Supplementary for Counterfactual Visual Explanations

Yash Goyal [1] Ziyan Wu [2] Jan Ernst [2] Dhruv Batra [1] Devi Parikh [1] Stefan Lee [1]

Abstract

In this supplement we provide additional results on the SHAPES (Andreas et al., 2016) dataset in Section 1.

1. Experiments on SHAPES

Dataset. To first evaluate our model on a simple setting, we created a dataset of SHAPES images (Andreas et al., 2016) for classification using the code released by the authors. This dataset consists of 3x3 grid images of size 30 pixels x 30 pixels. Only one out of the 9 cells contains a shape which can be either a circle, a square or a triangle, which is also the label of the image. Any of these shapes can take any of the three colors – blue, green and red. There is some small random perturbation in the size of each shape and in the pixel values of each color.

Classification model. We trained a simple CNN consisting of 1 convolutional layer followed by 2 fully connected layers with 3 output classes. The network achieves 100% test accuracy, which is unsurprising due to the simplicity of the task.

Experimental settings. For this task, the size of spatial features is 3 x 3 x 100. We randomly choose a distractor class c' different from the predicted class c, and a distractor image I' from the set of images for which the model predicts c'.

Results. Since these images are generated automatically, the cell location containing the shape is known for each image. Hence, the correct discriminative attention maps are known for each pair of (I, I') and the results of our approach can be quantitatively evaluated automatically. We found that approach is able to find the accurate attention maps 100% of the times. An example is shown in Fig. 1.

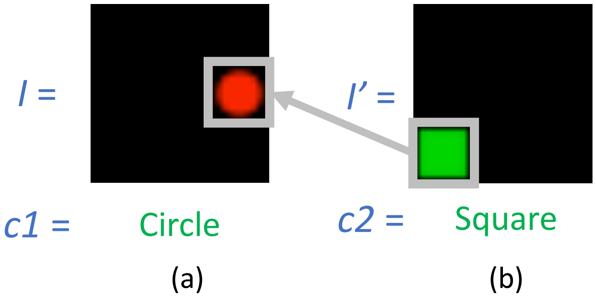

Figure 1. Results on SHAPES images. Each image is made up of 3x3 cells, one of which contains a shape. (a) Our approach highlights the middle right cell in the image I containing the *circle* shape which led the model to predict the class *Circle* instead of class *Square*. (b) In addition, our approach also highlights the bottom left cell containing the *square* shape in image I' of the distractor class *Square* such that if the middle right cell in image I looked like the bottom left cell in image I', the models prediction would have been *Square*.

References

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *CVPR*, 2016.

[1] Georgia Institute of Technology [2] Siemens Corporation. Correspondence to: Yash Goyal <ygoyal@gatech.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).