Grounded Video Description

Luowei Zhou[1,2], Yannis Kalantidis[1], Xinlei Chen[1], Jason J. Corso[2], Marcus Rohrbach[1]
[1] Facebook AI, [2] University of Michigan
github.com/facebookresearch/grounded-video-description

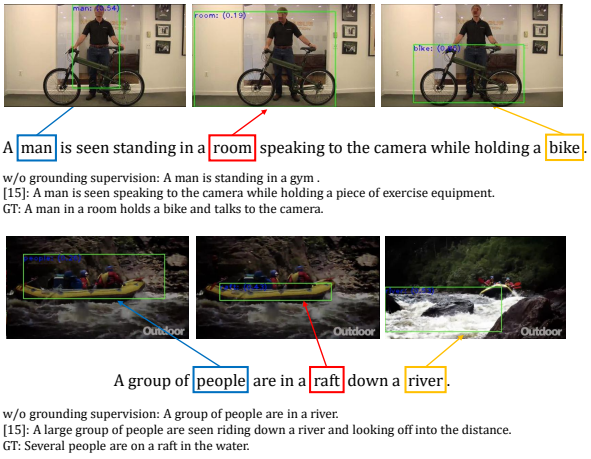


Figure 1: Word-level grounded video descriptions generated by our model on two segments from our ActivityNet-Entities dataset. We also provide the descriptions generated by our model without explicit bounding box supervision, [15] and the ground-truth descriptions (GT) for comparison.

1. Introduction

Image and video description models are frequently not well grounded [6] which can increase their bias [4] and lead to hallucination of objects [9], *i.e*. the model mentions objects which are not in the image or video *e.g*. because they might have appeared in similar contexts during training. This makes models less accountable and trustworthy, which is important if we hope such models will eventually assist people in need [1, 11]. Additionally, grounded models can help to explain the model's decisions to humans and allow humans to diagnose them [7]. While researchers have started to discover and study these problems for image description [6, 4, 9, 7],they are even more pronounced for video description due to the increased difficulty and diversity, both on the visual and the language side.

Fig. 1 illustrates this problem. A video description approach (without grounding supervision) generated the sentence "A man standing in a gym" which correctly mentions "a man" but hallucinates "gym" which is not visible in the video. Although a man is in the video it is not clear if the model looked at the bounding box of the man to say this word [4, 9]. For the sentence "A man [...] is playing the piano" in Fig. 2, it is important to understand that which "man" in the image "A man" is referring to, to determine if a model is correctly grounded. Such understanding is crucial for many applications when trying to build accountable systems or when generating the next sentence or responding to a follow up question of a blind person: *e.g*. answering "Is *he* looking at me?" requires an understanding which of the people in the image the model talked about.

The goal of our research is to build such grounded systems. As one important step in this direction, we collect ActivityNet-Entities (short as ANet-Entities) which grounds or links noun phrases in sentences with bounding boxes in the video frames (see Fig. 2 for an example). It is based on ActivityNet Captions [5], one of the largest benchmarks in video description.

Our new dataset allows us to introduce a novel grounding-based video description model that learns to jointly generate words and refine the grounding of the objects generated in the description. We explore how this explicit supervision can benefit the description generation compared to unsupervised methods that might also utilize region features but do not penalize grounding.

Our contributions are two-fold. First, we collect our large-scale ActivityNet-Entities dataset, which grounds video descriptions to bounding boxes on the level of noun phrases. Our dataset allows both *teaching* models on explicit object grounding and *evaluating* the grounding accuracy. Second, we propose a grounded video description framework which is able to learn from the bounding box supervision in ActivityNet-Entities and we demonstrate its superiority over baselines and prior work in generating grounded video descriptions.

2. ActivityNet-Entities Dataset

In order to train and test models capable of explicit grounding-based video description, one requires both language and grounding supervision. Although Flickr30k Entities [8] contains such annotations for images, no large-

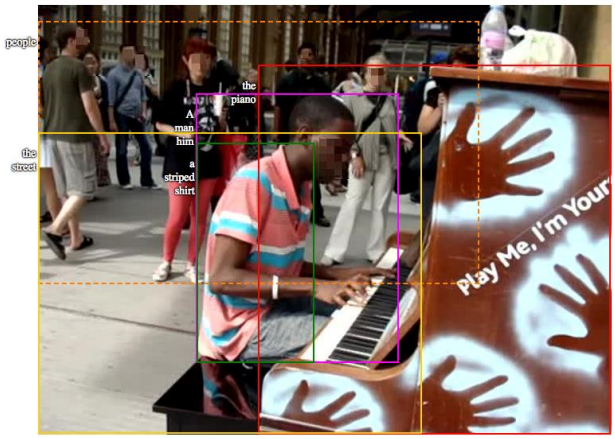
Method	Bleu@1	Bleu@4	METEOR	CIDEr	SPICE	Attn.	Grd.	$F1_{all}$	$F1_{loc}$	Cls.
Masked Transformer [15]	22.9	**2.41**	10.6	**46.1**	13.7	–	–	–	–	–
Bi-LSTM+TempoAttn [15]	22.8	2.17	10.2	42.2	11.8	–	–	–	–	–
Our Unsup. (w/o SelfAttn)	23.1	2.16	10.8	44.9	**14.9**	16.1	22.3	3.73	11.7	6.41
Our Sup. Attn.+Cls. (GVD)	**23.6**	2.35	**11.0**	45.5	14.7	**34.7**	**43.5**	**7.59**	**25.0**	**14.5**

(a) Results on ANet-Entities test set.

	vs. Unsupervised		vs. [15]	
Method	Judgments %	Δ	Judgments %	Δ
About Equal	34.9		38.9	
Other is better	29.3	6.5	27.5	6.1
GVD is better	**35.8**		**33.6**	

(b) Human evaluation of sentences.

Table 1: (a) Results on ANet-Entities test set. The top one score for each metric is in bold. (b) Human evaluation of sentence quality. We present results for our supervised approach vs. our unsupervised baseline and vs. Masked Transformer [15].



A man in a striped shirt is playing the piano on the street while people watch him.

Figure 2: An annotated example from our dataset. The dashed box ("people") indicates a group of objects.

scale description datasets with object localization annotation exists for videos. The large-scale ActivityNet Captions dataset [5] contains dense language annotations for about 20k videos from ActivityNet [2] but lacks grounding annotations. Leveraging the language annotations from the ActivityNet Captions dataset [5], we collected entity-level bounding box annotations and created the ActivityNet-Entities (ANet-Entities) dataset , a rich dataset that can be used for video description with explicit grounding. With 15k videos and more than 158k annotated bounding boxes, ActivityNet-Entities is the largest annotated dataset of its kind to the best of our knowledge.

When it comes to videos, region-level annotations come with a number of unique challenges. A video contains more information than can fit in a single frame, and video descriptions reflect that. They may reference objects that appear in a disjoint set of frames, as well as multiple persons and motions. To be more precise and produce finer-grained annotations, we annotate *noun phrases* (NP) rather than simple object labels. Moreover, one would ideally have dense region annotations at every frame, but the annotation cost in this case would be prohibitive for even small datasets. Therefore in practice, video datasets are typically sparsely annotated at the region level [3]. Favouring scale over density, we choose to annotate segments as sparsely as possible

Dataset	Domain	# Vid/Img	# Sent	# Obj	# BBoxes
Flickr30k Entities [8]	Image	32k	160k	480	276k
MPII-MD [10]	Video	$\ll$1k	$\ll$1k	4	2.6k
YouCook2 [14]	Video	2k	15k	67	135k
ActivityNet Humans [12]	Video	5.3k	30k	1	63k
ActivityNet-Entities (ours)	**Video**	**15k**	**52k**	**432**	**158k**
–train		10k	35k	432	105k
–val		2.5k	8.6k	427	26.5k
–test		2.5k	8.5k	421	26.1k

Table 2: Comparison of video description datasets with noun phrase or word-level grounding annotations.

and annotate every noun phrase only in one frame inside each segment. Dataset stats and comparisons with other related datasets can be found in Tab. 2.

3. Experiments

Details regarding methods and evaluations can be found at [13]. We show in Tab. 1a the results on description quality, object localization accuracy (indicated by Attn., Grd., $F1_{all}$, and $F1_{loc}$) and region classification accuracy (Cls.). All metrics are the higher the better. Our supervised method (GVD), *i.e.*, with box supervision during training, outperforms the unsupervised counterpart in four of the five language metrics and gets significantly boosts in all the localization/classification metrics. We also sets the new SotA on Bleu@1, METEOR and SPICE metrics, with relative gains of 2.8%, 3.9% and 6.8%, respectively over the previous best [15]. We observe slightly inferior results on Bleu@4 and CIDEr (-2.8% and -1.4%, respectively) but after examining the generated sentences, we see that prior work [15] generates repeated words way more often. The human evaluation in Tab. 1b further supports our discoveries from the automatic evaluation, that our GVD method generates descriptions with better quality.

Acknowledgement. The technical work was performed during Luowei's summer internship at Facebook AI Research. This work is also partly supported by DARPA FA8750-17-2-0125 and NSF IIS 1522904. Luowei Zhou and Jason Corso were partly supported by DARPA FA8750-17-2-0125 and NSF IIS 1522904 as part of their affiliation with University of Michigan. This article solely reflects the opinions and conclusions of its authors but not the DARPA or NSF.

References

[1] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM. 1

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[3] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE international conference on computer vision*, 2018. 2

[4] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787, 2018. 1

[5] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 1, 2

[6] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. Attention correctness in neural image captioning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 4176–4182, 2017. 1

[7] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[8] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 2

[9] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1

[10] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[11] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 2017. 1

[12] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1453–1462, 2017. 2

[13] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[14] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2

[15] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 1, 2