Supplementary Material
Learning Unsupervised Visual Grounding Through Semantic Self-Supervision

1 Introduction

In the supplementary material, we qualitatively highlight the following:

- How our generated heatmap differs from the VGG16 feature maps which are used as the visual input to our model.
- How the alignment of the chosen ground truth concept, the predicted concept and the actual entity to be grounded affects the quality of the phrase grounding.

2 Comparison with VGG16 feature maps

Our model uses a pre-trained VGG16 model to extract feature maps for our visual encoder. In this section we show that even though the visual features are fixed during training, our model can learn attention maps which are spatially distinct from the VGG16 feature maps used as input. We use the channel averaged VGG16 baseline model for visualizing the VGG16 heatmap. Figure 1 shows the comparison between this and our predictions. As evident from the examples, our method produces attention maps which can localize regions which were weak or even non-existent in the activations of the input feature maps. This shows that our model doesn't simply amplify the activations present in VGG16 channels but learns a phrase dependent attention map.

3 Alignment of the selected, predicted and true concept

In this section, we qualitatively cover four broad cases for summarizing the effects of the selected concept and the predicted concept and how these two relate to what the actual entity to be localized was. Note that our proxy loss is trained with the *selected concept* as the ground truth and predicts the *common concept* and *independent concept*. In Figure 2, *common concept* is denoted by red blocks, *independent concept* is denoted by blue blocks and the *selected concept* is denoted by gray blocks. For the remainder of this section we use the term *true concept* to refer to the actual entity to be localized.

First row: Correct grounding of phrase. In cases where the *selected concept* and all *predicted concepts* coincide with the *true concept* to be localized, the generated heatmap produces a good localization of the phrase. This is shown in the first row of Figure 2 with the concept *'headlight'* and *'picture'*.

Second row: Incorrect grounding due to wrong concept-selection. In cases where the *selected concept* is incorrect, *ie.* it's not the same as the *true concept*, even with the correct decoding, the localizations produced are wrong. For example in the second row of Figure 2, instead of selecting *'building'* and *'switch'*, the incorrect selection of *'top'* and *'wall'* leads to localization which is correct for the selected ground truth, but incorrect for the phrase.

Third row: Incorrect grounding due to wrong concept-learning. In these cases, the selected concept is correct but the decoder predicts incorrect common/independent concepts, due to which the final phrase grounding is affected. For example in the third row of Figure 2, even though *'window'* and *'tire'* are correct *selected concepts*, the concept-learning inaccurately predicts *'glass'* and *'car/vehicle'* which in turn generates a localization respecting the *predicted concept*.

Fourth row: Incorrect grounding due to challenging phrase-image pairs. Lastly, there are some cases where the entity to be localized is either ambiguous or simply too hard (due to a small size in

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

the image or due to a complicated phrase structure). In these cases, the grounding is incorrect across the different possibilities of alignment of the aforementioned concept. For example in the fourth row of Figure 2, the concept *'pole'* exists at multiple visual locations while in the other example, the concept *'lighter'* occupies a very small space in the visual region.

4 Additional outputs from our model

In Figure 3, the first two rows show a typical concept batch with *'ice'* and *'television'* as the respective common concepts. The third row shows some small and challenging entities to be grounded. Finally, the fourth and fifth row highlight the ability of the model to output completely different heatmaps for the same image having differing phrases. The grounded heatmap appear to identify regularities like localizing *'television'* towards the periphery near a wall or localizing *'phone'* near the hands of a person. We also note that since the concept batch is trained with concepts from very diverse contexts, the model is forced to learn high-level semantics about the image (see first row).

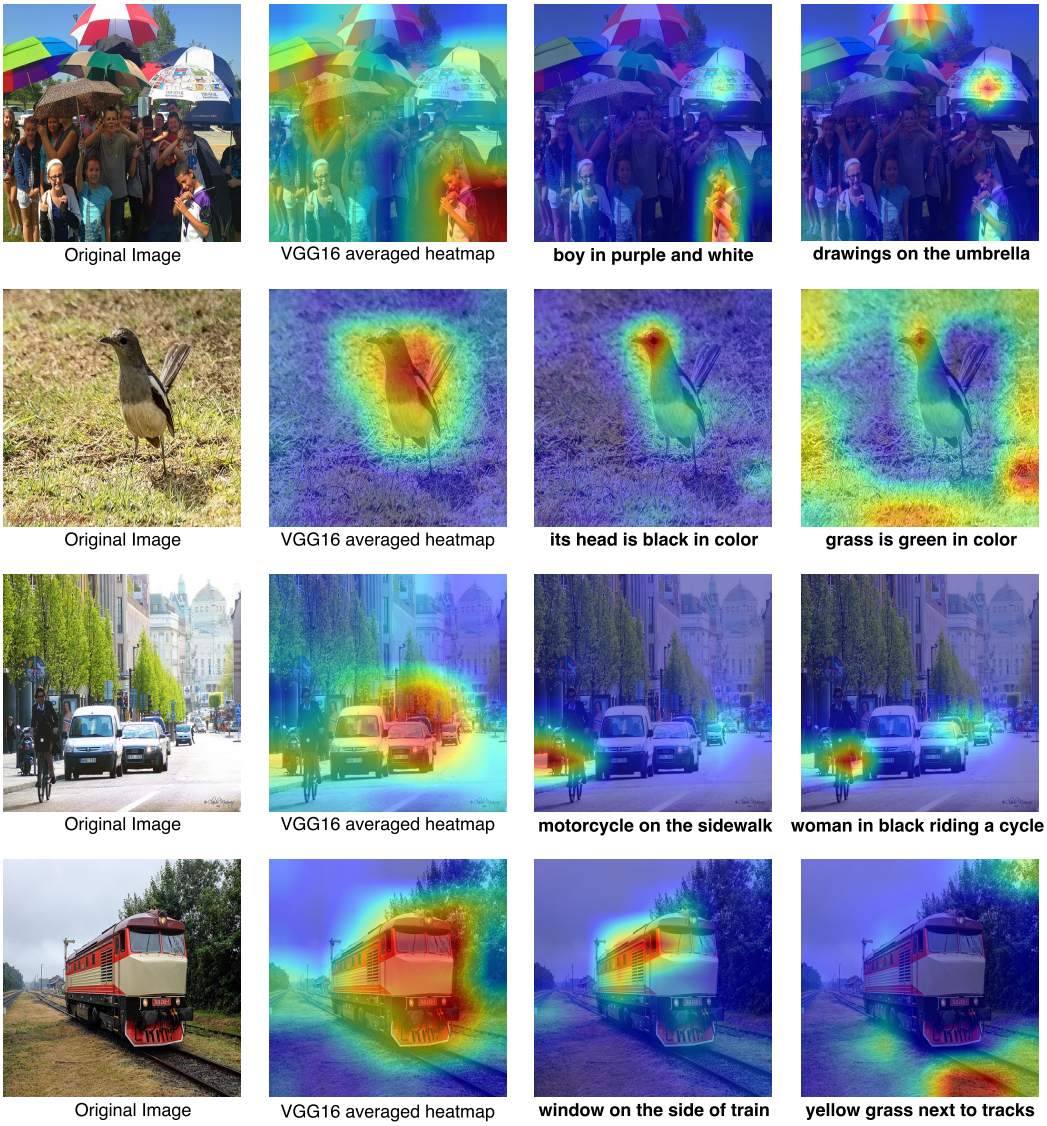

Figure 1: Comparison of VGG16 feature maps with our generated attention maps.

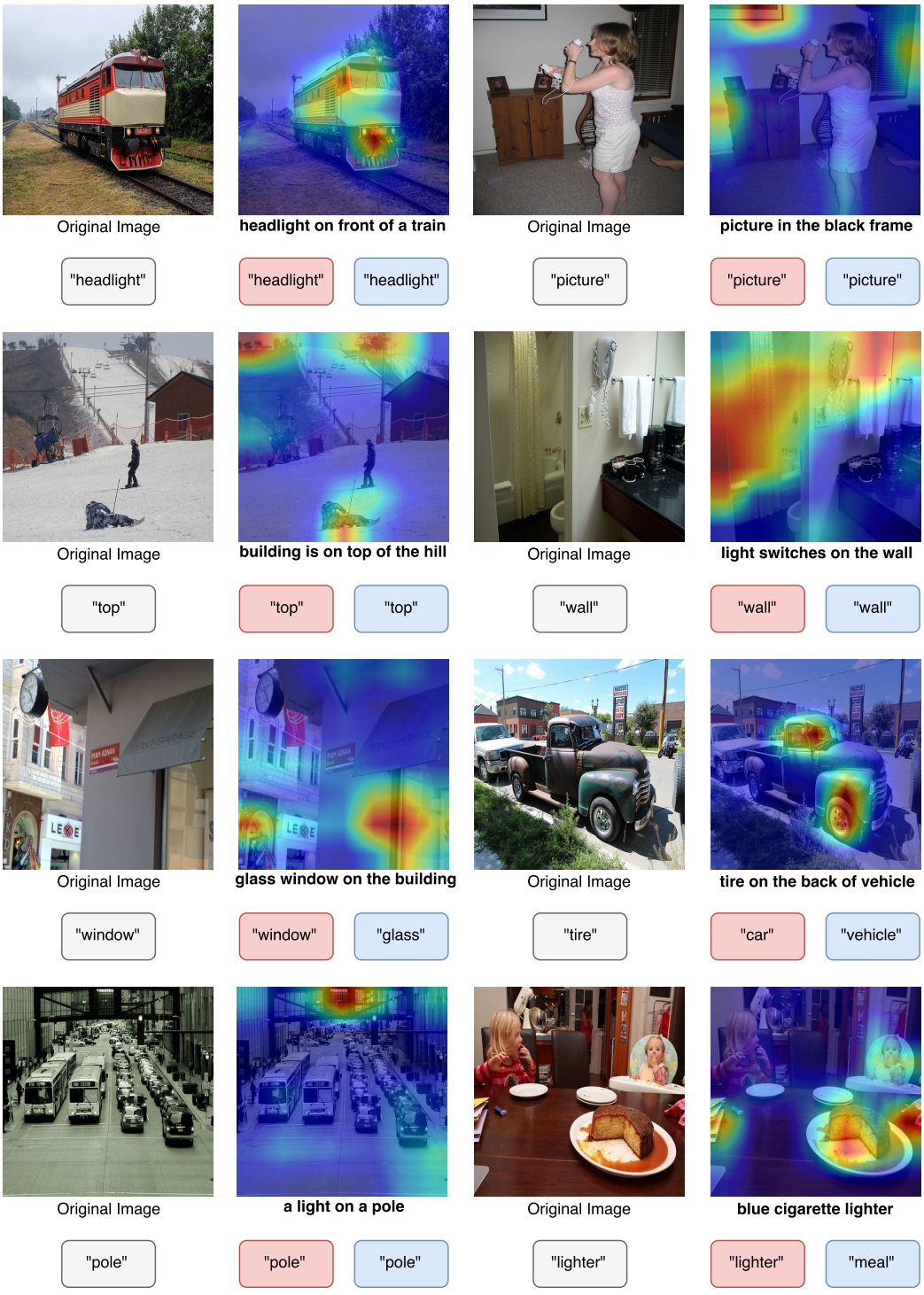

Figure 2: The figure shows how the quality of output heatmap changes with the alignment of the selected concept, predicted concept and the real entity to be grounded. For some sampled concept batch, the **gray box** refers to the chosen common concept, the **red box** refers to the predicted common concept and the **blue box** refers to the predicted independent concept. See section 3 for details about each row.

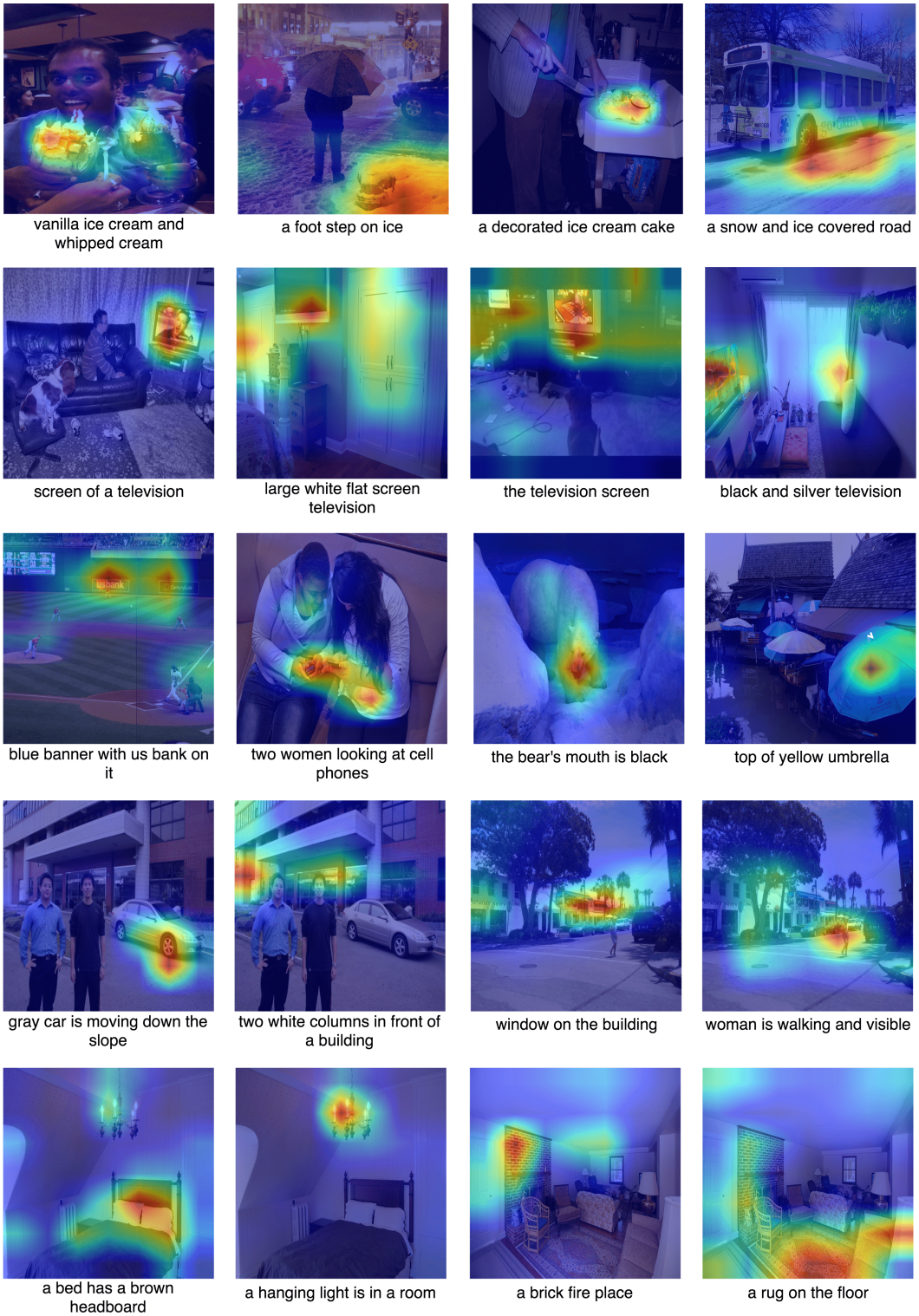

Figure 3: Additional qualitative examples