WSLLN: Weakly Supervised Natural Language Localization Networks

Mingfei Gao[1] * **Larry S. Davis**[1] **Richard Socher**[2] **Caiming Xiong**[2] †
[1]University of Maryland [2]Salesforce Research
{mgao,lsd}@umiacs.umd.edu, {rsocher,cxiong}@salesforce.com

Abstract

We propose weakly supervised language localization networks (WSLLN) to detect events in long, untrimmed videos given language queries. To learn the correspondence between visual segments and texts, most previous methods require temporal coordinates (start and end times) of events for training, which leads to high costs of annotation. WSLLN relieves the annotation burden by training with only video-sentence pairs without accessing to temporal locations of events. With a simple end-to-end structure, WSLLN measures segment-text consistency and conducts segment selection (conditioned on the text) simultaneously. Results from both are merged and optimized as a video-sentence matching problem. Experiments on *ActivityNet Captions* and *DiDeMo* demonstrate that WSLLN achieves state-of-the-art performance.

1 Introduction

Extensive work has been done on temporal action/activity localization (Shou et al., 2016; Zhao et al., 2017; Dai et al., 2017; Buch et al., 2017; Gao et al., 2017c; Chao et al., 2018), where an action of interest is segmented from long, untrimmed videos. These methods only identify actions from a pre-defined set of categories, which limits their application to situations where only unconstrained language descriptions are available. This more general problem is referred to as natural language localization (NLL) (Hendricks et al., 2017; Gao et al., 2017a). The goal is to retrieve a temporal segment from an untrimmed video based on an arbitrary text query. Recent work focuses on learning the mapping from visual segments to the input text (Hendricks et al., 2017; Gao et al., 2017a; Liu et al., 2018; Hendricks et al., 2018; Zhang et al., 2018) and retrieving segments based on the alignment scores. However, in order to successfully train a NLL model, a large number of diverse language descriptions are needed to describe different temporal segments of videos which incurs high human labeling cost.

We propose Weakly Supervised Language Localization Networks (WSLLN) which requires only video-sentence pairs during training with no information of where the activities temporally occur. Intuitively, it is much easier to annotate video-level descriptions than segment-level descriptions. Moreover, when combined with text-based video retrieval techniques, video-sentence pairs may be obtained with minimum human intervention. The proposed model is simple and clean, and can be trained end-to-end in a single stage. We validate our model on *ActivityNet Captions* and *DiDeMo*. The results show that our model achieves the state-of-the-art of the weakly supervised approach and has comparable performance as some supervised approaches.

2 Related Work

Temporal Action Localization in long videos is widely studied in both offline and online scenarios. In the offline setting, temporal action detectors (Shou et al., 2016; Buch et al., 2017; Gao et al., 2017c; Chao et al., 2018) predict the start and end times of actions after observing the whole video, while online approaches (De Geest et al., 2016; Gao et al., 2017b; Shou et al., 2018b; Xu et al., 2018; Gao et al., 2019) label action class in a per-frame manner without accessing future information. The goal of temporal action detectors is to localize actions in pre-defined categories. However, activities in the wild is very complicated and it is challenging to cover all the activities of interest by using a finite set of categories.

*Work done when the author was at Salesforce Research.
†Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1481–1487, Hong Kong, China, November 3–7, 2019.

Natural Language Localization in untrimmed videos was first introduced in (Gao et al., 2017a; Hendricks et al., 2017), where given an arbitrary text query, the methods attempt to localize the text (predict its start and end times) in a video. Hendricks *et al.* proposed MCN (Hendricks et al., 2017) which embeds the features of visual proposals and sentence representations in the same space and ranks proposals according their similarity with the sentence. Gao *et al.* proposed CTRL (Gao et al., 2017a), where alignment and regression are conducted for clip candidates. Liu *et al.* introduced TMN (Liu et al., 2018) which measures the clip-sentence alignment guided by the semantic structure of the text query. Later, Hendricks *et al.* proposed MLLC (Hendricks et al., 2018) that explicitly reasons about temporal clips of a video. Zhang *et al.* proposed MAN (Zhang et al., 2018) which utilizes Graph Convolutional Networks (Kipf and Welling, 2016) to model temporal relations among visual clips. Although these methods achieve considerable success, they need segment-level annotations for training. Duan *et al.* proposed WSDEC to handle weakly supervised dense event captioning in (Duan et al., 2018) by alternating between language localization and caption generation iteratively. WSDEC generates language localization as intermediate results and can be trained using video-level labels. Thus, we set it as a baseline, although it is not designed for NLL.

Weakly Supervised Localization has been studied extensively to use weak supervisions for object detection on images and action localization in videos (Oquab et al., 2015; Bilen and Vedaldi, 2016; Tang et al., 2017; Gao et al., 2018; Kantorov et al., 2016; Li et al., 2016; Jie et al., 2017; Diba et al., 2017; Papadopoulos et al., 2017; Duchenne et al., 2009; Laptev et al., 2008; Bojanowski et al., 2014; Huang et al., 2016; Wang et al., 2017; Shou et al., 2018a). Some methods use class labels to train object detectors. Oquab *et al.* discussed that object locations may be freely obtained when training classification models (Oquab et al., 2015). Bilen *et al.* proposed WSDDN (Bilen and Vedaldi, 2016), which focuses on both object recognition and localization. Their proposed two-stream architecture inspired several weakly supervised approaches (Tang et al., 2017; Gao et al., 2018; Wang et al., 2017) including our method. Li *et al.* presented an adaptation strategy in (Li et al., 2016) which uses the output of a weak detector as pseudo groundtruth to train a detector in a fully supervised way. OICR (Tang et al., 2017) integrates multiple instance learning and iterative classifer refinement in a single network. Some works use other types of weak supervisions to optimize detectors. In (Papadopoulos et al., 2017), Papadopoulos *et al.* used clicks to train detectors. Gao *et al.* utilized object counts for weakly supervised object detection (Gao et al., 2018). Instead of using temporally labeled segments, weakly supervised action detectors use weaker annotations, *e.g.*, movie script (Duchenne et al., 2009; Laptev et al., 2008), the order of the occurring action classes in videos (Bojanowski et al., 2014; Huang et al., 2016) and video-level class labels (Wang et al., 2017; Shou et al., 2018a).

3 Weakly Supervised Language Localization Networks (WSLLN)

3.1 Problem Statement

Following the setting of its strongly supervised counterpart (Gao et al., 2017a; Hendricks et al., 2017), the goal of a weakly supervised language localization (WSLL) method is to localize the event that is described by a sentence query in a long, untrimmed video. Formally, given a video consisting of a sequence of image frames, $\mathbf{V}_i = [I_i^1, I_i^2, ..., I_i^T]$, and a text query Q_i, the model aims to localize a temporal segment, $[I_i^{st}, ..., I_i^{ed}]$, which semantically aligns best with the query. st and ed indicate the start and end times, respectively. The difference is that WSLL methods only utilize video-sentence pairs, $\{\mathbf{V}_i, Q_i\}_{i=1}^N$, for training, while supervised approaches have access to the start and end times of the queries.

3.2 The Proposed Approach

Taking frame sequences, $[I_i^1, I_i^2, ..., I_i^T]$, as inputs, the model first generates a set of temporal proposals, $\{p_i^1, p_i^2, ..., p_i^n\}$, where p_i^j consists of temporally-continuous image frames. Then, the method aligns the proposals with the input query and outputs scores for proposals, $\{s_i^1, s_i^2, ..., s_i^n\}$, indicating their likelihood of containing the event.

Feature Description. Given a sentence query Q_i of arbitrary length, sentence encoders can be used to extract text feature, fq_i, from the query. For a video, $\mathbf{V}_i = [I_i^1, I_i^2, ..., I_i^T]$, features, $\mathbf{fv}_i = [fv_i^1, fv_i^2, ..., fv_i^T]$, are extracted from each frame. Following (Hendricks et al., 2017), the visual fea-

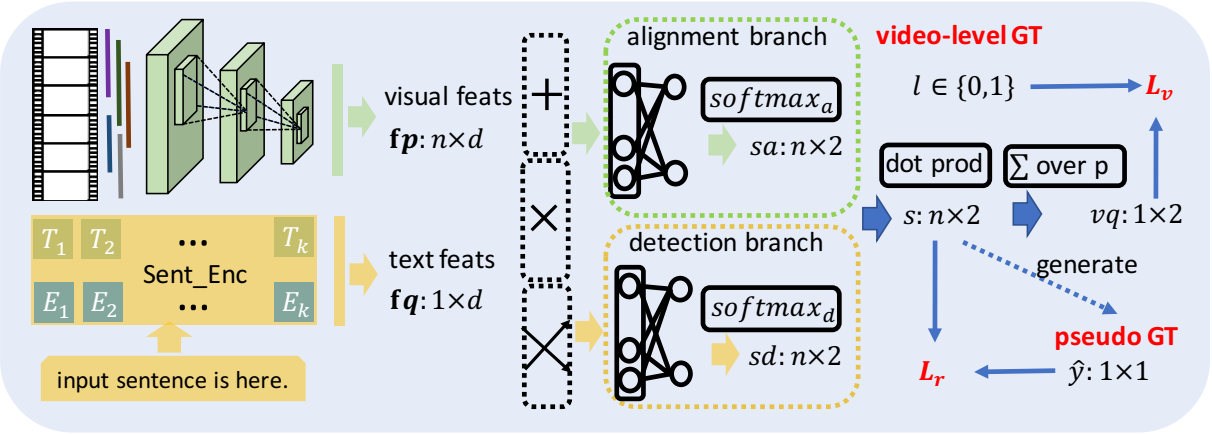

Figure 1: The workflow of our method. Visual and text features are extracted from n video proposals and the input sentence. Fully-connected (FC) layers are used to transform the features to the same length, d. The two features are combined by multi-modal processing (Gao et al., 2017a) and input to the two-branch structure. Scores from both parts are merged. Video-level scores, vq, are obtained by summing s over proposals. The whole pipeline is trained end-to-end using video-level and pseudo segment-level labels. $x \times z$ indicates dimensions.

ture, fp_i^j, of a proposal p_i^j is obtained using Eq. 1, where $pool(x, t_1, t_2)$ means average pooling features x from time t_1 to t_2, $||$ indicates concatenation, j_{st}/j_{ed} indicates start/end times of the proposal and $\bar{j}$ means time is normalized to $[0, 1]$.

$$pool(\mathbf{fv}_i, j_{st}, j_{ed})||pool(\mathbf{fv}_i, 0, T)||[\bar{j}_{st}, \bar{j}_{ed}] \quad (1)$$

We see that the feature of each proposal contains the information of its visual pattern, the overall context and its relative position in the video.

Following (Gao et al., 2017a), features of the sentence and a visual proposal are combined as in Eq. 2. The feature, fm, will be used to measure the matching between a candidate proposal and the input query.

$$fm = (fp + fq)||(fp \cdot fq)||FC(fp||fq) \quad (2)$$

The workflow of WSLLN is illustrated in Fig. 1. Inspired by the success of the two-stream structure in the weakly supervised object and action detection tasks (Bilen and Vedaldi, 2016; Wang et al., 2017), WSLLN consists of two branches, *i.e.*, alignment branch and selection branch. The semantic consistency between the input text and each visual proposal is measured in the alignment branch. The proposals are compared and selected in the detection branch. Scores from both branches are merged to produce the final results.

Alignment Branch produces the consistency scores, $sa_i \in \mathrm{R}^{n\times 2} = [sa_i^1, sa_i^2, ..., sa_i^n]$, for proposals of the video-sentence pair. sa_i in Eq. 3, measures how well each proposal matches the text. Different proposal scores are calculated independently where $softmax_a$ indicates applying the softmax function over the last dimension.

$$sa_i = softmax_a(\mathbf{W}_a fm_i) \quad (3)$$

Detection Branch performs proposal selection. The selection score, $sd_i \in \mathrm{R}^{n\times 2} = [sd_i^1, sd_i^2, ..., sd_i^n]$ in Eq. 4, is obtained by applying softmax function over proposals. Through softmax, the score of a proposal will be affected by those of other proposals, so this operation encourages competition among segments.

$$sd_i = softmax_d(\mathbf{W}_d fm_i) \quad (4)$$

Score Merging is applied to both parts to obtain the results by dot production, *i.e.*, $s_i = sa_i \cdot sd_i$, for proposals. s_i is used as the final segment-sentence matching scores during inference.

Training Phase. To utilize video-sentence pairs as supervision, our model is optimized as a video-sentence matching classifier. We compute the matching score of a given video-sentence pair by summing s_i^j over proposals, $vq_i = \sum_{j=1}^{n} s_i^j$. Then, L_v is obtained in Eq. 5 by measuring the score with the video-sentence match label $l_i \in \{0, 1\}$. Positive video-sentence pairs can be obtained directly. We generate negative ones by pairing each video with a randomly selected sentence in the training set. We ensure that the positive pairs are not included in the negative set.

$$L_v = loss(vq_i, l_i) \quad (5)$$

Results can be further refined by adding an auxiliary task L_r in Eq. 6 where $\hat{y}_i = \{0, 1, ..., n-1\}$ indicates the index of the segment that best matches the sentence during training. The real segment-level labels are not available, thus we generate pseudo labels by setting $\hat{y}_i = \text{argmax}_j s_i^j[:, 1]$. This loss further encourages competition among proposals.

$$L_r = loss(s_i^j, \hat{y}_i) \tag{6}$$

The overall objective is minimizing L in Eq. 7, where λ is a balancing scalar. $loss$ is cross-entropy loss.

$$L = loss(vq_i, l_i) + \lambda loss(s_i^j, \hat{y}_i). \tag{7}$$

4 Experiments

4.1 Experimental Settings

Implementation Details. BERT (Devlin et al., 2018) is used as the sentence encoder, where the feature of '[CLS]' at the last layer is extracted as the sentence representation. Visual and sentence features are linearly transformed to have the same dimension, $d = 1000$. The hidden layers for both branches have 256 units. For *ActivityNet Captions*, we take the $n = 15$ proposals over multiple scales of each video provided by (Duan et al., 2018) and use the C3D (Tran et al., 2015) features provided by (Krishna et al., 2017). For *DiDeMo*, we use the $n = 21$ proposals and VGG (Simonyan and Zisserman, 2014) features (RGB and Flow) provided in (Hendricks et al., 2017).

Evaluation Metrics. Following (Gao et al., 2017a; Hendricks et al., 2017), *R@k,IoU=th* and *mIoU* are used for evaluation. Proposals are ranked according to their matching scores with the input sentence. If the temporal IoU between at least one of the top-k proposals and the groundtruth is bigger or equal to th, the sentence is counted as matched. *R@k,IoU=th* means the percentage of matched sentences over the total sentences given k and th. *mIoU* is the mean IoU between the top-1 proposal and the groundtruth.

4.2 Experiments on ActivityNet Captions

Dataset Description. *ActivityNet Captions* (Krishna et al., 2017) is a large-scale dataset of human activities. It contains 20k videos including 100k video-sentences in total. We train our models on the training set and test them on the validation set.

Model	WS	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
CTRL	F	49.1	28.7	14.0	20.5
ABLR	F	73.3	55.7	36.8	37.0
WSDEC-S	F	70.0	52.9	37.6	40.4
WSDEC-W	T	62.7	42.0	23.3	28.2
WSLLN	T	75.4	42.8	22.7	32.2

Table 1: Comparison results based on $R@1$ on *ActivityNet Captions*. All baseline numbers are reprinted from (Duan et al., 2018). WS: weakly supervised.

$\lambda \rightarrow$	0.0	0.1	0.2	0.3	0.4	0.5
IoU=0.1	64.9	75.4	75.5	75.5	75.5	66.6
IoU=0.3	36.2	42.8	42.9	42.9	42.9	38.3
IoU=0.5	19.4	22.7	22.7	22.8	22.7	20.7
mIoU	27.4	32.2	32.3	32.3	32.3	28.8

Table 2: R@1 results of our method on *ActivityNet Captions* when λ in Eq. 7 is set to be different values.

Although the dataset provides segment-level annotation, we only use video-sentence pairs during training.

Baselines. We compare with strongly supervised approaches, *i.e.*, CTRL (Gao et al., 2017a), ABLR (Yuan et al., 2018) and WSDEC-S (Duan et al., 2018) to see how much accuracy it sacrifices when using only weak labels. Originally proposed for dense-captioning, WSDEC-W (Duan et al., 2018) achieves state-of-the-art performance for weakly supervised language localization. Although showing good performance, WSDEC-W involves complicated training stages, and alternates between sentence localization and caption generation for iterations.

4.2.1 Comparison Results

Comparison results are displayed in Tab. 1. It shows that WSLLN largely outperforms WSDEC-W by $\sim 4\%$ $mIoU$. When comparing with strongly supervised methods, WSLLN outperforms CTRL by over 11% $mIoU$. Using the $R@1, IoU = 0.1$ metric, our model largely outperforms all the baselines including strongly and weakly supervised methods which means that when a scenario is flexible with the IoU coverage, our method has great advantage over others. When th =0.3/0.5, our model has comparable results as WSDEC-W and largely outperforms CTRL. The overall results demonstrate good performance of WSLLN, even though there is still a big gap between weakly supervised methods and

some supervised ones, *i.e.*, ABLR and WSDEC-S. $mIoU$ (mean±std) of WSLLN across 3 runs is 32.2 ± 0.05 which demonstrates the robustness of our method.

4.2.2 Ablation Study

Effect of λ. We evaluate the effect of λ (see Eq. 7) in Tab. 2. As it shows, our model performs stable when λ is set from 0.1 to 0.4. When $\lambda = 0$, the refining module is disabled and the performance drops. When λ is set to a big number, *e.g.*, 0.5, the contribution of L_v is reduced and the model performance also drops.

Effect of Sentence Encoder. WSDEC-W uses GRU (Cho et al., 2014) as its sentence encoder, while our method uses BERT. It seems an unfair comparison, since BERT is powerful than GRU in general. However, we uses pretrained BERT model without fine tuning on our dataset, while WSDEC-W uses GRU but performed an end-to-end training. So, it is unclear which setting is better. To resolve this concern, we replace our BERT with GRU following WSDEC-W. The $R@1$ results when IoU is set to be 0.1, 0.3 and 0.5 are 74.0, 42.3 and 22.5, respectively. The mIoU is 31.8. It shows that our model with GRU has comparable results as that with BERT.

Effect of Two-branch Design. We create two baselines, *ie*, *Align-only* and *Detect-only*, to demonstrate the effectiveness of our design. To perform fair comparison, both of them are trained using only video-sentence pairs.

Align-only contains only the alignment branch. For positive video sentence pair, we give positive labels to all proposals. Negative pairs have negative labels for all the proposals. Loss is calculated between proposal scores and the generated segment-level labels.

Detect-only contains only the detection branch. Loss is calculated using the highest detection score over proposals and the video-level label at each training iteration.

Comparison results are displayed in Tab. 3. It shows that the two baselines underperform WSLLN by a large margin, which demonstrates the effectiveness of our design.

4.3 Experiments on DiDeMo

Dataset Description. *DiDeMo* was proposed in (Hendricks et al., 2017) for the language localization task. It contains 10k, 30-second videos including 40k annotated segment-sentence pairs.

Model	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
Align-only	40.0	18.9	7.5	13.4
Detect-only	33.7	18.3	10.4	13.6

Table 3: Ablation study based on $R@1$ on *ActivityNet Captions*. Both methods are trained using weak supervisions.

Model	WS	Input	R@1	R@5	mIoU
Chance	–	–	3.75	22.50	22.64
LOR	F	RGB	16.2	43.9	27.2
MCN	F	RGB	23.1	73.4	35.5
MCN	F	Flow	25.8	75.4	38.9
WSLLN	T	RGB	19.4	53.1	25.4
WSLLN	T	Flow	18.4	54.4	27.4

Table 4: Comparison results on *DiDeMo*. Following MCN, we set $th = 1.0$ for the IoU threshold. All baseline numbers are reprinted from (Hendricks et al., 2017). WS: weakly supervised.

Our models are trained using video-sentence pairs in the train set and tested on the test set.

Baselines. To the best of our knowledge, no weakly supervised method has been evaluated on *DiDeMo*. So, we compare with some supervised methods, *i.e.*, MCN (Hendricks et al., 2017) and LOR (Hu et al., 2016). MCN is a supervised NLL model. LOR is a supervised language-object retrieval model. It utilizes much more expensive (object-level) annotations for training. We follow the same setup of LOR as in (Hendricks et al., 2017) to evaluate LOR for our task.

Comparison Results are shown in Tab. 4. WSLLN performs better than LOR in terms of $R@1/5$. We also observe that the gap between our method and the supervised NLL model is much larger on *DiDeMo* than on *ActivityNet Captions*. This may be due to the fact that *DiDeMo* is a much smaller dataset which is a disadvantage for weakly supervised learning.

5 Conclusion

We propose WSLLN– a simple language localization network. Unlike most existing methods which require segment-level supervision, our method is optimized using video-sentence pairs. WSLLN is based on a two-branch architecture where one branch performs segment-sentence alignment and the other one conducts segment selection. Experiments show that WSLLN achieves promising results on *ActivityNet Captions* and *DiDeMo*.

References

Hakan Bilen and Andrea Vedaldi. 2016. Weakly supervised deep detection networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. 2014. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer.

Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. 2017. SST: Single-stream temporal action proposals. In *CVPR*.

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

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*.

Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. 2017. Temporal context network for activity localization in videos. In *ICCV*.

Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online action detection. In *ECCV*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. 2017. Weakly supervised cascaded convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5139.

Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3059–3069.

Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis R Bach, and Jean Ponce. 2009. Automatic annotation of human actions in video. In *ICCV*, volume 1, pages 3–2.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017a. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275.

Jiyang Gao, Zhenheng Yang, and Ram Nevatia. 2017b. RED: Reinforced encoder-decoder networks for action anticipation. In *BMVC*.

Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. 2017c. TURN TAP: Temporal unit regression network for temporal action proposals. *ICCV*.

Mingfei Gao, Ang Li, Ruichi Yu, Vlad I Morariu, and Larry S Davis. 2018. C-wsl: Count-guided weakly supervised localization. In *ECCV*.

Mingfei Gao, Mingze Xu, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. Startnet: Online detection of action start in untrimmed videos. *arXiv preprint arXiv:1903.09868*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.

De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2016. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer.

Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. 2017. Deep self-taught learning for weakly supervised object localization. *IEEE CVPR*.

Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. 2016. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *CVPR*.

Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. 2016. Weakly supervised object localization with progressive domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694.

Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Training object class detectors with click supervision. *CVPR*.

Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. 2018a. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171.

Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i Nieto, and Shih-Fu Chang. 2018b. Online action detection in untrimmed, streaming videos-modeling and evaluation. In *ECCV*.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*.

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

Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple instance detection network with online instance classifier refinement. *CVPR*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4325–4334.

Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. 2018. Temporal recurrent networks for online action detection. *arXiv:1811.07391*.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2018. To find where you talk: Temporal sentence localization in video with attention based location regression. *arXiv preprint arXiv:1804.07014*.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2018. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. *arXiv preprint arXiv:1812.00087*.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *ICCV*.