Supplementary Material for: Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen,
Carl Vondrick, and Shih-Fu Chang

Columbia University, New York, NY, USA
{ha2436,sk4089,sb4019,bc2754,cv2428,sc250}@columbia.edu

Abstract

We provide in this supplementary materials some additional visualization results and discuss what seems to be the strengths and weakness of our method based on those examples.

1. Visualization results

We report in Figure 1 some additional visualization of our method results where our model produces reasonably good heatmaps. Our model does a fairly good job at distinguishing the main focus of the image from the context scene (grass and beach) of objects (net and bridge).

Then, in Figure 2 we give an example of a complex sentence with multiple queries of a cluttered scene and the corresponding heatmaps. Even if the model (based on VGG16) operates on a feature map of just 18×18, it is able to properly distinguish the different elements of that cluttered scene.

Finally, in Figure 3 we show some localization failure examples. The first two rows show failures to detect instruments, more precisely a guitar in these examples, which seems to be one type of objects our model struggles to properly localize. The last two rows show failures for the word "dock", which our weakly supervised model seem not able to distinguish from the water nearby. Our assumption, at that time, is that these words co-occur often in the training set with other words, i.e "guitar" with the word "man" and "dock" with the word "water". If one word almost always co-appear in the training set with another concept, it can be challenging in our weakly supervised setting for the model to learn to distinguish these two co-occurring concepts.

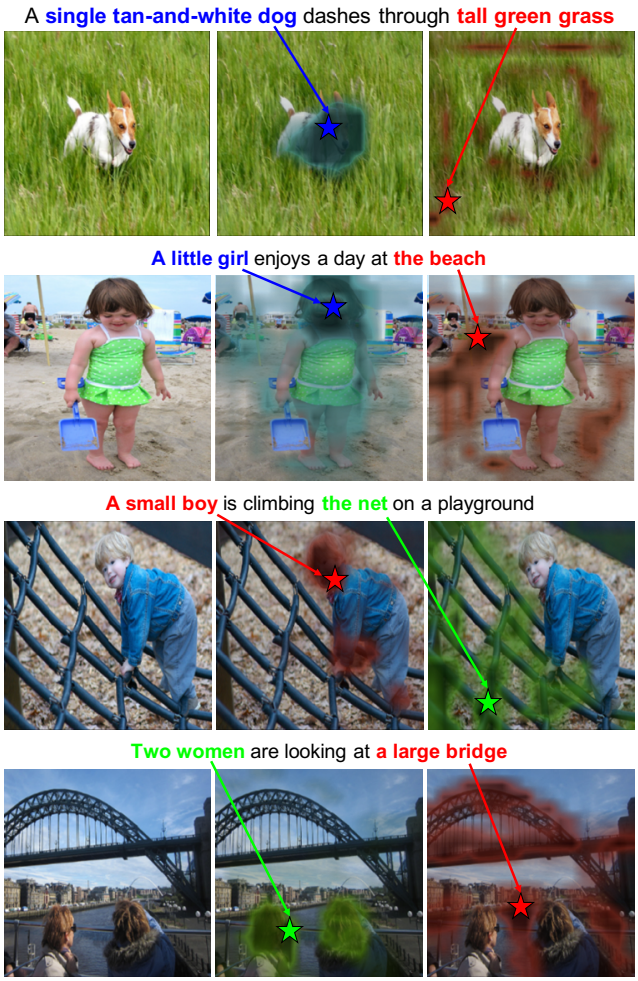

Figure 1. Some image-sentence pairs from Flickr30K, with two queries (colored text) and corresponding heatmaps and selected max value (stars).

Figure 2. Some image-sentence pairs from Flickr30K, with two queries (colored text) and corresponding heatmaps and selected max value (stars).

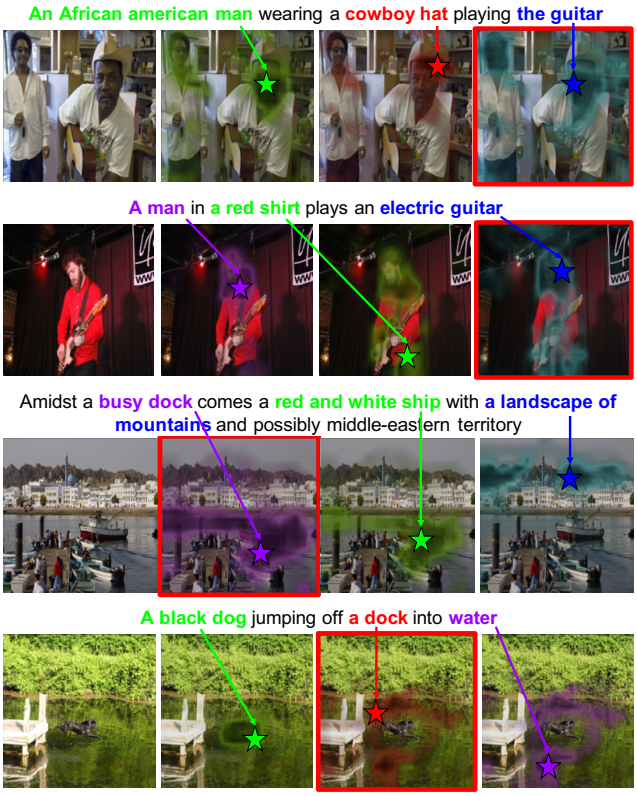

Figure 3. Image-sentence pairs from Flickr30K, with three queries (colored text) and corresponding heatmaps and selected max value (stars). In the first two rows the model fails to properly select the guitar, in the last two rows the model fails to properly localize the dock.