This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Not All Frames Are Equal: Weakly-Supervised Video Grounding with Contextual Similarity and Visual Clustering Losses

Jing Shi[†] Jia Xu[‡] Boqing Gong[‡] Chenliang Xu[†]
[†]University of Rochester [‡]Tencent AI Lab
[†]{j.shi,chenliang.xu}@rochester.edu [‡]xujianjucs@gmail.com [‡]boqinggo@outlook.com

Abstract

We investigate the problem of weakly-supervised video grounding, where only video-level sentences are provided. This is a challenging task, and previous Multi-Instance Learning (MIL) based image grounding methods turn to fail in the video domain. Recent work attempts to decompose the video-level MIL into frame-level MIL by applying weighted sentence-frame ranking loss over frames, but it is not robust and does not exploit the rich temporal information in videos. In this work, we address these issues by extending frame-level MIL with a false positive frame-bag constraint and modeling the visual feature consistency in the video. In specific, we design a contextual similarity between semantic and visual features to deal with sparse objects association across frames. Furthermore, we leverage temporal coherence by strengthening the clustering effect of similar features in the visual space. We conduct an extensive evaluation on YouCookII and RoboWatch datasets, and demonstrate our method significantly outperforms prior state-of-the-art methods.

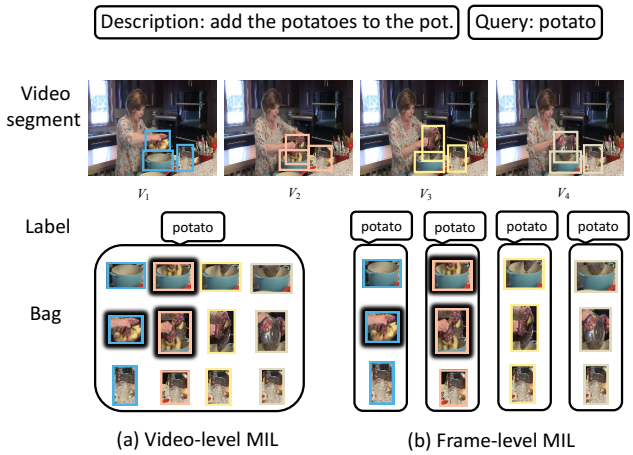


Figure 1. The illustration of (a) video-level MIL and (b) frame-level MIL. V_1 to V_4 are uniformly sampled from a video segment. Region proposals in different frames are distinguished by color. Video-level MIL puts region proposals from all frames into one bag while frame-level MIL constructs a bag for each frame. The positive instances are denoted with black shadow. Here is the dilemma: video-level MIL suffers from monotonically increased bag size w.r.t. the number of frames, while frame-level MIL may contain false positive bags such as bags for V_3, and V_4.

1. Introduction

Grounding textual signals to visual-spatial regions have various applications, e.g., robotics [3, 2], human-computer interaction [27] and image retrieval [11]. While visual grounding in static images has witnessed great progress [11, 24, 4, 34, 35], visual grounding in videos is still challenging—first, a video contains many frames, which induces the temporal visual-language alignment problem that is unique to video grounding; second, despite rich source of online videos, constructing a large-scale video dataset with grounding annotation is expensive and time-consuming. Therefore, in this paper, we aim to do weakly-supervised video grounding: localize language queries in video frames without object location annotation.

Kapathy and Fei-Fei [11] introduce a Multiple Instance Learning (MIL) based grounding method that only requires the alignment of images and sentences. It reasonably assumes that each image contains at least one region corresponding to the sentence query. If we define an image as the "bag," regions as instances in the bag and language query as the label of the bag, then the image satisfies the definition of the positive bag in MIL: a bag is positive if at least one of its instances is positive. However, directly extending MIL based grounding method from image to video easily falls into a dilemma as shown in Fig. 1. The first way is to regard each video as a bag, which contains all region proposals across frames as the instances. However, the bag size will drastically increase as video becomes longer. We call this *brute-force video-level MIL*. Another option is to construct a bag for every frame, and assign the same video label to all frame bags, but it is easy to trigger false-positive bags. This option is named as *frame-level MIL*. Zhou *et al*. [40]

try to jump out of the dilemma by choosing the frame-level MIL, but weight the loss function for each frame by measuring how "positive" each frame is. Namely, each frame loss is multiplied with a positive index which is defined by the similarity between the frame and the query. However, such method suffers from a problematic penalty term, which will indistinguishably enlarge the similarity score of both aligned and unaligned pairs and has very sensitive hyper-parameters.

To overcome the above limitations, we first compare the performance of vanilla brute-force video-level MIL and frame-level MIL and decide to follow the latter choice. Then, to better conquer the downsides of Zhou *et al*. [40], we propose a contextual similarity to measure the similarity score between the frame and the language query based on two intuitions:

1. *If a sentence contains multiple queries, then each query should focus on its most relevant frames.*

2. *If an object appears sparsely across frames, the no-object frames should be insignificant compared with the frames where the object appears.*

In the case of MIL, the contextual similarity can be viewed as an augmented similarity by considering the possibility of a frame to be the true positive bag of a query. Moreover, such possibility for one frame is calculated by looking at the other frames in the same video, which makes it more reliable. By replacing ordinary frame-sentence similarity with our contextual similarity, one can alleviate the difficulty of false positive bags in frame-level MIL.

Furthermore, the aforementioned methods fail to consider the visual consistency in the video, which is a unique property to video grounding; hence, we propose visual clustering to leverage the temporal information better. Visual clustering is inspired by the idea:

3. *If two regions have high similarity to a common query, then they should also be similar to each other.*

In this case, the visual similarity is not restricted to the adjacent frames, but can also work with sparsely sampled frames in a video segment.

We conduct extensive experiments on YouCookII dataset [41], which is the largest unconstrained instructional video dataset available for visual grounding. Experimental results demonstrate the effectiveness of our proposed techniques compared to other state-of-the-art methods. Furthermore, we show that our techniques can also lead to improved performance on RoboWatch dataset [26].

The rest of this paper is organized as follows. We review related work in visual grounding, weakly-supervised object localization and feature embedding in Sec. 2. We present formal description of contextual similarity and visual clustering in Sec. 3. Experimental settings and evaluation results are presented in Sec. 4. Finally, the paper is concluded in Sec. 5.

2. Related Work

Visual grounding. Supervised image grounding has been successfully explored in [21, 20, 37]; however, the task requires expensive labels for box location. Recently, weakly-supervised image grounding draws much attention from the community. Most weakly-supervised grounding methods can be classified as either proposal-based [11, 24, 4] or proposal-free [34, 35]. Given region proposals, Karpathy and Fei-Fei [11] formulated it as a ranking problem to rank the proposals according to visual-semantic similarity scores in a MIL fashion. Rohrbach *et al*. [24] encoded a phrase as its most similar region to reconstruct the region back to the phrase. Chen *et al*. [4] transferred the knowledge from the off-the-shelf object detector to help phrase grounding. For proposal-free methods, the region location is often obtained from phrase-salient map via subwindow search. Xiao *et al*. [34] generated the salient map by regarding language structure as additional supervision for the location relationship among objects. Raymond *et al*. [35] conducted hypothesis tests over the existence of image concept given words in a statistic view.

Weakly-supervised grounding has also been attempted in videos [36, 10, 40]. Yu and Siskind [36] grounded sentence to object in constraint lab-recorded videos. Huang *et al*. [10] addressed language reference and grounding together to enhance the grounding performance with the inspiration of graphical structure modeling [12, 38, 30] Zhou *et al*. [40] extended [11] to the video domain with frame-wise weighting and achieved the best performance so far on video visual grounding. In this work, we follow the proposal-based MIL methods [11, 40] due to the simplicity and effectiveness of the MIL learning framework.

Weakly-supervised object localization. Methods, e.g., [8, 6, 7, 18, 29], are related to visual grounding, but they typically localize an predefined object class or a video tag, while, in visual grounding, the target can be any words or phrases that are loosely defined. Most weakly-supervised object localization problem can be formulated as a MIL problem as well. The image that contains the label is regarded as positive instance and otherwise not. Among the methods, [15, 22] have studied the weakly-supervised video localization. Kwak *et al*. [15] combined object discovery and object tracking while Prest *et al*. [22] extracted candidate spatio-temporal tubes for a better localization. Comparing to these methods, we propose an easier way to employ temporal information on the feature level that does not require tracking or forming tubes, which are often computationally expensive.

Feature embedding. In metric learning, contrastive loss [9, 31] and triplet loss [25] are widely used to enhance the feature space with a clustering property. When it comes to the cross-model embedding, they are still feasible with the elements forming pairs and triplets coming from different modalities (e.g., language and image [11]). However, Collell and Moens [5] showed that the projection of the source modality does not resemble the target modality, in the sense of neighborhood topology, which drives researchers to develop more discriminative mappings. One way is to reduce the intra-class feature variations using center loss [33] which has been used in tasks such as face verification [17], and object retrieval [39]. However, center loss typically needs supervision and cannot fit into our task. Other methods such as the structure-preserving loss [32] would introduce extra hyper-parameters due to the margin and the neighborhood. Different from the above works, we employ temporal visual consistency as an additional cue to reduce intra-class feature variations.

3. Methodology

3.1. Problem Formulation

Given a video segment and its sentence description, we would like to locate each query in the sentence to each frame of the video, where the query can be either a word or a phrase. Formally, we denote a video segment as a set of T frames $\mathcal{V} = \{V_t\}_{t=1}^{T}$, and each frame V_t contains a set of N region proposals $\{v_n^t\}_{n=1}^{N}$, where the superscript t indexes the frames and the subscript n indexes the proposals on the current frame. We denote a sentence as a set of K queries $\mathcal{Q} = \{q_k\}_{k=1}^{K}$, and each q_k corresponds to one or more words in the sentence. Here, the visual feature and query feature are all encoded into a common d-dimensional space such that $v_n^t, q_k \in \mathbb{R}^d$.

Following [11] and [40], we define the similarity between the query q_k and the region v_n^t as:

$$a_k^{t,n} = q_k^{\mathbf{T}} v_n^t \ , \tag{1}$$

where $\mathbf{T}$ denotes transpose. We define the negative samples $\mathcal{Q}'$ and $\mathcal{V}'$ as queries and region proposals that are neither paired with $\mathcal{Q}$ nor $\mathcal{V}$. Next, we introduce two approaches for visual grounding: the brute-force video-level MIL and the frame-level MIL. Our final model builds upon the latter.

Brute-force video-level MIL. Brute-force video-level MIL regards a video as a bag and all regions across frames in the video as the instances in the bag, and then be trained with ranking loss on the bag level. Hence the similarity score between the video segment $\mathcal{V}$ and the description $\mathcal{Q}$ is written as:

$$S(\mathcal{V}, \mathcal{Q}) = \frac{1}{K} \sum_{k=1}^{K} \max_{t,n} a_k^{t,n} \ , \tag{2}$$

and the ranking loss with margin Δ is defined as:

$$\begin{aligned} L_{rank} = &\max(0, S(\mathcal{V}, \mathcal{Q}') - S(\mathcal{V}, \mathcal{Q}) + \Delta) + \\ &\max(0, S(\mathcal{V}', \mathcal{Q}) - S(\mathcal{V}, \mathcal{Q}) + \Delta) \ . \end{aligned} \tag{3}$$

Intuitively, Eq. (2) transforms the region-query similarity to video-sentence similarity, where max is the key operation in MIL to select the most positive instance from the positive bag, which can be paraphrased as to select the region from the video with the highest similarity to the query. Then the loss is constructed as a pair-wise ranking loss to embed the aligned video-sentence pairs with higher similarity than the unaligned pairs. However, such method has a fatal drawback—the bag size will monotonically increase as the number of frames in video increases. Nonetheless, we still compare it with our model in Sec. 4.2.

Frame-level MIL. The frame-level MIL is an alternative approach to the brute-force video-level MIL. Frame-level MIL regards a frame as a bag and all regions in the frame as the instances in the bag, and then be trained with ranking loss on the frame level. Here, we define the similarity score between sentence and frame as:

$$S(V_t, \mathcal{Q}) = \frac{1}{K} \sum_{k=1}^{K} \max_{n} a_k^{t,n}, \tag{4}$$

and the ranking loss on each frame with margin Δ is:

$$\begin{aligned} L_{rank}^t = &\max(0, S(V_t, \mathcal{Q}') - S(V_t, \mathcal{Q}) + \Delta) + \\ &\max(0, S(V_t', \mathcal{Q}) - S(V_t, \mathcal{Q}) + \Delta) \ . \end{aligned} \tag{5}$$

Therefore, the final ranking loss averages over all frames:

$$L_{rank} = \frac{1}{T} \sum_{t=1}^{T} L_{rank}^t \ . \tag{6}$$

Intuitively, frame-level MIL allows the queries to find their most similar regions in each frame to represent the similarity score. While this method has fixed bag size, it assumes that all frames in a video segment are positive bags. This assumption breaks when the queried object sparsely appears across frames and would trigger false positive bags, as shown in Fig. 1. We follow this framework because it makes use of more positive instances in a video segment; this can potentially increase the training samples and is more flexible. Next, we show how to alleviate these drawbacks of a vanilla frame-level MIL.

3.2. Contextual Similarity

We alleviate the false positive frame bag problem by creating a contextual similarity between frame and query; its high-level illustration is shown in Fig. 2. In the perspective of MIL, the contextual similarity can be viewed as a better similarity augmented by considering the possibility of a

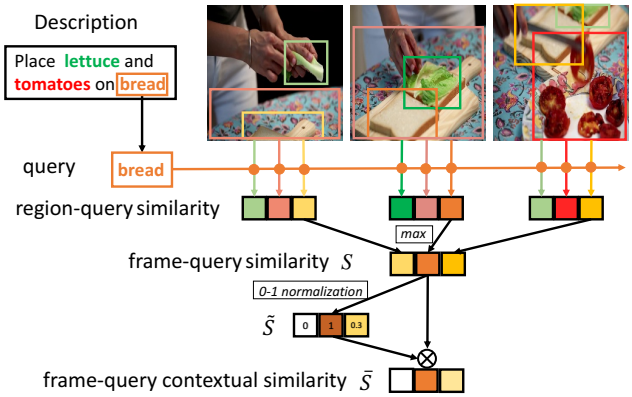


Figure 2. Diagram for evaluating the frame-query contextual similarity score. Region-query similarity scores are calculated between each region and the query "bread" with inner product similarity, which is best viewed in color. A 0-1 normalization along frames is applied to obtain its contextual gain.

frame to be the true positive bag of a query. Furthermore, such possibility for one frame is calculated by looking at the other frames in the same video making it more reliable. Concretely, we start by defining the original similarity between the frame V_t and the query q_k as:

$$S(V_t, q_k) = \max_n a_k^{t,n} \quad , \tag{7}$$

then the contextual similarity between frame and query is defined as:

$$\bar{S}(V_t, q_k) = S(V_t, q_k)\tilde{S}(V_t, q_k) \quad , \tag{8}$$

where $\tilde{S}(V_t, q_k)$ is expanded as:

$$\tilde{S}(V_t, q_k) = \frac{S(V_t, q_k) - \min_t S(V_t, q_k)}{\max_t S(V_t, q_k) - \min_t S(V_t, q_k)} \quad . \tag{9}$$

In fact, $\tilde{S}(V_t, q_k)$ is a 0-1 normalization of the original frame-query similarity over all frames in a video segment, but plays an important role as a weighting score over frames so as to guide the q_k to match its most correlated frames. Multiplying such weighting score to the original frame-query similarity yields the contextual score. Then, by averaging the contextual frame-query scores over all queries in a sentence, we obtain the sentence-frame score as:

$$S(V_t, \mathcal{Q}) = \frac{1}{K}\sum_{k=1}^{K} \bar{S}(V_t, q_k) \quad . \tag{10}$$

Next, we put Eq. (10) into Eq. (5) to get L_{rank}^t, and the video-level ranking loss is the same as Eq. (6).

The reason to design Eq. (9) is that the $\tilde{S}(V_t, q_k)$ guarantees the validity of the key frame with the highest frame-query score in the video segment, because it corresponds to $\tilde{S}(V_t, q_k) = 1$. And, it can directly abandon the trivial frame which has the lowest frame-query score since its $\tilde{S}(V_t, q_k) = 0$. Hence, we decay the importance of each frames by their relative importance to the key frame and the trivial frame. Furthermore, our formulation will not introduce additional hyper-parameters and is robust in training. Also, we find that letting gradient propagate to Eq. (9) leads to better performance.

3.3. Visual Clustering

Visual grounding is intrinsically a cross-model mapping problem. We would like to map the visual and textual features to a common space. In this sense, regions grounded by the same query should be embedded as a neighbor structure in feature space and will form a cluster. The visual clustering method assumes that the queried objects show similar appearance across video frames, and their visual features are within the same cluster. If we have region class label, it is natural to use center loss [33], which directly drives objects in the same class to be close to the class center. However, in weakly-supervised setting, the class label for each object is unknown. Instead, we first let query q_k select its most similar region proposal in frame t, and we denote the selected region as:

$$\hat{v}_{t,k} = \underset{v_t^n \in \{v_t^1, \ldots, v_t^N\}}{\arg\max} \; q_k^{\mathbf{T}} v_t^n \quad . \tag{11}$$

Then we want to further cluster all the visual features $\hat{v}_{t,k}$ in different frame t together because they all belong to the common query q_k. Hence, we minimize the negative cosine similarity of any two region features belonging to the same query in a video segment, which is defined as:

$$L_{vis} = -\sum_k \sum_{t<t'} \cos(\hat{v}_{t,k}, \hat{v}_{t',k}) \quad . \tag{12}$$

The cluster hypothesis tries to make use of the temporal connectivity so as to learn a more discriminative visual embedding.

Nonetheless, Eq. (12) has an implicit assumption that the queried object is required to appear in each frame of a video segment. According to the validation set of YouCookII dataset [40], the queried objects show up in 51.32% of the total frames, and in our experiment, such assumption does not hurt the performance. In order to better relax such assumption, we weight the visual similarity by the similarity between word feature and visual feature. Therefore, the contextual visual similarity is formulated as:

$$L_{vis}^{ctx} = -\sum_k \sum_{t<t'} \cos(\hat{v}_{t,k}, \hat{v}_{t',k})\tilde{S}(V_t, q_k)\tilde{S}(V_{t'}, q_k) \quad , \tag{13}$$

where $\tilde{S}(V_t, q_k)$ is defined in Eq. (9).

Finally, the full loss function combining the contextual similarity and visual clustering is constructed as:

$$L = \sum_{t=1}^{T} L_{rank}^{t} + \lambda L_{vis}^{ctx} \ , \tag{14}$$

where λ is the weighting parameter of the two parts of the loss function. The visual clustering can contribute to a more discriminative visual feature by reducing the intra-class variance, which achieves the same effect as the center loss. Moreover, we also have tried the loss that reduces the similarity of visual features of different classes, but the performance does not improve; hence we only force the similarity between similar visual features across frames in a same video segment.

3.4. Method Details

Learning & Inference. We employ two training strategies: Finite-Class Training (FCT) and Infinite-Class Training (ICT). In FCT, only words from a small size of vocabulary set are chosen to construct the ranking loss; whereas in ICT, any noun contributes to the loss. FCT has the advantage of higher grounding accuracy on the vocabulary but sacrifices the generalizability to other datasets. On the contrary, ICT generalizes easily by compromising accuracy in finite vocabulary set. We conduct both strategies in Sec. 4.

Visual embedding. To get visual embedding v, we first extract the 4096-dimensional v_{cnn} from the last fully-connected (FC) layer of a convolutional network, then add an additional FC layer with parameter W_v and hyperbolic tangent function to encode it to a 512-dimensional common space. In other words, $v = \tanh(W_v v_{cnn})$.

Textual embedding. Each word is first embeded with 200-dimensional GloVe [19] feature s_{glv}. For FCT, each word $s = \tanh(W_s s_{glv})$, where W_w is a linear layer. While for ICT, the ith word in a sentence is formulated as $s_i = \tanh(W_s[BiLSTM(s_i)]_i$, where $BiLSTM(\cdot)$ represents a bi-directional LSTM. If the query is a phrase, simply average the word features in the phrase, which has the same dimension as the visual feature.

Compare to other methods. We are not the first one to explore video grounding with MIL. Zhou *et al*. [40] constructed the frame-sentence similarity ranking loss with the weighted frame importance:

$$L = \frac{1}{T}\sum_{t=1}^{T}[\lambda S(V_t, \mathcal{Q}) L_{rank}^{t} + (1-\lambda)(-\log(2S(V_t, \mathcal{Q}))], \tag{15}$$

which does weighted-sum over each frame loss by the frame-sentence similarity $S(V_t, \mathcal{Q})$. Notice that it does not simply calculate L_{rank} like Eq. (6) because it tries to reduce the negative effect of false positive bags in frame-level MIL. The lower $S(V_t, \mathcal{Q})$ indicates the higher possibility that the frame V_t is a false positive bag. By multiplying this term to frame-wise ranking loss, the model down-weights the false negative bags and thus yields better result. Moreover, in order to avoid the trivial solution $S(V_t, \mathcal{Q}) = 0$, the second term in Eq. (15) is the penalty term that pulls $S(V_t, \mathcal{Q})$ to be greater than 0. While Eq. (15) tries to construct better positive sentence-frame pairs by applying a strong (weak) frame ranking loss if the frame has higher (lower) similarity to the sentence, such method has two obvious disadvantages: (1) the hyper-parameter λ is very sensitive to the model grounding accuracy; (2) the penalty term tries to penalize all the similarity of sentence-frame pairs even if the frame does not contain the queried object, which is not reasonable.

4. Experiments

4.1. Datasets and Evaluation Metric

We train our model in a weakly-supervised manner on YouCookII dataset [41] and conduct generalizability analysis on RoboWatch dataset [26].

YouCookII. YouCookII [41] is a large-scale dataset including 2000 YouTube cooking videos from 89 recipes. Each video recipe consists of 3 to 15 steps, where each step is annotated with a sentence description and temporal boundaries of the corresponding video segment. For evaluation and testing, [10] and [40] contribute to the bounding box annotation independently. [10] focuses on the union of grounding and co-reference, hence it annotates roughly 5 frames per object in a video segment with the reference of previous step with phrase. [40] aims at general video object grounding and thus annotates the boxes at 1 fps with 67 kinds of objects in vocabulary. We conduct experiments on YouCookII dataset following [40], which is more similar to our work.

RoboWatch. The test set of RoboWatch [26] contains 225 YouTube videos mainly about cooking. Similar to YouCookII, these videos are annotated with temporal intervals and description for each step. [10] extends the bounding box annotation for a part of those videos, and the query can be either word or phrase. One important difference of [10] compared with this paper is that [10] is a reference-aware grounding method which can ground a query to its unaligned video segment referred by such query, while our paper focuses more on the MIL strategy within a single video segment. Hence we only evaluate our model on the aligned video segment and query pairs in RoboWatch.

Evaluation metric. We evaluate the models using both Box accuracy [40] and Query accuracy [10]. For each query, we propose its top-1 grounded box. The box accuracy is defined as the ratio of correctly grounded boxes to all grounded boxes, where the correctly grounded boxes have more than 50% Intersection-over-Union (IoU) with ground-truth boxes. Query accuracy is defined as the ratio of cor-

Method	Box accuracy (%)				Query accuracy (%)			
	macro		micro		macro		micro	
	val	test	val	test	val	test	val	test
Upper Bound	62.42	62.41	-	-	-	-	-	-
Compared method								
GroundR [24]	19.63	19.94	-	-	-	-	-	-
DVSA_{frm}*[11]	36.90	37.55	44.26	44.16	38.48	39.31	46.27	46.14
DVSA_{vid}*[11]	36.67	36.30	43.62	42.87	38.20	37.98	45.60	44.79
Zhou *et al*. [40]	30.31	31.73	-	-	-	-	-	-
Zhou *et al*. *[40]	35.69	35.08	43.04	42.42	37.26	36.69	44.99	44.34
Our method								
VisClus	37.80	38.04	45.35	45.53	39.44	39.72	47.41	47.58
CtxSim	38.12	38.78	46.10	45.74	39.78	40.45	48.20	47.80
VisClus+CtxSim	**39.54**	**40.71**	**46.41**	**46.33**	**41.29**	**42.45**	**48.52**	**48.41**

Table 1. Weakly-supervised grounding results on YouCookII in FCT. The compared methods implemented by us are indicated with *.

rectly grounded queries to all queries, and a grounded query is correct if it is matched with correctly grounded box. Also, we denote macro-accuracy as the average of each class accuracy and denote micro-accuracy as the global accuracy without distinction of classes.

Implementation details. The description sentence is parsed by Stanford CoreNLP parser [16] into nouns. For each video segment, 5 frames are uniformly sampled and then fed into RPN [23] with VGG-Net [28] backbone pre-trained on [14] to get top-20 region proposals. The number of sampled frames and region proposals are set following [10] and [41]. We use Adam [13] with learning rate 0.01 for optimization, and dropout rate 0.1 for regularization. The hyper-parameters are searched by Bayesian optimization [1] as $\lambda = 4.13$ and $\Delta = 10$. At training phase, each batch contains 8 aligned video-sentence pairs and can form totally 64 pairs for ranking loss.

Grounding approaches. We compare the following models and variants of our model for visual grounding:

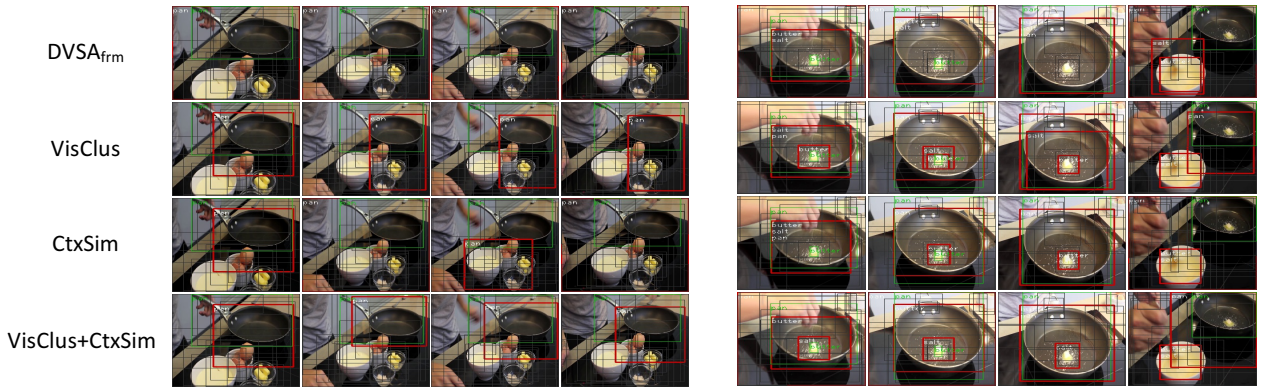
- *Deep Visual-Semantic Alignment (DVSA)* [11]. DVSA is the grounding by ranking method upon which we build our models. For a fair comparison, the image-based DVSA has been adapted to videos in both frame-level MIL (DVSA_{frm}) as in Eq. (3) and video-level MIL (DVSA_{vid}) as in Eq. (6).
- Zhou *et al*. [40]. This approach weights the frame loss by using Eq. (15) and is test on a limited word vocabulary. we re-implemented this method to draw a fair comparison.
- *RA-MIL* [10]. We compare this method for testing the generalizability of our model.
- *CtxSim*. Our model variant that uses only the contextual similarity loss defined in Sec. 3.2.
- *VisClus*. Our model variant that uses only the visual clustering loss in Sec. 3.3.
- *Upper Bound*. This is calculated by regarding all 20 proposed boxes as the grounded boxes of each query, rather than the top-1 box.

4.2. Main Results

We conduct FCT on YouCookII dataset for fair comparison with Zhou *et al*. [40], because the ground-truth object belongs to a finite set of words. And, for comparison with RA-MIL on generalization test on RoboWatch, ICT is employed on YouCookII, due to the ground-truth query is a word or a phrase without constraints. Quantitative results on YouCookII in FCT mode are shown in Table 1. We report macro-accuracy and micro-accuracy on both box and query accuracy.

Frame-level MIL and video-level MIL. We report both video-level and frame-level MIL extensions of DVSA and show that the frame-level MIL outperforms video-level MIL. To further analyze the reason, we go through the validation set of YouCookII and find that the queried objects show up in 51.32% of the total frames, suggesting that half of the frame-level MIL bags are false positive. On the other side, frame-level MIL has intrinsically more bags than video-level MIL, which is equivalent to say frame-level MIL have more training data. Experimental results show that even half of the positive bags are false positive, frame-level MIL still outweighs its video counterpart due to more training samples.

Contextual similarity and visual clustering. Both contextual similarity and visual clustering outperform DVSA and Zhou *et al*.'s method. Experimental results show that contextual similarity has larger improvement compared with visual clustering. We suspect that it is because visual clustering relies more heavily on the occurrence of objects in frames. Our full model outperforms the individual models in all the metrics, and higher Box macro-accuracy than DVSA_{frm}, i.e., 3.16%, which demonstrates that visual clustering and contextual similarity are mutually beneficial. Notice that Zhou *et al*. [40]'s implementation has a lower accuracy than ours, which can be partially attributed to our



(a) Put a **pan** on medium to high heat (b) Heat the **butter** and some sea **salt** flakes in the **pan**

Figure 3. Visualization of grounding results from frame-level DVSA and our proposed methods on YouCookII. Bold words are queries. Red, green and grey boxes represent model prediction, ground-truth and region proposals, respectively.

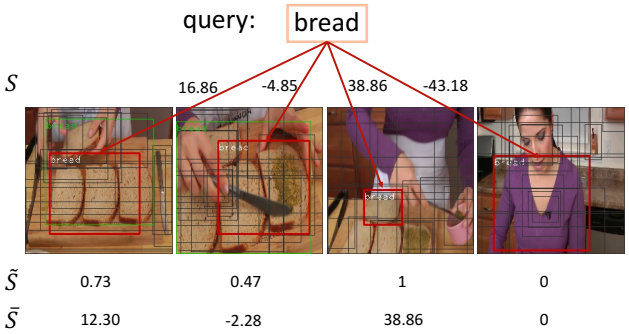
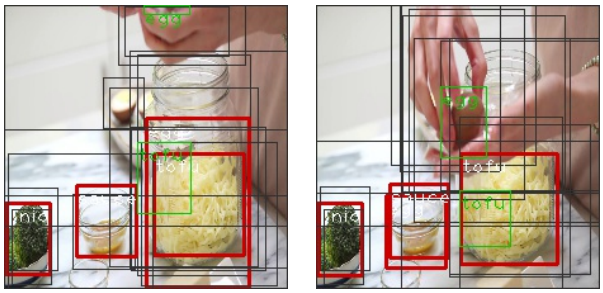


Figure 4. Example to show how the contextual similarity works with numerical demos. The grounded boxes are localized by our full model and box color is defined the same as Fig. 2.



Wrap **sauce egges tofu** and green **onions** and place in jar.

Figure 5. Failure cases. The grounded box are localized by our full model and box color is defined the same as Fig. 3.

higher upper bound then theirs. Our RPN is pretrained on Visual Gnome [14], which has richer visual semantic and enables RPN to generate better proposals. In addition, our implementation of Zhou *et al*. [40] is lower than $\mathrm{DVSA}_{\mathrm{frm}}$ baseline, because of the sensitivity of the hyper-parameter in its loss function rendering the difficulty for parameter tuning.

Analysis. We analyze the reason why our proposed methods work with the help of qualitative results. Fig. 3 visualizes two sequences of video frames from YouCookII Dataset. As expected, in both sequences, our proposed methods ground better than DVSA basline and our full model looks better than the individual ones. We observe that visual clustering performs better when localizing temporally consistent object. For example, in video segment (a) in Fig. 3, visual clustering and full model capture all the pans across frames while contextual similarity missed half of them. This observation indicates that visual clustering will push the model learn a more discriminative visual feature embedding. Also, the small objects such as butter in video segment (b), which cannot be grounded by DVSA, are correctly matched in our proposed method. This is another evidence that our model has higher recognition ability.

Contextual similarity qualitative analysis. Contextual loss attaches the normalized weights to different frames, which is experimentally proved effective. For instance, in Fig. 4, the query "bread" is not a positive label for the last frame, which will mislead the model by feeding the model with wrong samples. Fortunately, with the help of contextual loss, the normalization $\tilde{S}$ assigns the false positive frame bag with lower importance, with 0 in extreme.

Failure cases. Figure 5 presents the common failure cases: the object occlusion, object out of scene, and small size object. For example, the egg is out of scene in the left figure and occluded by hand in the right figure, which are quite often in cooking, e.g., the camera is set statically and human-object interaction can easily deviate from the screen. Also, ingredients or foods can be easily occluded by hand due to manual operation over them. In addition, objects become smaller while camera zooming out, which also adds difficulty to the grounding task. Overall, objects falling into

Method	YouCookII		RoboWatch
	val (%)	test (%)	test (%)
Upper Bound	62.42	62.41	-
Compared method			
$DVSA_{frm}$ [11]	35.87	37.33	28.25
RA-MIL [10]	-	-	19.80
Our method			
VisClus	36.44	37.80	28.68
CtxSim	**37.99**	37.67	31.08
VisClus+CtxSim	37.43	**38.49**	**31.68**

Table 2. Generalizability to unseen video classes (RoboWatch) in ICT. The score for YouCookII is the box macro-accuracy, for RoboWatch is the query micro-accuracy.

the three typical failure cases are usually not covered by region proposals, which is also true for the egg in Fig. 5.

4.3. Generalizability Test

To further test the generalization ability, we do ICT on YouCookII and test it on videos in RoboWatch including different recipes and other miscellaneous videos such as "How to remove gum from clothes," and "How to tie a tie." And, there is no recipe or video overlap with YouCookII.

The generalization performance is shown in Table 2 with metric of query micro-accuracy. For consistency consideration, the number reported on YouCookII is still evaluated in [40]'s box annotation, even if [10] also annotated the box in YouCookII. We observe that visual clustering and contextual similarity both show good generalizability and our full model outperforms all the other methods on the testing set of both datasets, with 3.43% higher than frame-level MIL baseline in RoboWatch, proving our method has a good generalization ability. Contextual similarity has a higher score on the validation set of YouCookII, but lower on testing set, suggesting it overfits to the validation set. Though we are not in an absolute fair comparison with [10] due to the reference-awareness, we list [10]'s result as a reference. Different from the experiment set up in [10], we have filtered out those ground-truth boxes corresponding to language queries in unpaired descriptions, which means the number of testing samples are smaller than [10]. However, we still can observe an improvement of our method by testing accuracy on RoboWatch.

The qualitative results of RoboWatch are shown in Fig. 6. We observe that the model has comparative grounding ability for the queries known by YouCookII, but for some unseen query such as "hanger," the model can still correctly ground it. Also, we find the model tend to localize hand, so we suspect that "hanger" has similar textual embedding with "hand" thus the model transfers the knowledge of hand toward hanger.

FCT and ICT. FCT and ICT are adopted respectively over all methods in Tables 1 and 2. Comparing their performance on YouCookII uncovers that the ICT is inferior to FCT according to accuracy but stronger then FCT on generalizability. The accuracy gain of FCT can be explained by a reduced complexity in feature embedding space for FCT because it only need to push the visual feature embedded to finite word feature cluster centers.

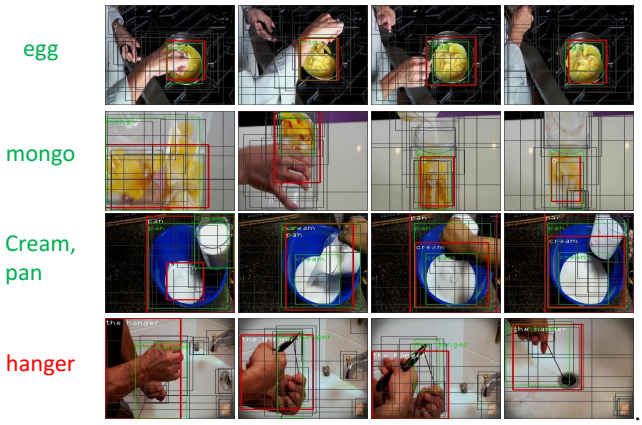


Figure 6. Visualization of grounding results with our full model on RoboWatch. The green queries have been seen in YouCookII while the red one has not. Box colors are defined in Fig. 3.

5. Conclusion

In this paper, we propose two techniques to improve the video grounding accuracy. Contextual similarity remedies the overly-strong assumption that each frame in a video segment needs to contain the grounded object. Visual clustering better exploits the temporal consistency in video and embeds a more discriminative visual feature. Experimental results on two prevalent datasets demonstrate the effectiveness and generalizability of our methods.

Limitations. As pointed out in the failure case, our model is limited by the quality of region proposals, which constrains the model's upper bound. The model's ability is also confined by the quality of the pretrained visual and textual feature encoders. With shallow learnable embedding layers, our model mainly relies on pretrained deep feature extractor.

Future work. This work tries to improve frame-wise MIL and incorporate temporal visual information extractor. However, the visual consistency is now only employed at feature level. We plan to add the visual consistency constraint at spatial level as future work.

Acknowledgement. This work was supported in part by NSF IIS 1813709, IIS 1741472, and the Tencent Rhino-Bird gift. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] Bayesian optimization. https://github.com/fmfn/BayesianOptimization.

[2] M. Al-Omari, P. Duckworth, D. C. Hogg, and A. G. Cohn. Natural language acquisition and grounding for embodied robotic systems. In *AAAI*, 2017.

[3] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In *ECCV*, 2018.

[4] K. Chen, J. Gao, and R. Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, 2018.

[5] G. Collell and M.-F. Moens. Do neural network cross-modal mappings really bridge modalities? In *ACL*, 2018.

[6] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3), 2012.

[7] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.

[8] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[10] D.-A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. Finding it: Weakly-supervised reference-aware visual grounding in instructional videos. In *CVPR*, 2018.

[11] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[12] C. Kiddon, G. T. Ponnuraj, L. Zettlemoyer, and Y. Choi. Mise en place: Unsupervised interpretation of instructional recipes. In *EMNLP*, 2015.

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

[14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1), 2017.

[15] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid. Unsupervised object discovery and tracking in video collections. In *ICCV*, 2015.

[16] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (system demonstrations)*, 2014.

[17] Z. Ming, J. Chazalon, M. Muzzamil Luqman, M. Visani, and J.-C. Burie. Simple triplet loss based on intra/inter-class metric learning for face verification. In *ICCV*, 2017.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[19] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[20] B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik. Conditional image-text embedding networks. In *ECCV*, 2018.

[21] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[22] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[24] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[26] O. Sener, A. R. Zamir, S. Savarese, and A. Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015.

[27] M. Shridhar and D. Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems*, 2018.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[29] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014.

[30] L. Song, Y. Zhang, Z. Wang, and D. Gildea. A graph-to-sequence model for amr-to-text generation. In *ACL*, 2018.

[31] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014.

[32] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[33] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[34] F. Xiao, L. Sigal, and Y. Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017.

[35] R. A. Yeh, M. N. Do, and A. G. Schwing. Unsupervised textual grounding: Linking words to image concepts. In *CVPR*, 2018.

[36] H. Yu and J. M. Siskind. Sentence directed video object codiscovery. *IJCV*, 124, 2017.

[37] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.

[38] Y. Zhang, Q. Liu, and L. Song. Sentence-state lstm for text representation. In *ACL*, 2018.

[39] X. Zheng, R. Ji, X. Sun, Y. Wu, F. Huang, and Y. Yang. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In *IJCAI*, 2018.

[40] L. Zhou, N. Louis, and J. J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*, 2018.

[41] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.