Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction

Luowei Zhou[1]
luozhou@umich.edu
Nathan Louis[2]
natlouis@umich.edu
Jason J. Corso[1,2]
jjcorso@umich.edu

[1] Robotics Institute
University of Michigan
Ann Arbor, USA

[2] Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, USA

Abstract

We study *weakly-supervised video object grounding*: given a video segment and a corresponding descriptive sentence, the goal is to localize objects that are mentioned from the sentence in the video. During training, no object bounding boxes are available, but the set of possible objects to be grounded is known beforehand. Existing approaches in the image domain use Multiple Instance Learning (MIL) to ground objects by enforcing matches between visual and semantic features. A naive extension of this approach to the video domain is to treat the entire segment as a bag of spatial object proposals. However, an object existing sparsely across multiple frames might not be detected completely since successfully spotting it from one single frame would trigger a satisfactory match. To this end, we propagate the weak supervisory signal from the segment level to frames that likely contain the target object. For frames that are unlikely to contain the target objects, we use an alternative penalty loss. We also leverage the interactions among objects as a textual guide for the grounding. We evaluate our model on the newly-collected benchmark YouCook2-BoundingBox and show improvements over competitive baselines.

1 Introduction

Grounding language in visual regions provides a fine-grained perspective towards visual recognition and has become a prominent research problem in the computer vision and natural language processing communities [6, 19, 20, 24]. In this paper, we study the problem of *video object grounding*, where a video (segment) and an associated sentence are given and the goal is to localize the objects that are mentioned in the sentence in the video. This task is often formulated as a *visual-semantic alignment* problem [7] and has broad applications including retrieval [7, 8], description generation [20, 25], and human-robot interaction [1, 22].

Like most fine-grained recognition problems [15, 18], grounding can be extremely data intensive, especially in the context of unconstrained video. On the other hand, video-sentence pairs are easier to obtain than object region annotations (*e.g.*, YouTube Automatic Speech Recognition scripts). We focus on the weakly-supervised version of the grounding problem where the only supervision is sentence descriptions; no spatially-aligned object bounding boxes are available for training. Sentence grounding can involve multiple interacting objects, which sets our work apart from the relatively well-studied weakly-supervised object localization problem, where one or more objects are localized independently [10, 17].

Existing work on visual grounding falls into two categories: multiple instance learning [6, 7] and visual attention [19]. In either case, the visual-semantic similarity is first measured between the target object/phrase and all the image-level, *i.e.* spatial, object region proposals. Then, either a ranking loss or a reconstruction loss—both of which we refer to here as matching losses—measures the quality of the matching. A naive extension of the existing approaches to the video domain is to treat the entire video segment as a bag of spatial object proposals. However, this presents two issues. First, existing methods rely on the assumption that the target object appears in *at least one* of the proposal regions. This assumption is weak when it comes to video, since a query object might appear sparsely across multiple frames[1] and might not be detected completely. The *segment-level supervision*, *i.e.* object labels, could be potentially strengthened if applied to individual frames. Second, a video segment can last up to several minutes. Even with temporal down-sampling, this can bring in tens or hundreds of frames and hence thousands of proposals, which compromise the visual-semantic alignment accuracy.

To address these two issues, we propose a frame-wise loss weighting framework for video grounding. We ground the target objects on a frame-by-frame basis. We face the challenge that the segment-level supervision is not applicable to individual frames where the query object is off-screen, occluded, or just not present in the proposals for that frame. Our solution is to first estimate the likelihood that the query object is present in (a proposal in) each video frame. If the likelihood is high, we judge the matching quality mainly on the matching loss. Otherwise, we down-weight the matching loss while bringing in a penalty loss. The lower the confidence, the higher the penalty. With the conditioned frame-wise grounding framework, the proposed model can avoid being flooded with massive proposals even when the sampling rate is high and only make predictions for applicable frames.

We propose two approaches to estimate frame-wise object likelihood (confidence) scores. The first one is conditioned on both visual and textual inputs, namely, the maximum visual-semantic similarity scores in each frame. The second approach is inspired by the fact that the combination of objects can imply their order of appearance in the video. For example, when a sequence of objects “tomatoes”, “pan” and “plate” appears in the description, the video scene is likely to include a shot of tomatoes being grilled in the pan at the beginning, and a shot of tomatoes being moved to the plate at the end. In the temporal domain, “pan” appears mostly ahead of “plate” while “tomatoes” intersects with both. We implicitly model the object interaction with self-attention [23] and use textual guidance to estimate the frame-wise object likelihood.

For evaluation, due to lack of existing video grounding benchmarks, we have collected annotations over the large-scale instructional video dataset YouCook2, which provides over 15,000 video segment-description pairs. We sample the validation and testing videos at 1fps and draw bounding box for the 67 most frequent objects when they are present in both the video segment and the description. We compare our methods against competitive baselines on video grounding and our proposed methods achieve state-of-the-art performances.

[1] In YouCook2-BoundingBox, the target object appears in 60.7% of the total frames, on average.

Our contributions are twofold: 1) we propose a novel frame-wise loss weighting framework for the video object grounding problem that outperforms competitive baselines; 2) we provide a benchmark dataset for video grounding.

2 Related Work

Grounding in Image/Video. Supervised grounding or referring has been intensively studied [15, 16, 27] in the image domain. These methods require dense bounding box annotations for training, which are expensive to obtain. Recently, an increasing amount of attention has shifted towards the weakly-supervised grounding problem [6, 7, 8, 19, 24], where only descriptive phrases, no explicit target grounding locations, are made accessible during training. Karpathy and Fei-Fei [7] propose to pair image regions to words in a sentence by computing a visual-semantic similarity score, finding the word that best describes the region. Rohrbach *et al.* [19] ground textual phrases in images by reconstructing the original phrase through visual attention. Yu and Siskind [26] ground objects from text in constrained videos. De-An *et al.* [6] extend [7] to the video domain and further improve the work by modeling the reference relationships among segments. In this work, we tackle the problem from a novel aspect as fully exploiting the visual-semantic relations within each segment, *i.e.* frame-wise supervisions and object interactions.

Weakly-supervised Object Localization. Weakly-supervised object localization has been explored in both the image [2, 4, 5, 14, 21] and the video domain [10, 17]. Unlike object grounding from text, object localization typically involves localizing an object class or a video tag in the visual content. Existing works in the image domain naturally pursue a multiple instance learning (MIL) approach to this problem. Positive instances are images where the label is present, and negative instances are given as images with the label absent. In the video domain, the existing methods [10, 17] approach this problem by taking advantage of motion information and similarity between frames to generate spatio-temporal tubes. Note that these tubes are much more expensive to obtain compared with spatial proposals, hence we only consider the latter option.

Object Interaction. Object interaction was initially proposed to detect fine-grained visual details for action detection, such as the temporal relationships between objects in a scene, to overcome changes in illumination, pose, occlusion, etc. Some works have modeled object interaction using pairwise or higher-order relationships [11, 12, 13]. Ni *et al.* [13] consolidate object detections at each step by modeling pair-wise object relationships and hence enforce the temporal object consistency in each additional step. Ma *et al.* [12] implicitly model the higher-order interactions among object region proposals, using groups and subgroups rather than just pairwise interactions. Inspired by recent work [3, 24], where the linguistic structure of the input phrase is leveraged to infer the spatial object locations, we propose to model object interaction from a linguistic perspective as a textual guidance for grounding.

3 Methods

We start this section by introducing some background knowledge. In Sec. 3.2, we describe the video object grounding baseline. We then propose our framework in Sec. 3.3 by extending the segment-level object label supervision to the frame-level. Two novel approaches are proposed in judging under what circumstances the frame-level supervision is applicable.

3.1 Background

In this section we provide some background on visual-semantic alignment framework (grounding by ranking) and self attention, which are building blocks of our model.

Grounding by Ranking. We start by describing ranking-based grounding approach from [7]. Given a sentence description including O query objects/phrases and a set of N object region proposals from an image, the goal is to target each referred object in the query as one of the object proposals. Queries and visual region proposals are first encoded in a common d-dimensional space. Denote the object query feature vectors as $\{q_k\}$, $k = 1,2,\ldots,O$ and the region proposal feature vectors as $\{r_i\}$, $i = 1,2,\ldots,N$. We pack the feature vectors into matrices $Q = (q_1,\ldots,q_O)$ and $R = (r_1,\ldots,r_N)$. The visual-semantic matching score of the description and the image is formulated as:

$$S(Q,R) = \frac{1}{O}\sum_{k=1}^{O}\max_i a_k^i, \tag{1}$$

where $a_k^i = q_k^\top r_i$ measures the similarity between query q_k and proposal r_i. Defining negative samples Q' and R' as the query and proposal from texts and images that are not paired with R nor Q, the grounding by ranking framework minimizes the following margin loss:

$$L_{rank} = \sum_{R' \neq R}\sum_{Q' \neq Q}[\max(0, S(Q,R') - S(Q,R) + \Delta) + \max(0, S(Q',R) - S(Q,R) + \Delta)], \tag{2}$$

where the first ranking term encourages the correct region proposal matching and the second ranking term encourages the correct sentence matching. Δ is the ranking margin. During inference, the proposal with the maximal similarity score a_k^i with each object query is selected.

Self Attention. We now describe the scaled dot-product attention model. Define a set of queries $q_j \in \mathbb{R}^d$, a set of keys $k_t \in \mathbb{R}^d$ and values $v_t \in \mathbb{R}^d$, where $j = 1,2,\ldots,O$ is the query index, $t = 1,2,\ldots,T$ is the key/value index. Given an arbitrary query q_k, scaled dot-product attention computes the output as a weighted sum of values v_t, where the weights are determined by the scaled dot-products of query q_j and keys k_t, as formulated below:

$$A(q_j,K,V) = \text{Softmax}(q_j^\top K/\sqrt{d})V^\top, \tag{3}$$

where the authors pack k_t and v_t into matrices $K = (k_1,\ldots,k_T)$ and $V = (v_1,\ldots,v_T)$, respectively. *Self-attention* [23] is a special case of the scaled dot-product attention where the queries, keys and values are all identical. In our case, they are all object encoding vectors and self-attention encodes the semantic relationships among the objects. We adopt a multi-head version of the self-attention layer [23, 29] for modeling object relationships, which deploys multiple paralleled self-attention layers.

3.2 Video Object Grounding

We adapt the Grounding by Ranking framework [7] to the video domain, and this adaptation will serve as our baseline. Denote the set of T frames in a video segment as $\{f_t\}$ and the object proposals in frame t as r_i^t, $i = 1,2,\ldots,N$. As before, define the object queries as q_k, we compute the similarity between the query object and all the proposals $\{r_i^t\}$ in a segment. Note that the similarity dot product might grow large in magnitude as d increases [23]. Hence, we scale the dot-product by $\frac{1}{\sqrt{d}}$ and restrict $a_k^{t,i}$ to be between 0 and 1 with a Sigmoid function.

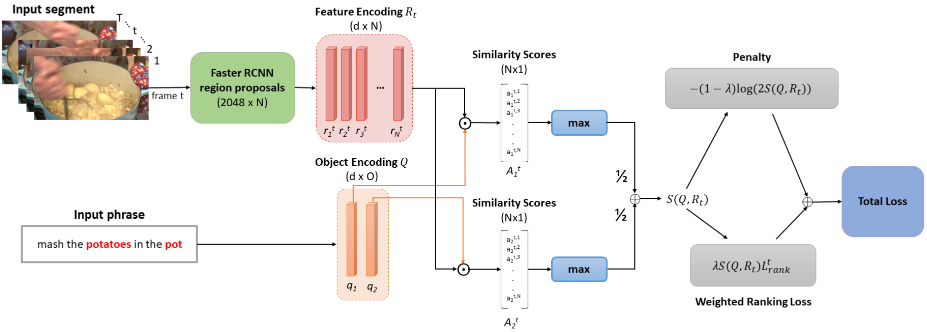

Input segment
T
t
2
1
frame t
Faster RCNN region proposals (2048 x N)
Feature Encoding R_t (d x N)
r_1^t r_2^t r_3^t ... r_N^t
Object Encoding Q (d x O)
q_1 q_2
Input phrase
mash the potatoes in the pot
Similarity Scores (Nx1)
$a_1^{t,1}$ $a_1^{t,2}$ $a_1^{t,3}$ $a_1^{t,N}$
A_1^t
Similarity Scores (Nx1)
$a_2^{t,1}$ $a_2^{t,2}$ $a_2^{t,3}$ $a_2^{t,N}$
A_2^t
max
max
½
½
$S(Q,R_t)$
Penalty
$-(1-\lambda)\log(2S(Q,R_t))$
$\lambda S(Q,R_t)L_{rank}^t$
Weighted Ranking Loss
Total Loss

Figure 1: An overview of our framework. Inputs to the system are a video segment and a phrase that describes the segment. The objects from the phrase are grounded for each sampled frame t. Object and proposal features are encoded to size d and visual-semantic similarity scores are computed. The ranking loss is weighted by a confidence score which combined with the penalty form the final loss. The object relations are further encoded to guide the loss weights (see Sec. 3.4 for details). During inference, the region proposal with the maximum similarity score with the object query is selected for grounding.

The similarity function and segment-description matching score are then:

$$a_k^{t,i} = \text{Sigmoid}(q_k^\top r_i^t/\sqrt{d}), \quad S(Q,R) = \frac{1}{O}\sum_{k=1}^{O} \max_{t,i} a_k^{t,i}, \tag{4}$$

where matrix $R = (r_1^1,\dots,r_N^1,r_1^2,\dots,r_N^T)$ indicates the pack of all proposal features.

This "brute-force" extension of Grounding by Ranking framework to the video domain presents two issues. First, depending on the video sampling rate, the total number of proposals per segment ($T \times N$) could be extremely large. Hence this solution does not scale well to long frame sequences. Second, an object existing sparsely across multiple frames might not be detected completely since successfully spotting it from one single frame would trigger a satisfactory match. We explain next how we propagate this weak supervisory signal from the segment level to frames that likely contain the target object.

3.3 Frame-wise Loss Weighting

In our framework, each frame is considered separately to ground the same target objects. Fig. 1 shows an overview of our model. We first estimate the likelihood that the query object is present in each video frame. If the likelihood is high, we judge the matching quality mainly on the matching loss (e.g., ranking loss). Otherwise, we down-weight the matching loss while bringing in a penalty loss. The lower the confidence, the higher the penalty. For clarity, we explain our idea when the matching loss is the ranking loss L_{rank} but note that this can be generalized to other loss functions.

Let the ranking loss for frame t be L_{rank}^t and the similarity score between query k and proposal i be $a_k^{t,i}$. Let $Q = (q_1,\dots,q_O)$ and $R_t = (r_1^t,\dots,r_N^t)$. We define the *confidence score* of the prediction at frame t as the visual-semantic matching score:

$$C_t = \frac{1}{O}\sum_{k=1}^{O} \max_i (a_k^{t,i}) \equiv S(Q,R_t), \tag{5}$$

where $S(\cdot,\cdot)$ is defined in Eq. 1. The corresponding *penalty* is:

$$D_t = -\log(2C_t) = -\log[\frac{2}{O}\sum_{k=1}^{O}\max_i(a_k^{t,i})], \tag{6}$$

inspired by [9]. The final loss for the segment is a weighted sum of frame-wise ranking losses and penalties:

$$L = \frac{1}{T}\sum_{t=1}^{T}[\lambda C_t L_{rank}^t + (1-\lambda)D_t], \tag{7}$$

$$L_{rank}^t = \sum_{R_t'\neq R_t}\sum_{Q'\neq Q}[\max(0, S(Q,R_t') - S(Q,R_t) + \Delta) + \max(0, S(Q',R_t) - S(Q,R_t) + \Delta)], \tag{8}$$

where λ is a static coefficient to balance the ranking loss and the penalty and can be validated on the validation set. A low λ might cause the system to be over-confident on the prediction.

3.4 Object Interaction

We assume that the object types and their order in the language description can roughly determine when they appear in the video content, as motivated in Sec. 1. We show that this language prior can work as the frame-wise confidence score. To consider the interaction among objects, we further encode each object query feature q_k as:

$$J(q_k) = \text{MA}(q_k, Q, Q), \tag{9}$$

where $\text{MA}(\cdot,\cdot,\cdot)$ is the multi-head self-attention layer [23], taking in the (query, key, value) triplet. It represents each query as the combination of all other queries based on their inter-relations. The built-in positional encoding layer [23] in multi-head attention captures the order of objects appearing in the description. Note that the formulation is non-autoregressive, *i.e.*, all the objects in the same description can interact with each other.

We evenly divide each video segment into T' snippets and predict the confidence score for object k to appear in each snippet based upon the concatenation of $J(q_k)$ and q_k. Note that T' is a pre-specified constant that satisfies $T' \leq T$. The language-based confidence score $C_{lang} = (C_{lang}^1, \ldots, C_{lang}^{T'})$ is formulated as:

$$C_{lang} = \frac{1}{O}\sum_{k=1}^{O}\text{Sigmoid}(W_{lang}[J(q_k);q_k] + b_{lang}), \tag{10}$$

where $[\cdot\,;\,\cdot]$ indicates the feature concatenation, $W_{lang} \in \mathbb{R}^{T'\times 2d}$ and $b_{lang} \in \mathbb{R}^{T'}$ are embedding weights and biases. We average the language-based and the similarity-based confidence score and rewrite Eq. 7 as:

$$L = \frac{1}{T}\sum_{t=1}^{T}[\lambda\frac{1}{2}(C_t + C_{lang}^{t_s})L_{rank}^t - (1-\lambda)\log(C_t + C_{lang}^{t_s})] \tag{11}$$

where $t_s = \min(\lceil t/\lceil \frac{T}{T'}\rceil\rceil, T)$ is the snippet index and $\lceil\cdot\rceil$ stands for the ceiling operator.

4 Experiments

4.1 Dataset

YouCook2-BoundingBox. YouCook2 [28] consists of 2000 YouTube cooking videos from 89 recipes. Each video has recipe steps temporally annotated (*i.e.* start timestamp and end

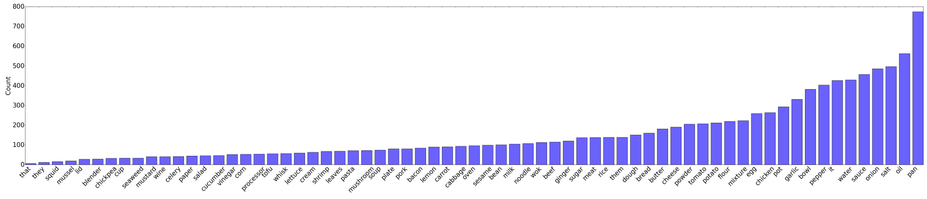


Figure 2: Frequency count of each class label (including referring expressions).

timestamp) and each segment is described by a natural language sentence. The average segment duration is 19.6s. Our training set is the same as the YouCook2 training split, only paired sentences are provided. For each segment-description pair in the validation and testing set however, we provide bounding box annotations for the most frequently appearing objects from the dataset, *i.e.* the top 63 recurring objects along with four referring expressions: *it, them, that, they* (see Fig. 2). These are used only during evaluation.

From YouCook2, we split each recipe step into a separate segment and sample it at 1 fps. We use Amazon Turk workers to draw bounding box around the objects in the video segment using the highlighted words in the sentence (from the 67 objects in our vocabulary). All annotations are further verified by the top 30 annotators. Please see the Appendix for more details on annotations and quality control.

4.2 Baselines and Metrics

Baselines. We include two competitive baselines from published work: DVSA [7] and GroundeR [19]. DVSA is the Grounding by Ranking method which we build all our methods upon. For fair comparison, all the approaches take in the same object proposals generated by Faster-RCNN [18] (pre-trained on MSCOCO). Following the convention from [6, 7], we select the top $N = 20$ proposals per frame and sample $T = 5$ frames per segment unless otherwise specified. We also evaluate the Baseline Random, which chooses a random proposal as the output.

Metrics. We evaluate the grounding quality by bounding box localization accuracy (denoted as Box Accuracy). The output is positive if the proposed box has over 50% IoU with the ground-truth annotation, otherwise negative. We compute accuracy for each object and average across all the object types.

4.3 Implementation Details

The number of snippets T' in Sec. 3.4 is set to 5. The encoding size d is 128 for all the methods. Object labels are represented as one-hot vectors, which are encoded by a linear layer without the bias term. The loss factor λ is cross-validated on the validation set and is set to 0.9. The ranking margin Δ is set to 0.1. For training, we use stochastic gradient descent (SGD) with Nesterov momentum. The learning rate is set at 0.05 and the momentum is 0.9. We implement the model in PyTorch and train it using either a single Titan Xp GPU with SGD or 4 GPUs with synchronous SGD, depending on the validation accuracy. The model typically takes 30 epochs, *i.e.* 4 hours to converge. More details are in the Appendix.

Table 1: Evaluation on localizing objects from the grounding-truth captions.

Method	Box Accuracy (%) Val.	Test
Compared methods		
Baseline Random	13.30	14.18
GroundeR [19]	19.63	19.94
DVSA [7]	30.51	30.80
Our methods		
Loss Weighting	30.07	31.23
Object Interaction	29.61	30.06
Full Model	30.31	31.73
Upper bound	57.77	58.56

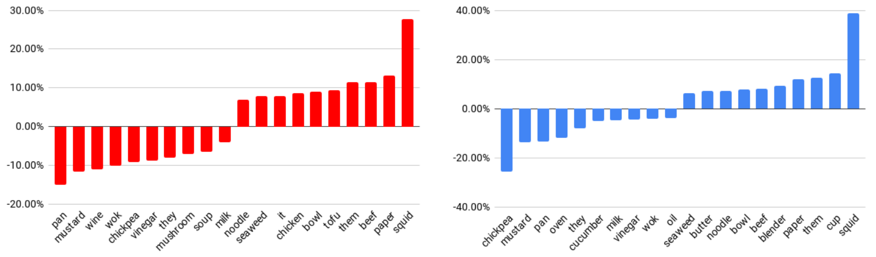


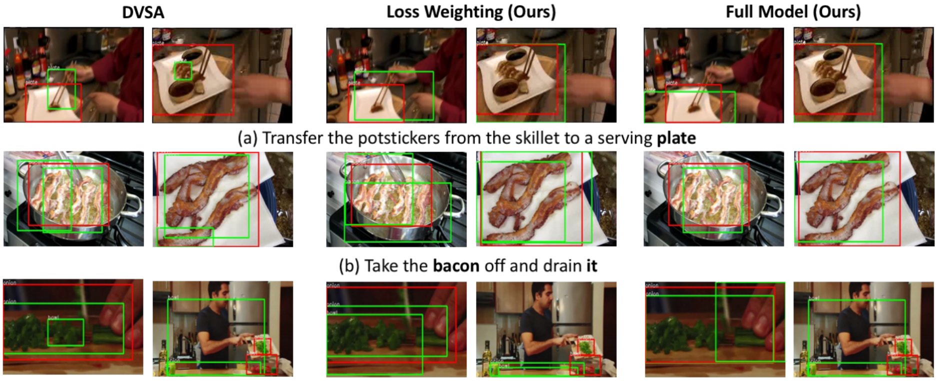
Figure 3: Top 10 accuracy increases & decreases by object category. (Left) Improvements of our Loss Weighting model over DVSA. (Right) Improvements of our Full Model over DVSA.

4.4 Results on Object Grounding

The quantitative results on object grounding are shown in Tab. 1. The model with the highest score on the validation set is evaluated on the test split. We compute the upper bound as the accuracy when proposing all 20 proposals, to see how far the methods are from the performance limit. Note that the upper bound reported here is lower than that in [19]. This is largely due to the domain shift from general scenes to cooking scenes and the large variance in our object states, e.g. zoom-in and zoom-out views, onions v.s. fried onion rings.

We show results on our proposed models, where the "Loss Weighting" model computes the confidence score with visual-semantic matching and the "Object Interaction" model computes the confidence score with textual guidance (Sec. 3.4). Our full model averages these two scores as the final confidence score (Eq. 11). The proposed methods demonstrate a steady improvement from the DVSA baseline, with a relative 1.40% boost from loss weighting and another 1.62% from combining object interaction, a total improvement of 3.02%. On the other hand, the baseline has a higher validation score, which indicates model overfitting. Note that text guidance alone ("Object Interaction") works slightly worse than the baseline, showing that both visual and textual information are critical for inferring the frame-wise loss weights. Our methods also outperform other compared methods, GroundeR and Baseline Random by a large margin.

Analysis. We show in Fig. 3 the top 10 accuracy increases and decreases of our methods over the DVSA baseline, by object category. Our methods make better predictions on static



(a) Transfer the potstickers from the skillet to a serving **plate**

(b) Take the **bacon** off and drain **it**

(c) Negative Example: Chop green **onion** and to **bowl**

Figure 4: Visualization of localization output from baseline DVSA and our proposed methods. Red boxes indicate ground-truths and green boxes indicate proposed regions. The first two rows show examples where our methods perform better than DVSA. The last row displays a negative example where all methods perform poorly. Better viewed in color.

objects such as "squid", "beef", and "noodle" and worse predictions on cookwares, such as "wok", "pan", and "oven", which involves more state changes, such as containing/not containing food or different camera perspectives. Our hypothesis is, our loss weighting framework favors consistent objects across frames, due to the shared frame-wise supervision.

Impact of Sampling Rate. We investigate the impact of high video sampling rate on grounding accuracy by increasing the total number of frames per segment (T) from 5 to 20. The accuracy from DVSA drops from 30.80% to 29.90% and the accuracy from our Loss Weighted model drops from 31.23% to 30.93%. We expected these inferior performances, due to the excessive object proposals. However, our loss weighted method only compromises 0.96% of the accuracy while the accuracy from DVSA drops by 2.92%, showing that our method is less sensitive to high sampling rate and predicts better on long frame sequences.

Qualitative Results. Fig. 4 visualizes the grounded objects with DVSA and our proposed methods. The first two rows show some positive examples. In Fig. 4 (a), we see with DVSA baseline the "plate" object is grounded to the incorrect regions in the frames. However our methods correctly select regions with a large IOU with the ground truth box. In Fig. 4 (b) the labels "bacon" and "it" refer to the same target object. Per our annotation requirements, there is only one ground truth box instead of two. The full model correctly combines both "bacon" and "it" grounds them to the same region proposal. The last row that shows where all methods fail to ground the target objects adequately. This may be a result of errors in the top object proposals proposed since the scene is rather complicated. An additional explanation may be bias in the dataset, where during training the "bowl" object typically occupies the majority of the frame.

Limitations. There are two limitations in our method we hope to address in our future work. First, even though the frame-wise loss can to some degree enforce the temporal consistency between frames, we do not explicitly model the relation between frames, for instance motion information. The transition between object states across frames, e.g., raw meat to cooked meat, should be further studied. Second, our grounding performance is upper-bounded by the object proposal accuracy and we have no control over the errors from the proposals. An end-to-end version of the proposed method that solves both the proposing and the grounding problem can potentially improve the grounding accuracy.

5 Conclusion

We propose a frame-wise loss weighted grounding model for video object grounding. Our model applies segment-level labels to the frames in each segment, while being robust to inconsistencies between the segment-level label and each individual frame. We also leverage object interaction as textual guidance for grounding. We evaluate the effectiveness of our models on the newly-collected video grounding dataset YouCook2-BoundingBox. Our proposed methods outperform competitive baseline methods by a large margin. Future directions include incorporating the video motion information and exploring an end-to-end solution for video object grounding.

Acknowledgement

This work has been supported by DARPA FA8750-17-2-0112. This article solely reflects the opinions and conclusions of its authors but not DARPA. We thank Tianhang Gao, Ryan Szeto and Mohamed El Banani for their helpful discussions.

References

[1] Muhannad Al-Omari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition and grounding for embodied robotic systems. In *AAAI*, pages 4349–4356, 2017.

[2] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416. IEEE, 2014.

[3] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. *AAAI*, 2018.

[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012.

[5] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pages 3270–3277, 2014.

[6] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional video. *To appear in CVPR*, 2018.

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

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

[9] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 2017.

[10] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised object discovery and tracking in video collections. In *ICCV*, pages 3173–3181. IEEE, 2015.

[11] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*, pages 36–52. Springer, 2016.

[12] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. *arXiv preprint arXiv:1711.06330*, 2017.

[13] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *CVPR*, pages 1020–1028, 2016.

[14] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, pages 685–694, 2015.

[15] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649. IEEE, 2015.

[16] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. *arXiv preprint arXiv:1711.08389*, 2017.

[17] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, pages 3282–3289. IEEE, 2012.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017.

[19] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, pages 817–834. Springer, 2016.

[20] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. *arXiv preprint arXiv:1704.01518*, 3, 2017.

[21] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

[22] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning*, pages 67–76, 2017.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

[24] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. *arXiv preprint arXiv:1705.01371*, 2017.

[25] Haonan Yu and Jeffrey Mark Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63, 2013.

[26] Haonan Yu and Jeffrey Mark Siskind. Sentence directed video object codiscovery. *International Journal of Computer Vision*, 124(3):312–334, 2017.

[27] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. *arXiv preprint arXiv:1801.08186*, 2018.

[28] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. *AAAI*, 2018.

[29] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. *arXiv preprint arXiv:1804.00819*, 2018.