Deep Neural Networks Rely on Distinct Semantic Features of Same-Category Examplars Not Predicted By Low-Level Image Statistics

MohammadHossein NikiMaleki[1], Hamid Karimi-Rouzbahani[2,3]

[1]Faculty of Computer Science and Engineering, Shahid Beheshti University, Iran

[2]Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, UK

[3]Department of Computing, Macquarie University, Australia

Deep Convolutional Neural Networks (DCNNs) are among the most accurate and brain-plausible models of human object recognition. It has been shown that humans rely on specific segments of objects (called minimal recognizable configurations or MIRCs) for recognition. However, DCNNs did not show such sensitivity to identical MIRCs (Ullman et al., 2016). Therefore, it remains unclear if humans and DCNNs use different mechanisms for object recognition. Specifically, we have shown previously that while humans used relatively consistent/invariant sets of object features across variations (in-depth and in-plane rotation, size and translation), DCNNs relied on relatively inconsistent/distinct object features across the variations of the same objects (Karimi-Rouzbahani et al., 2017). This suggests that, as opposed to humans, DCNNs seem to rely on semantically distinct object features across object variations for recognition. This might be a more general mechanism suggesting that DCNNs may even use relatively more distinct object features to recognize the examplars from the same semantic object category (e.g. different examplars of an elephant), compared to humans. To test this hypothesis, we obtained MIRCs for one of the most brain-like DCNNs (VGG16) using the well-established Bubbles method (Gosselin and Schyns, 2001). As an advantage to previous procedures, which detected MIRCs from pre-selected discrete image parts, Bubbles sweeps the whole image using continuous masks, allowing data-driven contribution of all pixels to recognition. We extracted MIRCs from 12 semantic object categories (e.g. elephant, hammer, pot, etc., each with 16 examplars) of the ImageNet dataset (Deng et al., 2009). Results clearly showed different MIRCs for distinct examplars of the same object category, reflecting the exemplar-specific nature of feature selection in DCNNs. This may underlie the robust object recognition observed for DCNNs under variations in objects and examplers. To provide a mechanistic account of how feature selection might happen in DCNNs, we then asked if the MIRCs found for DCNNs could be predicted by low-level image statistics. Specifically, we wondered if the MIRCs were simply salient segments of an image as detected by computational models of saliency. These models use local low-level image statistics (e.g. color, orientation, contrast) to predict the location of human overt attention (gaze) on the image (Kimura et al.,

2013), and can indicate image areas that are visually rather than semantically distinct from other areas. Alternatively, MIRCs could be object segments which potentially contain semantic information which diagnoses the category of the object. To test this hypothesis, we obtained the salient segments of all the images in our dataset using 5 of the most brain-plausible saliency-based models e.g. Itti et al., 1998. Results showed that the MIRCs obtained from the DCNN and the salient regions obtained from the saliency models were quantitatively and qualitatively different. This suggests that, rather than relying on salient low-level image statistics, DCNNs may rely on object segments which probably contain semantic category information relevant for object recognition. We are collecting human data to quantitatively compare to the results from our DCNN and the computational models of attention.

References:

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

Gosselin, F. and Schyns, P.G., 2001. Bubbles: a technique to reveal the use of information in recognition tasks. Vision research, 41(17), pp.2261-2271.

Itti, L., Koch, C. and Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 20(11), pp.1254-1259.

Karimi-Rouzbahani, H., Bagheri, N. and Ebrahimpour, R., 2017. Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. Scientific reports, 7(1), pp.1-24.

Kimura, A., Yonetani, R. and Hirayama, T., 2013. Computational models of human visual attention and their implementations: A survey. IEICE TRANSACTIONS on Information and Systems, 96(3), pp.562-578.

Ullman, S., Assif, L., Fetaya, E. and Harari, D., 2016. Atoms of recognition in human and computer vision. Proceedings of the National Academy of Sciences, 113(10), pp.2744-2749.