

# Modularized Textual Grounding for Counterfactual Resilience

In this appendix, we provide details of the Person Attribute Counterfactual Grounding dataset (PACG), and show more visualizations of modular groundings and results on counterfactual queries.

## 1. PACG

To collect counterfactual textual grounding test set, we select 2k images from Flickr 30k Entities Dataset [2] and RefCOCO+ Dataset [1]. We first extracted all attribute words from the ground truth captions as existing attributes. Whereafter, we further manually check and complement all the missing attributes in the images on an interactive user interface. We demonstrate several examples from PACG in Figure 3, and their counterfactual attributes/queries. We adopt person-related semantic attributes and colors on them as the grounding queries. All colors and words from our corpus, that are not related to the person showing up the image, are considered as counterfactual (CF) attributes in the specific image (see Figure 3). To conduct textual grounding evaluations, we substitute the original attribute words in the ground truth captions with CF attributes (semantic-attributes or colors) as our CF queries. We generate all the possible CF queries per image, and more than 20k CF queries are obtained in total.

## 2. Additional Grounding Results from Sub-modules

We show additional grounding outputs from our semantic attribute grounding module ( $M_a$ ) and color grounding module ( $M_c$ ) in Figure 1 and Figure 2. For semantic grounding module, we visualize the attention maps of four most frequent input queries “man”, “woman”, “boy”, and “girl” (see Figure 1). We see clearly from Figure 1 that, our module distinguishes between textual words with different gender and age. It indicates that the semantic attribute grounding module is able to learn the concepts of gender and age through weakly supervised training.

## 3. Additional Grounding Results on Counterfactual Queries

We demonstrate additional intermediate grounding outputs on PACG dataset based on our modularized network in



Figure 1: Examples of our grounding results from the semantic attribute grounding module.

Figure 4. Specifically, besides showing the grounding outputs on ground truth attributes in the left four rows, we also show CF attribute grounding outputs from sub-modules in the last four rows. From the attention maps, we observe that on these examples (1) our model precisely give out the region of ground truth attributes; (2) our model successfully reject the counterfactual queries in these cases.

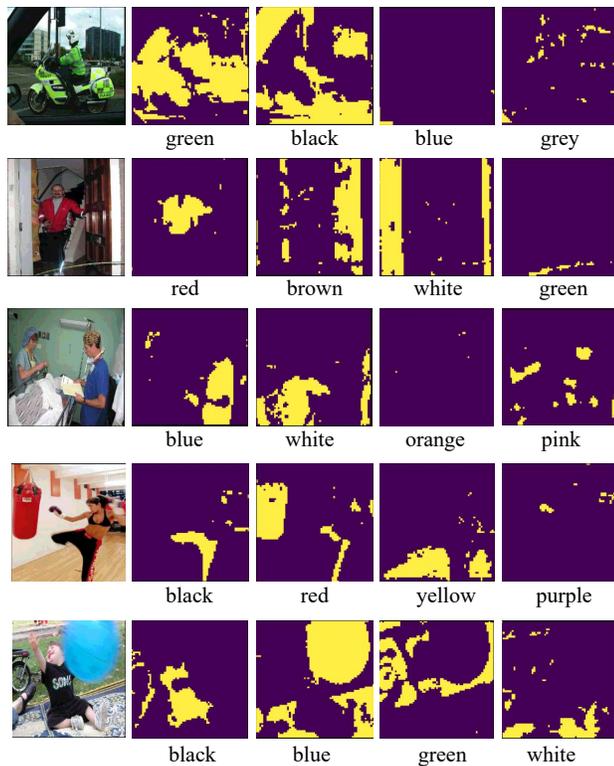


Figure 2: Examples of our grounding results from color grounding module.

## References

- [1] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Refer-itgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1
- [2] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

