Weakly-supervised Visual Grounding of Phrases with Linguistic Structures

Fanyi Xiao, Leonid Sigal*, and Yong Jae Lee

University of California, Davis

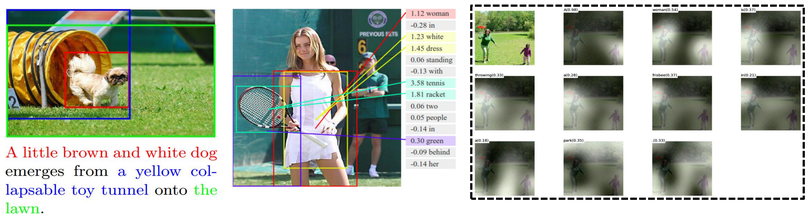
*Disney Research

Goal

Produce visual groundings of linguistic phrases at all levels (words, phrases, sentences), by training with weak supervision (img-cap pairs).

a man | sandwich | a man that is cutting sandwich on a table

Previous works



A little brown and white dog emerges from a yellow collapsable toy tunnel onto the lawn.

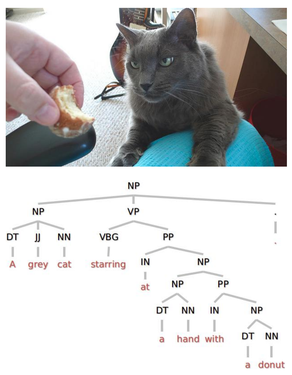
- Treat the caption as a sequence of word tokens [*Xu 2014, Rohrbach 2016, ...*].
- Require phrase-region pairs or pretrained detectors [*Karpathy 2015, Plummer 2015, ...*].

Motivation & Key ideas

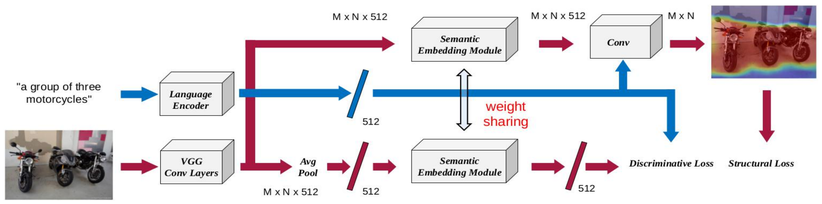
- Hard, if not impossible, to annotate a large collection of phrase-segment pairs.
- A sentence is not simply a sequence of token. We leverage structure in natural language for weakly supervised visual grounding.

Benefits of exploiting linguistic structures

- Avoid grounding non-sensical tokens ("a", "the", etc.)
- Enrich training data, in a linguistically sound way (i.e., words, phrases, sentence)
- Transfer linguistic structure to visual domain



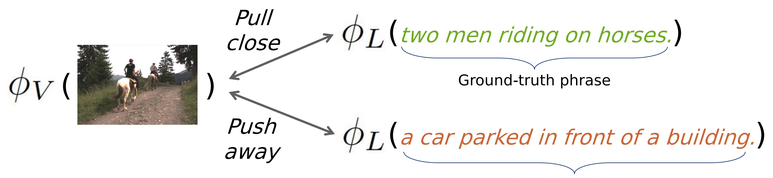
Network architecture



(1) visual encoder (2) language encoder (3) semantic embedding module (4) loss functions

Loss functions

$$L = L_{PC} + L_{SIB} + L_{disc}$$

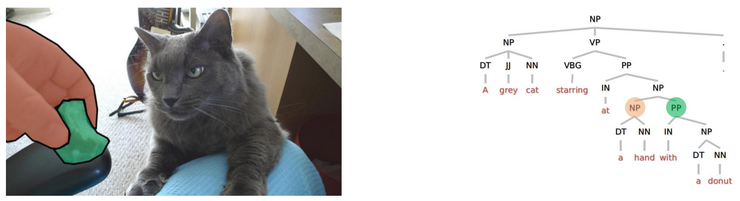


$$L_{disc} = -Y_i^j \cdot Sigmoid(\phi_V(I_i) \cdot \phi_L(P_i^j))$$

Transferring linguistic structures

- Sibling exclusivity constraint: Masks for nodes that are siblings should be *exclusive*:

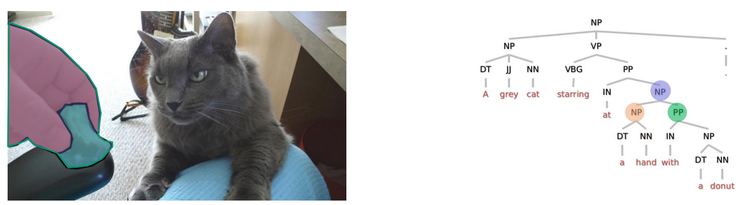
$$L_{SIB} = -\frac{1}{|S|} \sum_{m \in S} \sum_{pixels \in A} W_m \cdot \log \frac{\max_n A_{m,n}}{\sum_n A_{m,n}}$$



A grey cat starring at a hand with a donut

- Parent-child inclusivity constraint: Mask of parent node should be *union* of children's masks:

$$L_{PC} = \frac{1}{|P|} \sum_{k \in P} ||A_k - \max_{l \in child(k)} A_l||^2$$

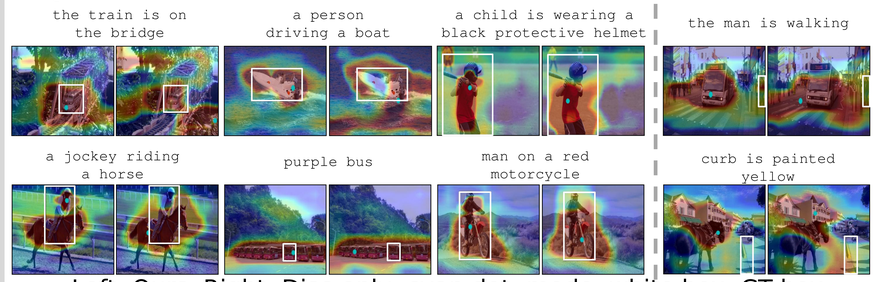


A grey cat starring at a hand with a donut

Results

Train models on MS COCO; test on both MS COCO and Visual Genome

- Qualitative results on Visual Genome



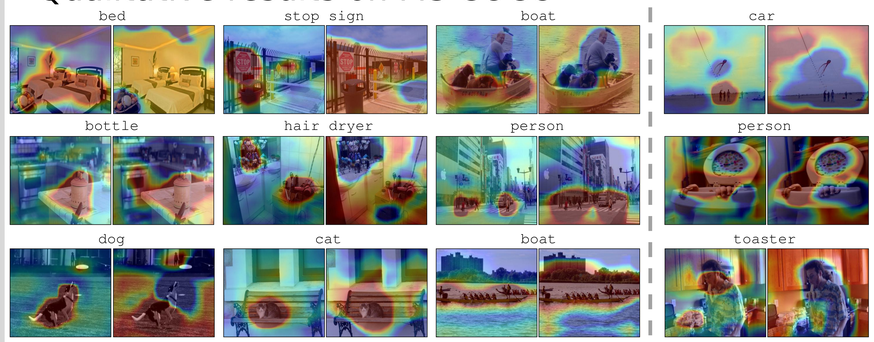
Left: Ours, Right: Disc-only, cyan dot: mode, white box: GT box

- Quantitative results on Visual Genome

	Random	Disc-only	Token	PC	SIB	Ours
Accuracy	0.115	0.230	0.222	0.236	0.231	0.244

Pointing Game [*Zhang 2016*] measure -- the frequency of the mode being contained by the GT box

- Qualitative results on MS COCO



Our results (left) are more accurate (bottle, dog) and cleaner (cat, boat)

- Quantitative results (Seg mAP) on MS COCO

	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

Multiple queries on same input image