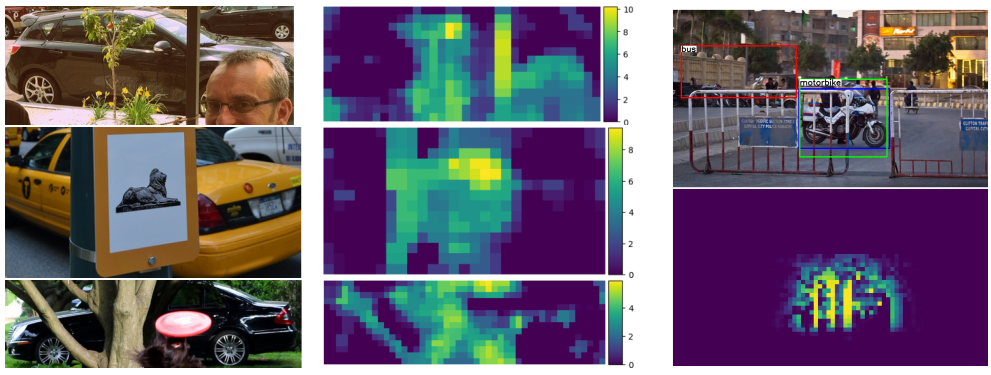
COMPOSITIONAL GENERATIVE NETWORKS: A ROBUST AND INTERPRETABLE MODEL FOR OBJECT RECOGNITION UNDER OCCLUSION

Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, Alan Yuille
Department of Computer Science
Johns Hopkins University

ABSTRACT

Computer vision systems in real-world applications need to be robust to partial occlusion. In this work, we show that black-box deep convolutional neural networks (DCNNs) have only limited robustness to partial occlusion. We overcome these limitations by unifying DCNNs with part-based models into Compositional Generative Networks (CGNs) - a deep architecture with innate robustness to partial occlusion. Specifically, we propose to replace the fully connected classification head of DCNNs with a differentiable compositional model that can be trained end-to-end. The structure of the compositional model enables CompositionalNets to decompose images into objects and context, as well as to further decompose object representations in terms of individual parts and the objects' pose. The generative nature of our compositional model enables it to localize occluders and to recognize objects based on their non-occluded parts. Our experiments in terms of image classification and object detection on images of partially occluded vehicles show that CGNs made from several popular DCNN backbones (VGG-16, ResNet50, ResNext) improve by a large margin over their non-compositional counterparts. Furthermore, they can localize occluders accurately despite being trained with class-level supervision only. Finally, we demonstrate that CompositionalNets provide human interpretable predictions as their individual components can be understood as detecting parts and estimating an objects' viewpoint.



(a) (b)

Figure 1: Example of images for the classification (a) and detection (b) of partially occluded objects from the MS-COCO dataset. A standard DCNN misclassifies the images in (a) and does not detect the motorbike in (b), while also having a false-positive detection of a bus in the background. In contrast, Compositional Generative Networks (CGNs) provide correct predictions in all cases. Intuitively, a CGN can localize the occluders (see visualization of occlusion scores) and subsequently focus on the non-occluded parts of the object to make a robust prediction.