Supplementary Material – Weakly-supervised Visual Grounding of Phrases with Linguistic Structures

Anonymous CVPR submission

Paper ID 2570

In this document, we provide additional materials to supplement our main submission. In the first section, we provide further details on how we choose the weights (λ's) for the different loss terms. In the second section, we show additional qualitative examples on COCO segmentation and Visual Genome phrase localization, and in particular, examples that have multiple labels/phrases per image; this demonstrates that our model is conditioning the attention mask generation process on the input language, instead of simply learning a generic saliency map.

1. Setting λ_{PC} and λ_{SIB}

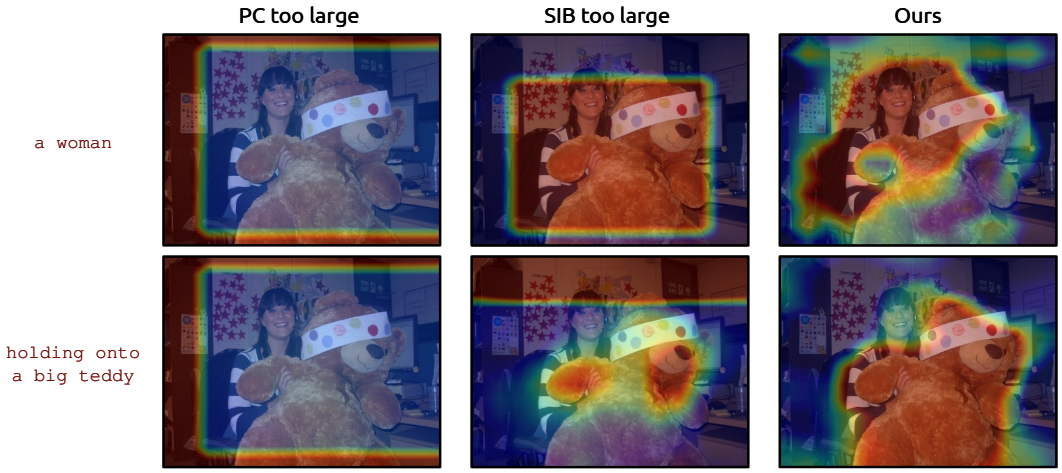

Figure 1. Visualization of attention masks for setting λ_{PC} and λ_{SIB}. For two phrases "a woman" and "holding onto a big teddy", we show the corresponding attention masks, for three different settings. In the first column, we show the attention masks for a model trained with λ_{PC} being too large, which leads to the result that all attention masks are roughly the same. In the second column, we show the effect of having too large λ_{SIB} – here, weird artifacts are generated to enforce exclusivity. Finally, with a properly set λ_{PC} and λ_{SIB}, we obtain a reasonable visualization without artifacts. Through these visualizations, we can correctly set the weights for each loss term.

In Sec. 3.3.1 of our main paper, we have the following equation: $L_{struct} = \lambda_{PC} L_{PC} + \lambda_{SIB} L_{SIB}$. Here we discuss how we select the appropriate λ for the different terms.

Since we are working in a weakly-supervised problem setting, we do not have any ground-truth region-phrase annotations to validate our model (i.e., computing accuracy on the held-out validation data). Therefore, we instead set the λ's based on qualitative visualizations.

Specifically, we find that models trained with a too high λ_{PC} start to take a shortcut – they simply predict all attention masks to be the same, which definitely satisfies all of the parent-child constraints as implied by Eq. (2) in our main submission. This phenomenon is shown in the first column of Fig. 1. On the other hand, if we set λ_{SIB} to be too big, the model starts to generate weird artifacts to satisfy exclusivity between different regions of the image, without taking into account the image content, as shown in the second column of Fig. 1. In order to generate reasonable attention masks (as shown in the last column

of Fig. 1), we need to properly set the weights. We rely on this attention mask visualization during training to determine the appropriate weights; in our case, $\lambda_{PC} = 0.01$ and $\lambda_{SIB} = 0.0001$.

2. Additional qualitative results

We next provide additional qualitative results with a focus on demonstrating the localization of *multiple* phrases per image. We first present results of localizing phrases on Visual Genome, and then show results of localizing object labels on MS COCO.

Visual Genome The phrase localization results on Visual Genome are shown in Fig. 2. Our model generates attention masks conditioned on the input phrase, instead of simply outputting a generic saliency map. For example, for "the clock tower is tall", our model correctly highlights the clock tower in the image, whereas it is not highlighted anymore when the input phrase is changed to "buildings by the street".

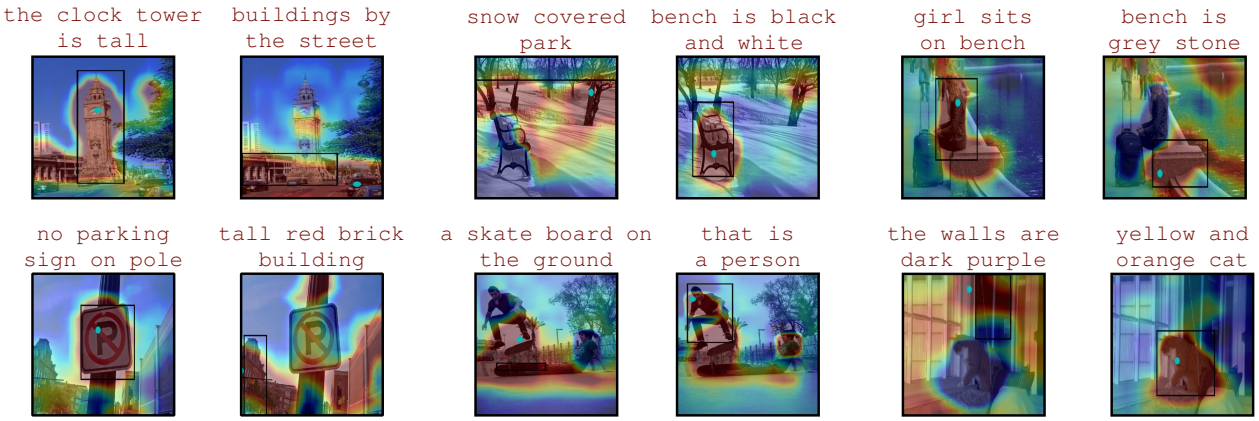

Figure 2. Results on Visual Genome. Our model generates different attention masks for different input phrases. The input phrases are shown above the respective attention masks, in red font. The black box in each image is the ground-truth bounding box corresponding to the phrase. The cyan dot denotes the maximum confidence point from our predicted attention mask. Clearly, our model outputs the attention masks conditioned on the phrase input.

MS COCO Finally, we demonstrate that our model can generate different attention masks for different object label/tag inputs as well. See Fig. 3. For example, we clearly see the different attention masks generated for "bus" and "sheep" in the first pair of images, and this holds for other examples as well. These results demonstrate that our model is clearly conditioning its attention mask generation process on the input language.

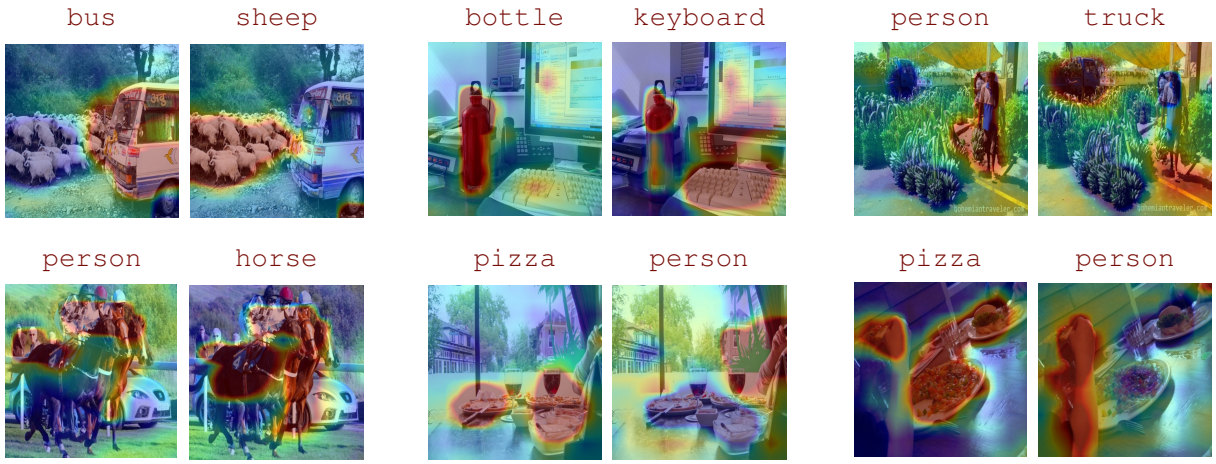

Figure 3. Results on MS COCO. For each pair of images, the object labels are shown above the respective attention masks in red font. For the same image, our model can generate different attention masks for different object labels.