Weakly-supervised Visual Grounding of Phrases with Linguistic Structures

Zhang, Liqiang

May 31, 2018

Abstract

In this paper, the author propose a weakly-supervised approach that takes image-sentence pairs as input and learns to visually ground arbitrary linguistic phrases, in the form of spatial attention masks. Specifically, the model is trained with images and their associated image-level captions, without any explicit region-to-phrase correspondence annotations. To this end, author introduce an end-to-end model which learns visual groundings of phrases with two types of carefully designed loss functions.

1. Introduction

Visual recognition research has made tremendous strides in recent years, achieving unprecedented performance in various tasks including image classification [1], object detection [2], semantic segmentation, and image captioning. However, traditional supervised frameworks for these tasks often rely on large datasets with expensive bounding box or pixel-level segmentation annotations. As the field pushes toward solving larger-scale and more complex problems, obtaining massive annotated datasets is becoming a critical bottleneck.

Figure 1. From the phrase a man that is cutting sandwich, we can infer that a man and sandwich should be exclusive to each other spatially.

While great progress has been made, learning from a list of category tags ignores the rich semantics and structure in natural language that humans use to describe visual data. For example, in Fig. 1, a tag-based description would simply list {man, sandwich, table} whereas a natural language description might say "a man that is cutting sandwich on a table".Importantly, the natural language description provides *structure*, which can benefit a weakly-supervised learning algorithm [3].

In this paper, the key idea is to utilize the rich structure in a natural language description by transforming it into a hierarchical parse tree of phrases (see Fig. 2). In this way, it could extract two types of linguistic structural constraints for visual grounding: (1) compositionality of attention masks among children and their parent phrases, and (2) complementarity among the attention masks between sibling phrases.

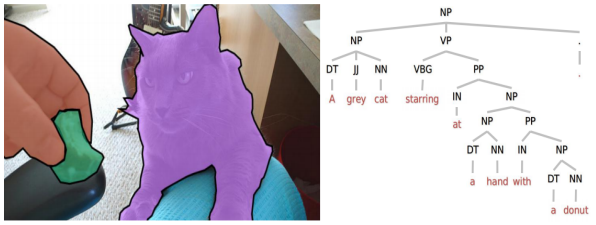

Figure 2. An illustration of the concept.We exploit structures present in natural language to provide regularities and constraints for grounding free-form language on images. Note that we do not use any ground-truth masks during training.

2. Method

This loss function is used to match the corresponding image-phrase pairs. Given an input image I_i and a set of corresponding phrases (both positive and negative ones)$\{P_i^1, P_i^2, P_i^3, ..., P_i^n\}$, the author compute the discriminative loss as:

$$L_{disc} = -Y_i^j \cdot Sigmoid(\phi_V(I_i) \cdot \phi_L(P_i^j), \qquad (1)$$

where $Y_i^j \in \{-1, 1\}$ is the indicator variable denoting whether P_i^j is a negative/positive match to I_i and $\phi_V(I)$ and $\phi_V(P)$ denote the visual and language code, respectively. The positive phrases are those in the parse tree associated

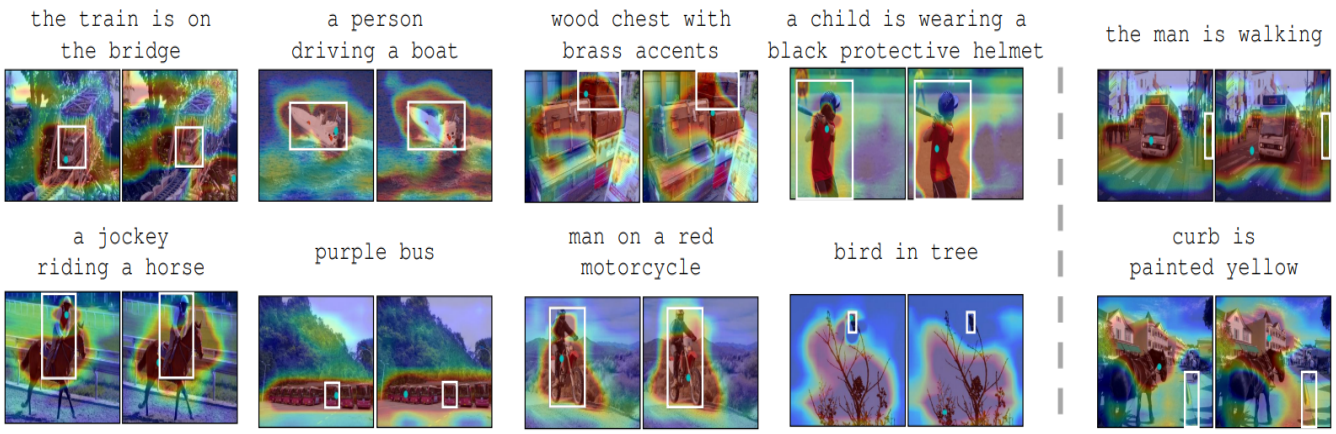

Figure 3. Comparing with the baseline model trained without structural constraints (Disc-only). In each image pair, the left is result and the right is the baselines result. The ground-truth bounding box is annotated with a white solid line, whereas the maximum point of the prediction is denoted with a cyan dot. The ground-truth phrase associated with each image is shown on top of each image. The last column shows difficult examples containing small or infrequent objects.

with the input image I_i while the negative phrases are randomly sampled from those in the parse tree associated with any other image.

3. Results

Fig. 3 shows qualitative example predictions on the Visual Genome dataset. For example, for the train is on the bridge the model accurately pinpoints the train and the bridge whereas the Disc-only baseline produces high responses on many irrelevant pixels. A similar thing happens for a person driving a boat. Furthermore, in some cases, even though the maximum point of the baseline attention mask falls within the ground-truth bounding box, we can clearly see that the generated mask is not as clean as that of this model; This is likely because the baseline model does not have any constraints to exploit other than the discriminative loss, whereas our model explicitly enforces structural priors onto the generated attention masks.

	IOU@0.3	IOU@0.4	IOU@0.5	Avg mAP
Disc-only	0.302	0.199	0.110	0.203
PC	0.327	0.213	0.118	0.219
SIB	0.316	0.203	0.114	0.211
Token	0.334	0.240	0.138	0.238
Ours	0.347	0.246	0.159	0.251

Table 1. Segmentation mAP on MS COCO across all 80 categories.

Table 1 shows the segmentation results, in term of mean Average Precision (mAP) over all 80 MS COCO categories at different IOU thresholds. Both *PC* and *SIB* provide consistent improvement over *Disc-only* across different IOU thresholds [4]. This demonstrates that both structural constraints effectively transfer their respective structure from the language domain to the visual domain.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and K. Dan. Neural module networks. In *CVPR*, 2016. 1

[2] S. Antol, A. Agrawal, J. Lu, and M. Mitchell. VQA: Visual question answering. In *ICCV*, 2015. 1

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 1

[4] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015. 2