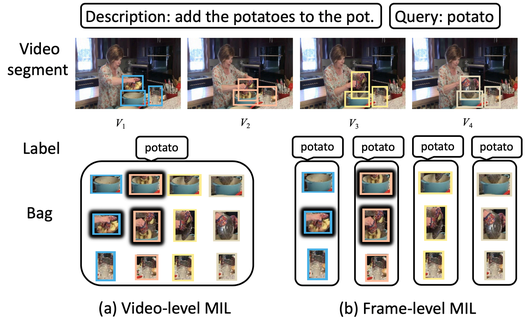
Not All Frames Are Equal: Weakly-Supervised Video Grounding with Contextual Similarity and Visual Clustering Losses

Jing Shi[1] Jia Xu[2] Boqing Gong[2] Chenliang Xu[1]
[1]University of Rochester [2]Tencent AI Lab

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- **Video Grounding Definition:** Given a video segment with its language description, the aim is to localize objects query from the description to the video.
- **Motivation:** We investigate the problem of weakly-supervised video grounding, where only video-level sentences are provided. Both video-level MIL and frame-level MIL can potentially tackle the problem, but video-level MIL has increasingly large bag sizes as frame number increases, and frame-level MIL frequently triggers false-positive bags. Moreover, the video temporal consistency should be considered.



- **Frame-level MIL with Ranking loss:** We denote the similarity of a frame and a sentence as $S(V_t, \mathcal{Q})$, it is calculated by first computing the similarity between each query and its matched region, then averaging the similarity of all queries in the sentence. The frame ranking loss is

$$L^t_{rank} = \max(0, S(V_t, \mathcal{Q}') - S(V_t, \mathcal{Q}) + \Delta) + \max(0, S(V'_t, \mathcal{Q}) - S(V_t, \mathcal{Q}) + \Delta) \ .$$

- **Contributions**
  - Following frame-level MIL, we design a contextual similarity between query and frame to deal with sparse objects association across frames.
  - We propose the visual clustering loss that can leverage temporal coherence in the video by strengthening the clustering effect of similar visual features.

Methodology

Contextual Similarity

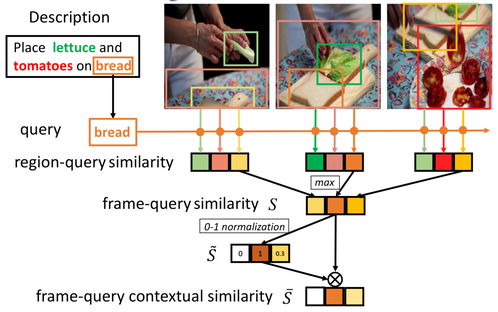


Diagram for calculation of the frame-query contextual similarity score

- Contextual similarity reweighs the frame importance according to frame-query similarity, so as to alleviate false positive bag.

$$S(V_t, q_k) = \max_n a^{t,n}_k \qquad \tilde{S}(V_t, q_k) = \frac{S(V_t, q_k) - \min_t S(V_t, q_k)}{\max_t S(V_t, q_k) - \min_t S(V_t, q_k)}$$

$$\bar{S}(V_t, q_k) = S(V_t, q_k)\tilde{S}(V_t, q_k) \qquad S(V_t, \mathcal{Q}) = \frac{1}{K}\sum_{k=1}^{K} \bar{S}(V_t, q_k)$$

Visual Clustering

- Visual clustering forces the similarity between similar visual features across frames to learn a more discriminative visual feature.

$$\hat{v}_{t,k} = \underset{v^n_t \in \{v^1_t, \ldots, v^N_t\}}{\arg\max} \ q^{\mathbf{T}}_k v^n_t$$

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

- Final loss: combination of ranking loss and visual clustering loss

$$L = \sum_{t=1}^{T} L^t_{rank} + \lambda L^{ctx}_{vis}$$

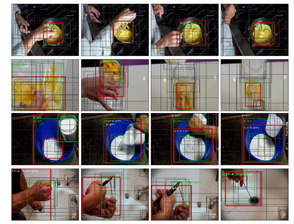
Experiments

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**

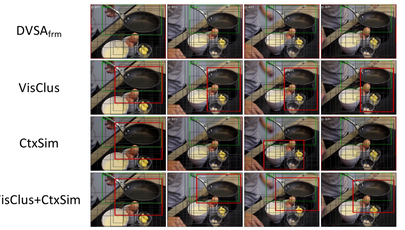
Comparison with other methods on YouCookII in Finite-Class Training.

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**

Generalizability Test, trained on YouCookII in ICT and testing on RoboWatch



Qualitative result on RoboWatch.



Qualitative comparison on YouCookII.
Description: “Put a pan on medium to high heat”, query: “pan”

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.

Reference

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