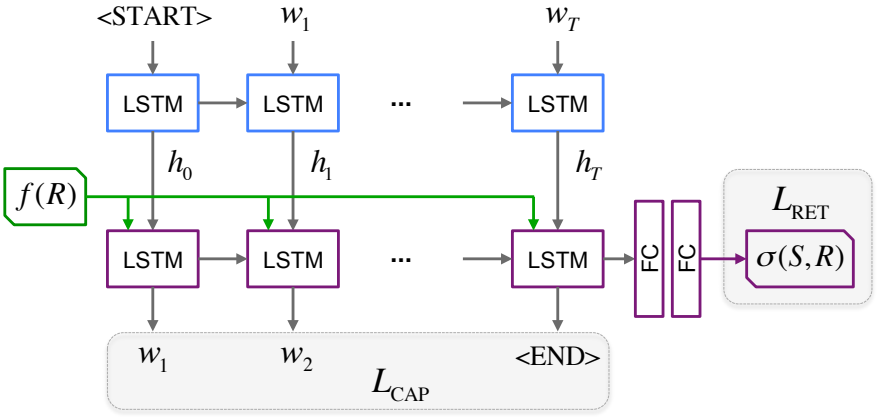


Fig. 3. Our early fusion model with multi-task loss. We add a captioning loss L_{CAP} to our top-layer LSTM, which enforces it to re-generate the input sentence query as a caption for the retrieved video clip. This serves as a verification for the retrieval task, and also helps to learn better fusion features, resulting in improved retrieval performance.

words generated at all T unrolled time steps, over all K ground truth matching sentence-clip pairs:

$$L_{\text{CAP}} = -\frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\log P(w_t^k | f(R), h_{t-1}^{(2)}, w_1^k, ..., w_{t-1}^k). \tag{3}$$

With our early fusion approach, we can ensure that gradients from both losses reach the same set of underlying layers, and act on the visual and sentence representations at the same time. The early fusion model with multi-task loss is illustrated in Fig. 3.

3.4 Implementation Details

Our multi-task model optimizes a weighted combination of retrieval loss and captioning loss, with a weighting parameter λ:

$$L = L_{\text{RET}} + \lambda L_{\text{CAP}}. \tag{4}$$

We choose $\lambda = 0.5$ through cross-validation. The margin parameter η is set to 0.2 in the retrieval loss L_{RET}. During training, each minibatch contains 32 matching sentence-clip pairs sampled from the training set, which are then used to construct triplets. We use the Adam optimizer [18] with learning rate 0.001 and early stopping on the validation set, for 30 epochs in total.

For the sentence embedding LSTM, we use `word2vec` [26] as the input word representation. The word embeddings are 300-dimensional, and trained from

scratch on each dataset. The hidden state size of the LSTM is set to 512. The size of common embedding space in the late fusion retrieval model is 1024.

For the early fusion model, which outputs a nonlinear similarity score, we take the hidden state corresponding to the last word in the second LSTM, and pass it through two fully-connected (FC) layers to produce a scalar value σ, as shown in Fig. 3. The two FC layers reduce the dimensionality from 512 to 64 to 1. A sigmoid activation is applied after the FC layers.

At test time, retrieving clips in untrimmed videos involves searching over all possible proposal segments. Candidate proposal segments generated from the proposal network are filtered by non-maximum suppression with threshold 0.7, and the top 100 proposals in each video are kept.

4 Experiments

We evaluate our proposed models on two recent datasets designed for the text-to-clip retrieval task: Charades-STA [9] and DiDeMo [11]. We consider several methods for comparison. First, `Random` is a baseline that randomly selects among candidate clips. `LateFusion` is another baseline that directly measures similarity between visual and sentence-level embedding vectors using the cosine similarity metric. `LateFusion+Cap` is the `LateFusion` model with captioning loss. Our proposed `EarlyFusion` model merges visual features and word-level embeddings at an early stage, and finally, `EarlyFusion+Cap` is our full model with the captioning loss.

We follow the evaluation setup in [9], which is adapted from a similar task in the image domain, namely the task of object retrieval with natural language descriptions [14]. Specifically, we consider a set of temporal Intersection-Over-Union (tIoU, or simply IoU) thresholds. For each threshold τ, we compute the Recall@K metric, defined as the fraction of sentence queries having at least one correct retrieval (having tIoU greater than τ with ground truth) in the top K retrieved video clips. Following standard practice, we use $\tau \in \{0.3, 0.5, 0.7\}$ and $K \in \{1, 5, 10\}$. We present experimental details and results on the Charades-STA dataset in Sec. 4.1, and on the DiDeMo dataset in Sec. 4.2.

4.1 Experiments on the Charades-STA Dataset

Dataset and Setup: The Charades-STA dataset was introduced by Gao *et al.* [9] for evaluating temporal localization of events in video given natural language queries. The original Charades dataset [37] only provides a paragraph description for each video. To generate sentence-clip annotations used in the retrieval task, the authors of [9] decomposed the original video-level descriptions into shorter sub-sentences, and performed keyword matching to assign them to temporal segments in videos. The alignment annotations are further verified manually. The released annotations comprise 12,408 sentence-clip pairs for training, and 3,720 for testing.

Methods	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 [9]	–	–	–	8.5	37.1	–	3.0	14.1	–
CTRL(reg-np) [9]	–	–	–	23.6	58.9	–	8.9	29.5	–
`LateFusion`	43.9	83.5	89.7	26.3	63.9	78.2	10.9	35.6	50.5
`LateFusion+Cap`	44.7	83.4	90.6	27.0	63.5	77.8	10.6	35.4	50.4
`EarlyFusion`	51.6	**95.5**	**99.0**	32.8	76.3	**92.5**	14.0	43.2	60.7
`EarlyFusion+Cap`	**53.0**	94.6	98.5	**33.8**	**77.3**	91.6	**15.0**	**43.9**	**60.9**

Table 1. Results on the Charades-STA dataset [9]. R@K stands for Recall@K. Our early fusion retrieval model `EarlyFusion` significantly outperforms baselines, while the multi-task `EarlyFusion+Cap` further improves results.

We keep all the words that appear in the training set to build a vocabulary of size 1,111. The maximum caption length is set to 10. We sample frames at 5 fps for this dataset and set the number of input frames to 768, breaking arbitrary-length input videos into 768-frame chunks, and zero-padding them if necessary. To initialize our segment proposal network, we finetune a 3D ConvNet model [41] pretrained on Sports-1M, with the ground truth activity segments of 157 classes in the training videos of the Charades activity detection dataset. We then extract proposal visual features, and train the retrieval model from random initialization.

Results: Table 1 shows the results on the text-to-clip retrieval task for Charades-STA. First, it is interesting to note that our baseline `LateFusion` retrieval model already outperforms the best model in [9], CTRL (reg-np), by a noticeable margin. We believe there are two reasons for this. First, our segment proposal network offers finer temporal granularity, and therefore provides cleaner visual feature representations compared to the sliding windows approach in CTRL. Second, we use a triplet-based loss that more effectively captures ranking constraints, compared to CTRL's binary classification loss. On the other hand, adding the multi-task captioning loss to the late fusion model (`LateFusion+Cap`) attains nearly the same result as `LateFusion`. We note that since late fusion uses a sentence-level wholistic embedding derived from the hidden states of the lower-level sentence LSTM, the higher-level captioning loss does not have a direct effect.

Our `EarlyFusion` model significantly outperforms the late fusion approaches. Due to the direct sharing of parameters between two tasks in the fusion LSTM layer, `EarlyFusion+Cap` is able to further improve results. These improvements are more salient with respect to higher IoU thresholds.

We provide an ablation study of the different forms of sentence embedding in `LateFusion+Cap`, shown in Table 2. Instead of simply using the last hidden state from the sentence embedding LSTM, using a weighted average of all hidden states (mean pooling or self-attention [22]) can give marginal improvements, but results are still significantly below those of `EarlyFusion`. Further ablations of

Sentence Embedding	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
Last hidden state	44.7	83.4	90.6	27.0	63.5	77.8	10.6	35.4	50.4
Mean pooling	43.9	89.2	93.3	26.2	68.5	82.4	11.1	34.5	51.2
Self attention	43.8	89.1	93.6	26.4	68.0	84.4	11.1	35.4	50.4

Table 2. Comparison between different forms of sentence embedding for producing the sentence embedding $f(S)$ in the `LateFusion+Cap` method, measured on the Charades-STA dataset. R@K stands for Recall@K.

Loss Weight	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
$\lambda = 0.5$	**53.0**	94.6	**98.5**	**33.8**	**77.3**	**91.6**	**15.0**	**43.9**	**60.9**
$\lambda = 1$	50.8	94.5	98.1	32.5	76.1	91.2	14.1	41.9	59.2
$\lambda = 2$	50.6	**94.9**	**98.5**	33.5	76.5	91.3	14.3	43.4	60.3

Table 3. The effect of loss weight λ in the `EarlyFusion+Cap` method, measured on the Charades-STA dataset. R@K stands for Recall@K. As our main task is retrieval, we consistently underweight the captioning loss with $\lambda = 0.5$ in our experiments.

the captioning loss weight λ in Eq. 4 for training the `EarlyFusion+Cap` method are shown in Table 3. As our main task is retrieval, we choose $\lambda = 0.5$ in our experiments.

Two example videos from the Charades-STA dataset along with query localization results are shown in Figure 4(a). The correct prediction is marked as green, while the wrong one is marked as red. Please note that the prediction is in fact correct for the query *Person takes out a towel*, but is marked incorrect due to inaccurate ground truth.

4.2 Experiments on the DiDeMo Dataset

Dataset and Setup: The DiDeMo dataset was recently proposed by Hendricks *et al.* [11], specifically for the temporal localization of events in video given natural language descriptions, using videos from Flickr [40]. To reduce the complexity of annotation, videos in this dataset are trimmed to a maximum of 30 seconds, split into 5-second segments, and each clip (called a "moment") includes one or more 5-second segments. The sentence descriptions in DiDeMo are ensured to be referring expressions so that they point to specific moments in each video, and so that each description refers to a single moment. The training, validation and test sets contain 8,395, 1,065 and 1,004 videos, respectively, with a total of 26,892 clips and 40,543 sentences; a clip could be associated with multiple descriptions. Compared to object retrieval and video summarization datasets, sentences in the DiDeMo dataset contain more indicators of camera movement and temporal transition, as well as verbs, which are more informative for understanding actions in time.

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`	14.2	49.1	69.0	6.3	26.5	43.9	2.0	10.4	19.3
`LateFusion`	**22.1**	63.9	84.3	10.9	37.9	59.1	4.5	17.7	30.3
`LateFusion+Cap`	21.9	64.9	84.0	11.0	39.3	59.6	4.3	17.4	30.2
`EarlyFusion`	20.9	69.4	**90.3**	10.6	40.6	**66.4**	4.3	17.8	31.9
`EarlyFusion+Cap`	21.9	**70.6**	89.3	**11.9**	**42.5**	66.1	**5.5**	**19.0**	**33.6**

Table 4. Results on the DiDeMo dataset [11]. R@K stands for Recall@K. Our early-fusion retrieval model with captioning supervision `EarlyFusion+Cap` significantly outperforms other baselines.

Method	Rank@1	Rank@5	mIoU
Random [11]	3.75	22.5	22.64
LSTM-RGB-local [11]	13.1	44.82	25.13
`LateFusion`	11.04	43.27	26.38
`LateFusion+Cap`	10.40	42.28	26.23
`EarlyFusion`	12.81	45.14	27.42
`EarlyFusion+Cap`	**13.23**	**46.98**	**27.57**

Table 5. Results on the DiDeMo dataset, using the evaluation protocol in [11]. Our early-fusion retrieval model with captioning supervision `EarlyFusion+Cap` outperforms other baselines, using RGB input for fair comparison.

We keep all the words in the training set to build a vocabulary of size 6,664, and set the maximum caption length to 25. We sample frames at 12.5 fps, and set the maximum number of input frames in a video to be 512, considering the fact that all the videos are around 30 seconds long. Again, a 3D ConvNet model [41] pretrained on the Sports-1M dataset is used to initialize our segment proposal network.

We also would like to discuss the evaluation metrics in the DiDeMo dataset. As mentioned earlier, DiDeMo only has coarse localization annotation, where each video is divided into 5-second segments. For a 30-second video, there are only 21 possible combinations of contiguous segments to assign to a clip. The evaluation procedure in [11] is specifically designed for this scenario: at test time, a model predicts similarity scores of all the 21 clips for a sentence query, and is evaluated against the ground truth in a "hit-or-miss" fashion, instead of a more commonly used soft criterion based on temporal intersection over union (tIoU). Since our method does not rely on coarse heuristics for localization, using more accurate segment proposals could actually be penalized in such a rigid evaluation protocol, which does not consider soft matches. Therefore, we report results using the more standard "IoU=τ, Recall@K" protocol used above for our methods on DiDeMo.

Results: Results using the standard evaluation protocol are given in Table 4. Similar trends can be observed for the four variants of our model, as

in the Charades-STA experiments. `EarlyFusion` significantly outperforms both baselines, `LateFusion` and `LateFusion+Cap`, whose relative performances are similar. Also, with the assistance of the captioning loss, the multi-task model `EarlyFusion+Cap` does better than `EarlyFusion`, which is more evident with higher tIoU thresholds.

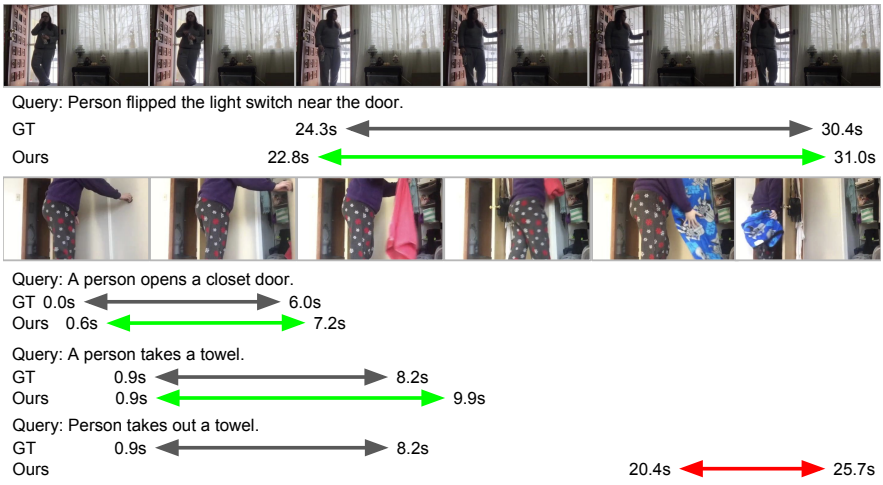
In addition, results using [11]'s "hit-or-miss" evaluation protocol are summarized in Table 5. As our models are trained only on RGB input, for conducting fair comparisons, we compare to the "LSTM-RGB-local" model trained on RGB input from [11], and note that [11]'s fusion models additionally use optical flow and a "temporal endpoint feature" as input. In Table 5, "mIoU" stands for the average tIoU of the top-1 retrieved segment with respect to ground truth annotation. Using this soft-match metric, all of our model variants actually outperform "LSTM-RGB-local", with `EarlyFusion+Cap` being the top performer. On the other hand, for the more rigid Rank@1 and Rank@5 metrics that only consider exact matches, `EarlyFusion+Cap` also outperforms "LSTM-RGB-local".

Two example retrieval results from the DiDeMo dataset can be found in Figure 4(b). In the first example, our model very accurately localizes the precise moment described by the query sentence, *Roller coaster first begins to move*. In the second example, it also correctly identifies the event corresponding to the query *Group of people exit frame left*, however, the temporal overlap is deemed less than 0.5 with human annotation. Note that the ground truth in this dataset is always specified in terms of 5-second segments, while our method is able to generate variable-length temporal localizations.

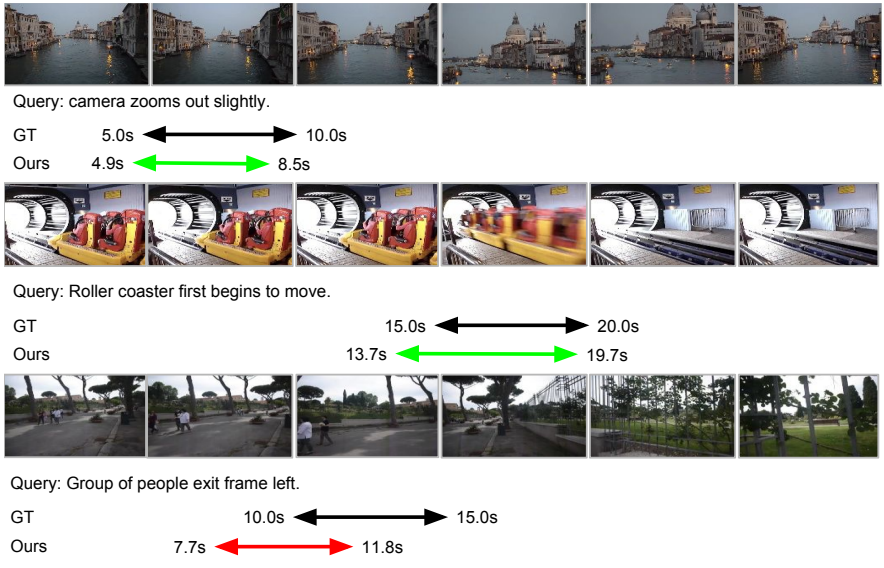
5 Conclusion

In this paper, we address the problem of text-to-clip retrieval: temporal localization of events within videos that match a given natural language query. We introduce an early fusion technique, which modulates the the integration of word-level language features using visual information in a recurrent LSTM model, and improves upon commonly used late fusion approaches that are based on vector embeddings. Motivated by the interplay between vision-language tasks, we also propose to add re-captioning as an auxiliary task, and we make use of a segment proposal network to filter out unlikely clips. Evaluated on two challenging datasets, our approach performs more accurately than existing methods when retrieving clips from many possible candidates in untrimmed videos. For example, on the Charades-STA dataset, we achieve a significant improvement in the recall at top 5 retrievals with 0.5 temporal overlap, from 58.9% in [9] to 77.3% with our model. We also provide detailed ablation studies to confirm the benefits of our proposed formulations.

An interesting future direction is to improve the segment proposal network by conditioning it on the input sentence query, in order to produce fewer, but better, query-guided proposal segments in the subsequent retrieval. Also, as our early fusion model explored the modulation of language features using visual



(a) Charades-STA retrieval examples



(b) DiDeMo retrieval examples

Fig. 4. Qualitative visualization of the retrieval results of our `EarlyFusion+Cap` method on the Charades-STA dataset (a) and the DiDeMo dataset (b). Ground truth clips are marked with black arrows. Predicted clips are marked in green for correct predictions (temporal IoU more than 0.5 with ground truth) and in red for incorrect ones. Corresponding start-end times are shown. (Best viewed in color)

information, we are also interested in the other direction, namely, using language features to modulate the extraction of visual features, similar to [5].

References

1. Arandjelović, R., Zisserman, A.: Objects that sound. arXiv preprint arXiv:1712.06651 (2017)
2. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
3. Bargal, S.A., Zunino, A., Kim, D., Zhang, J., Murino, V., Sclaroff, S.: Excitation backprop for RNNs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
4. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Niebles, J.C.: SST: Single-stream temporal action proposals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 6373–6382
5. de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. In: Advances in Neural Information Processing Systems (NIPS). (2017) 6594–6604
6. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 2625–2634
7. Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B.: DAPs: Deep action proposals for action understanding. In: European Conference on Computer Vision (ECCV). (2016)
8. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1473–1482
9. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: Temporal activity localization via language query. In: IEEE International Conference on Computer Vision (ICCV). (2017)
10. Gkioxari, G., Malik, J.: Finding action tubes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 759–768
11. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: IEEE International Conference on Computer Vision (ICCV). (2017)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, Springer (2015) 84–92
13. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (t-cnn) for action detection in videos. In: IEEE International Conference on Computer Vision (ICCV). (2017)
14. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 4555–4564
15. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence **172**(16-17) (2008) 1897–1916
16. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
17. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: IEEE International Conference on Computer Vision (ICCV). (2017)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: IEEE International Conference on Computer Vision (ICCV). (2017)
20. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proc. ACM Conference on Multimedia. (2017) 988–996
21. Lin, X., Parikh, D.: Leveraging visual question answering for image-caption ranking. In: European Conference on Computer Vision (ECCV). (2016) 261–277
22. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
23. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision (ECCV). (2016) 21–37
24. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: IEEE International Conference on Computer Vision (ICCV). (2015) 2623–2631
25. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in LSTMs for activity detection and early detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS). (2013) 3111–3119
27. Ramanishka, V., Das, A., Zhang, J., Saenko, K.: Top-down visual saliency guided by captions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). (2015)
29. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision, Springer (2016) 817–834
30. Saha, S., Singh, G., Cuzzolin, F.: Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 4414–4423
31. Saha, S., Singh, G., Sapienza, M., Torr, P.H., Cuzzolin, F.: Deep learning for detecting multiple space-time action tubes in videos. arXiv preprint arXiv:1608.01529 (2016)
32. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 815–823
33. Serena Yeung afinally, n.O.R., Mori, G., Fei-Fei, L.: End-to-end Learning of Action Detection from Frame Glimpses in Videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 2678–2687
34. Shen, Z., Li, J., Su, Z., Li, M., Chen, Y., Jiang, Y.G., Xue, X.: Weakly supervised dense video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
35. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 1417–1426

36. Shou, Z., Wang, D., Chang, S.F.: Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
37. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision (ECCV). (2016)
38. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
39. Teney, D., Hengel, A.v.d.: Visual question answering as a meta learning task. arXiv preprint arXiv:1711.08105 (2017)
40. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. Communications of the ACM **59**(2) (2016) 64–73
41. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: IEEE International Conference on Computer Vision (ICCV). (2015)
42. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. ICLR (2016)
43. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: IEEE International Conference on Computer Vision (ICCV). (2015)
44. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2015) 1494–1504
45. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 3156–3164
46. Wang, L., Yu Qiao, Y., Tang, X.: Action Recognition and Detection by Combining Motion and Appearance Features. ECCV THUMOS Workshop (2014)
47. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2018)
48. Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Learning to track for spatio-temporal action localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
49. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International Conference on Machine Learning. (2016) 2397–2406
50. Xu, H., Das, A., Saenko, K.: R-C3D: Region convolutional 3D network for temporal activity detection. In: IEEE International Conference on Computer Vision (ICCV). (2017)
51. Xu, H., Li, B., Ramanishka, V., Sigal, L., Saenko, K.: Joint event detection and description in continuous video streams. arXiv preprint arXiv:1802.10250 (2018)
52. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision, Springer (2016) 451–466

53. Xu, H., Venugopalan, S., Ramanishka, V., Rohrbach, M., Saenko, K.: A multi-scale multiple instance video description network. arXiv preprint arXiv:1505.05914 (2015)
54. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. (2015) 2048–2057
55. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 21–29
56. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: IEEE International Conference on Computer Vision (ICCV). (2015) 4507–4515
57. Yu, G., Yuan, J.: Fast action proposals for human action detection and search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
58. Yu, Y., Ko, H., Choi, J., Kim, G.: End-to-end concept word detection for video captioning, retrieval, and question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 3261–3269
59. Yuan, J., Ni, B., Yang, X., Kassim, A.A.: Temporal action localization with pyramid of score distribution features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
60. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: IEEE International Conference on Computer Vision (ICCV). (2017)
61. Zhu, H., Vial, R., Lu, S.: Tornado: A spatio-temporal convolutional regression network for video action proposal. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 5813–5821