Emergence of Topographical Correspondences between Deep Neural Network and Human Ventral Visual Cortex

Yalda Mohsenzadeh* (yalda@mit.edu)
Computer Science & AI Lab., 32 Vassar Street
Cambridge, MA 02139 USA

Caitlin Mullin* (crmullin@mit.edu)
Computer Science & AI Lab., 32 Vassar Street
Cambridge, MA 02139 USA

Dimitrios Pantazis (pantazis@mit.edu)
McGovern Institute for Brain Research, 43 Vassar Street
Cambridge, MA 02139 USA

Aude Oliva (oliva@mit.edu)
Computer Science & AI Lab., 32 Vassar Street
Cambridge, MA 02139 USA

Abstract:

Recent computer vision work dissecting information from within the layers of deep neural networks revealed emergence of human-interpretable concepts within these artificial units. In the current study, using representational similarity analysis, we compare convolutional layers of DNNs trained for object and scene recognition (hybrid AlexNet) with regions along ventral visual pathway to ask whether these layers and regions share topographical correspondence. Results reveal the emergence of a brain inspired topographical organization in this hybrid-net, such that layer-units showing strong central-bias were associated with cortical regions with foveal tendencies, and layer-units showing greater selectivity for image boundaries and backgrounds were associated with cortical regions showing strong peripheral preference. The emergence of a categorical topographical correspondence between deepnets and visual regions of interests further strengthens the role of deepnets as models of the inner workings of perceptual networks in the brain.

Keywords: Neural categorical representations; deep convolutional neural networks; representational similarity analysis; fMRI; topographical maps

Introduction

Decades of human neuroscience research has revealed that distinct cortical regions are differentially activated by separate image categories. Moreover, these category selective regions demonstrate an organizing principle such that some show a central visual field bias while others a peripheral bias (Levy et al., 2001).

Recent works at the intersection of computer vision and neuroscience have suggested that artificial visual systems, such as deepnets, learn and interpret visual features along a hierarchy much like the human visual system, with different hidden units within the network spontaneously learning representations of features that guide the accuracy of the output (Khaligh-Razavi et al., 2014; Yamins et al., 2014; Cichy et al., 2016; Bau et al., 2017). For instance, deepnets trained on scene categorization showed the spontaneous emergence of object representations in certain layers of the network (Zhou et al., 2014; Bau et al., 2017).

Given that distinct visual categories are mapped according to a center/periphery rule in the human visual system, we asked whether deepnets might also spontaneously learn this topographical organization. To test this hypothesis, we probed the individual layers of a hybrid network trained on both object and scene categorization with representations pulled from several levels of the visual hierarchy in the human brain. We predict that layer-units associated with object emergence will show a strong central-bias, while layer units associated with global representations will demonstrate peripheral selectivity.

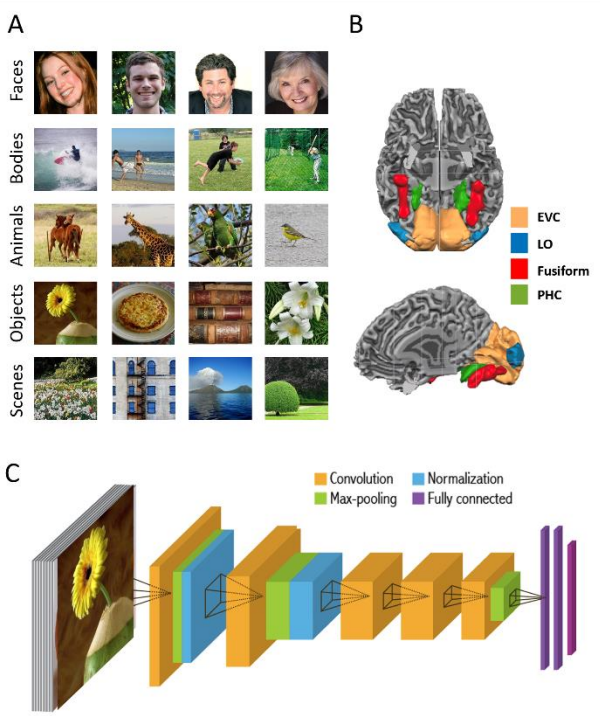

Figure 1: A) Examples of stimuli in five categories. B) Regions of interest along the ventral visual pathway. C) A deep neural network model trained both on imageNet and Places datasets (hybrid-AlexNet).

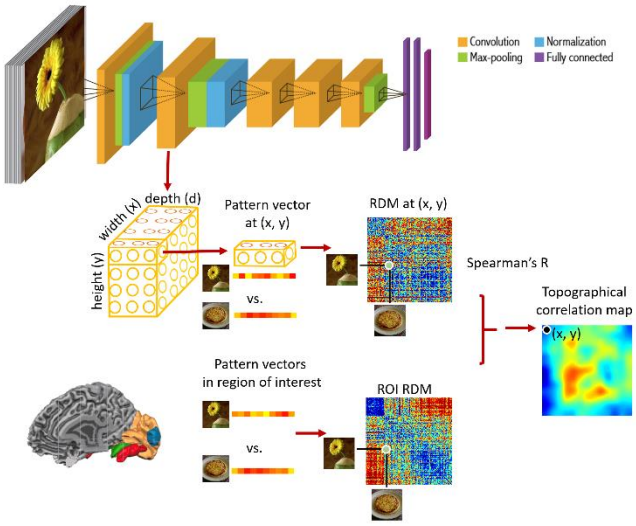

Figure 2: Topographical maps. DNN layer RDM matrices are computed by pairwise comparison of the network activation patterns extracted at (x,y) location in the feature map for each image condition. The ROI RDM matrices are computed by pairwise comparison of t-value pattern vectors in that area. Comparison of neural network RDMs at (x,y) position with the brain ROI RDM by computing Spearman's Rho correlation yield a topographical correlation map which is then up-sampled to the image size.

Method

Experiment Design and Stimulus Set

To study the categorical representations in the human brain we selected 156 natural images organized in 5 categories (faces, bodies, animals, objects and scenes). Examples of stimuli are shown in Figure 1A. We collected fMRI data while participants (N=16) viewed these images presented at the center of the screen at 6° visual angle for the duration of 0.5s with 2.5s interstimulus intervals and performed an orthogonal task (detecting a color change in the fixation cross). The participants completed two fMRI sessions of 5-8 runs each (11-15 runs over the two sessions). Each image was presented once per run in randomized order.

Convolutional Neural Network Architecture and Training

To investigate the correspondence of topography in category-specific cortical regions and deep convolutional neural network, we compared fMRI data extracted from regions of interest along ventral visual stream with a DNN with AlexNet architecture (Figure1C) trained both on object and scene image categories (*hybrid-AlexNet*, Zhou et al., 2014).

Brain and DNN Topographical Maps

To compare category-specific neural and computational model representations we used representational similarity analysis (Kriegeskorte et al., 2008). We defined anatomically four regions of interest (ROIs) along the ventral stream, early visual cortex (EVC), lateral occipital area (LO), fusiform area, and parahippocampal area (PHC) (see Figure1B).

In each ROI and for each of 156 image conditions we extracted the t-value activation patterns, arranged them into vector patterns, and then computed the pairwise dissimilarity of these 156 vector patterns by calculating 1 minus Pearson correlations. This yielded a 156×156 representational dissimilarity matrix (RDM) for each subject and ROI.

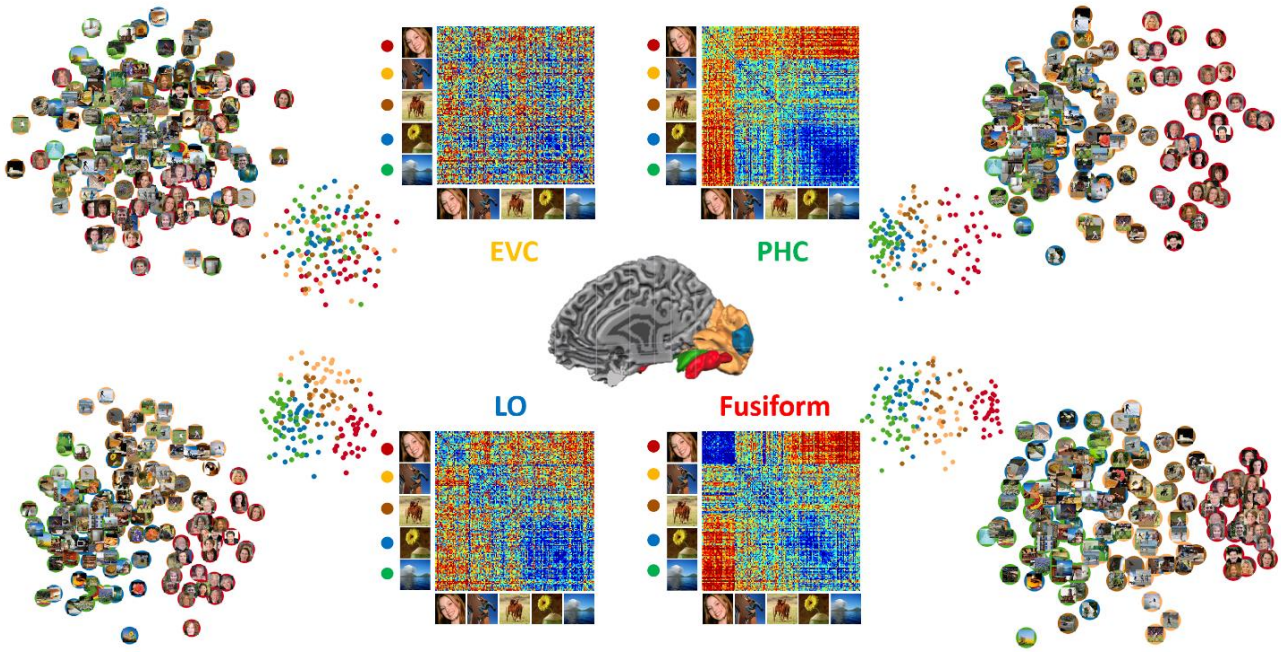

Figure 3: Neural representations along ventral visual pathway. RDM matrices, and 2D multidimensional scaling visualization of stimuli depicted for early visual cortex (EVC), lateral occipital area (LO), parahippocampal area (PHC) and fusiform area.

To create the topographical maps, we fed the images to the deep network and extracted the 3D activation patterns from its convolutional layers. For each entered image the first 2 Dimensions have a spatial relation with the image space (width and height). As illustrated in Figure 2, at each (x,y) position in feature maps, we extracted a pattern vector with the length equivalent to its depth and constructed the RDM matrix from the neural network activity patterns at each (x,y) location. Comparison of these RDM matrices of each layer with a brain ROI RDM results in a 2D correlation map which we then up-sample it to the image size and call it a topographical map. This analysis yields a topographical map for comparison of each convolutional layer and each brain region.

Results and Discussion

The averaged ROI RDMs and their 2D multidimensional scaling visualizations are shown in Figure 3. As expected, EVC shows a random pattern, LO depicts a clear animate/inanimate distinction, fusiform clusters face images strongly and PHC groups together scene images.

Figure 4 shows the topographical correlation maps of the five convolutional layers of the hybrid-AlexNet model with the four fMRI ROIs. EVC shows a random topographical correlation map pattern in the first two layers. This is in line with previous studies showing earlier layers of network being significantly correlated with EVC. Our results further show topographically these low level features scattered over the image. LO shows a dispersed correlation map in the mid-level layers and becomes more and more centralized over the layers, depicting a mid to high-level representation transformation. Correlation maps with PHC show a background/surrounding organization in layers 1 to 4 resulting into a scattered distributed representation in the very last layer. This shows while the peripheral units of the network have similar representation to PHC in the earlier and mid-level layers, in the last layer they capture more distributed zones in the image, more adapted to scene representations. Finally, correlation maps of fusiform and layers of the network demonstrates very strong center-selective patterns, consistent with the foveal-bias representations in fusiform area.

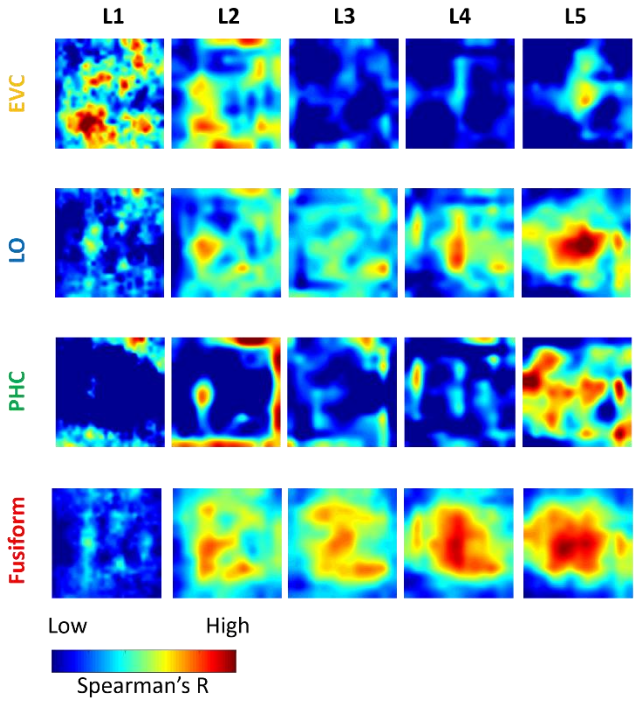

Figure 4: Topographical correspondence between convolutional layers of deepnets and human ventral visual regions. Each row shows the correlation maps for each brain ROI (EVC, LO, PHC and Fusiform) and each column corresponds to a convolutional layer in the hybrid-AlexNet.

Conclusion

Previous studies have shown hierarchical and temporal correspondences between the regions in ventral visual pathway and layers of deepnets (Khaligh-Razavi et al., 2014; Yamins et al., 2014; Cichy et al., 2016). In the current study, using RSA analysis, we showed a topographical correspondence between the brain regions and units of the network. Specifically, fovealy biased fusiform highly correlated with units of the network selective to the center of the visual field and peripherally biased PHC strongly correlated with units of the network selective to the background/surrounding of the image. The hierarchical, temporal and topographical correspondences between deepnets and visual cortex, further motivate the use of deepnets as relevant models of the visual ventral stream.

Acknowledgments

Funding from the Vannevar Bush Faculty Fellowship program by the ONR to A.O. (N00014-16-1-3116). Study conducted at the Athinoula A. Martinos Imaging Center, MIBR, MIT.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network Dissection: Quantifying the Intepretability of Deep Visual Representations. In *Computer Vision and Pattern Recognition*.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6, 27755.

Khaligh-Razavi, S., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *Plos Computational Biology*, 10(11): e1003915. doi:10.1371/journal.pcbi.1003915.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2(November), 4-4. doi:10.3389/neuro.06.004.2008.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature neuroscience*, 4(5), 533.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619-8624.

Zhou, B., Lapedriza , A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning Deep Features for Scene Recognition using Places Database. *NIPS*.