The role of textural statistics vs. outer contours in deep CNN and neural responses to objects

Bria Long (bria@stanford.edu)
Department of Psychology, Stanford University, 450 Serra Mall
Stanford, CA 94305 USA

Talia Konkle (tkonkle@fas.harvard.edu)
Department of Psychology, Harvard University, 33 Kirkland Street
Cambridge, MA 02140 USA

Abstract:

Deep convolutional neural networks (CNNs) are providing new insight into the high-dimensional feature space that supports object representations in the ventral stream. Here, we examined what specific visual features underlie deep CNN's ability to predict occipitotemporal cortex responses to images of animals and objects of different sizes. To do so, we measured activations from a widely-used convolutional neural network (Krizhevsky et al., 2012) to four variants of the same image set: (i) original images, (ii) silhouetted images, (iii) phase-scrambled images, and (iv) texforms images (which preserve a combination of texture and coarse form; Long, Yu, & Konkle, 2017). We found that the predictive power of CNN features in the ventral stream was better accounted for by textural rather than outer contour properties. These results point towards textural statistics as an important dimension in characterizing the representational layout of object representations in object selective cortex, and underscore the importance of controlled image sets for examining when and why deep CNN features hold predictive power.

Keywords: occipitotemporal cortex; visual features; CNNs

Introduction

At present, deep convolutional neural network activations are the best predictors of object responses in object selective cortex (Khaligh-Razavi & Kriegeskorte, 2014; Güclü & van Gerven, 2015; Long, Yu, & Konkle, 2017; Yamins et al., 2014). However, we still have a relatively impoverished understanding of what features underlie their predictive success.

Recently, we leveraged a new class of stimuli—*texforms*, to explore the nature of the feature tuning in object selective cortex (see Figure 1; Long, Yu, & Konkle, 2017). We found that texforms and recognizable images yielded similar patterns of neural activity across occipitotemporal cortex (OTC) for both the animacy and object size distinctions. This result suggests that mid-level features underlie a substantial portion of neural responses to objects. However, as texforms preserve both coarse form and texture information, this work leaves open their relative contributions to object responses.

Thus, in the present work, we examined the role of texture vs. outer contour information using a method developed by Bonner & Epstein (2018). Specifically, here we examined the extent to which deep CNN activations to recognizable images might reflect textural vs. outer contour properties by testing whether deep CCN activations to texforms (Long, Yu, & Konkle, 2017) silhouettes, and phase-scrambled images also explain similar variance.

Methods

Stimuli Each variant of the stimulus set was constructed from an original set consisting of 30 big objects, 30 small objects, 30 big animals, and 30 small animals. Recognizable images and their texform counterparts were the same images used in Long, Yu, & Konkle, 2017. Silhouettes and phase scrambled images were generated using custom scripts in Matlab 2017a.

Multi-voxel patterns in OTC Neural patterns were taken from a pre-existing dataset (Long, Yu, & Konkle, 2017), and consisted of activations across occipito temporal cortex to 24 conditions: animals/objects x big/small real-world size x 6 mini groups. (The images were grouped into 6 sets of 5 images, based on how well their texform counterparts were classifiable by animacy and size in the real-world; see Long, Yu, & Konkle, 2017).

Feature extraction Feature activations were extracted using the neural network toolbox implemented in Matlab 2017a and the standard AlexNet architecture (Krizhevsky et al., 2012); the network was not fine-tuned for any of the image sets. For each image and each convolutional filter, the summed activation map

of the filter was computed. In order to compare these feature activations with the occipitotemporal responses, these activation maps were averaged across all five images in each mini group; RDMs were then constructed by computing the correlation distance between activation vectors for each set, allowing direct comparison with neural RDMs.

Shared variance analyses Following Bonner & Epstein (2018), we combined the standard RSA approach with commonality analysis (Nimon & Oswald, 2013). This analysis estimates the portion of the shared variance between OTC and CNN activations to originals images that can be accounted for by CNN activations to other kinds of image sets. Formally, this entails conducting multiple linear regressions using ordinary least squares with CNN activations to two sets of images together (e.g., originals and texforms) as well as separately for each image set, and then using a variance partitioning procedure to estimate the shared variance. See Bonner & Epstein (2018) for further details.

Results

The representational dissimilarity matrix (RDM) for recognizable images in occipitotemporal cortex is shown in Figure 1. A relatively strong block-diagonal structure can be seen, with an overall division between animals and objects, and a smaller subdivision (in the upper left) between big and small objects. This structure is also visible in the multi-dimensional scaling plot, which shows a tripartite division between all animals, big inanimate objects, and small inanimate objects, replicating prior work (Konkle & Caramazza, 2013).

Next, we examined the RDMs computed from the activation in the last convolutional layer to each of the stimulus set variants (Figure 1). Doing so enables us to examine the degree to which they recapitulate the OTC structure in the absence of any reweighting of their features. While the model RDMs for texforms and originals appear relatively similar, the model RDMs for silhouettes vs. phase-scrambled images differ in systematic ways. For example, phase scrambled big

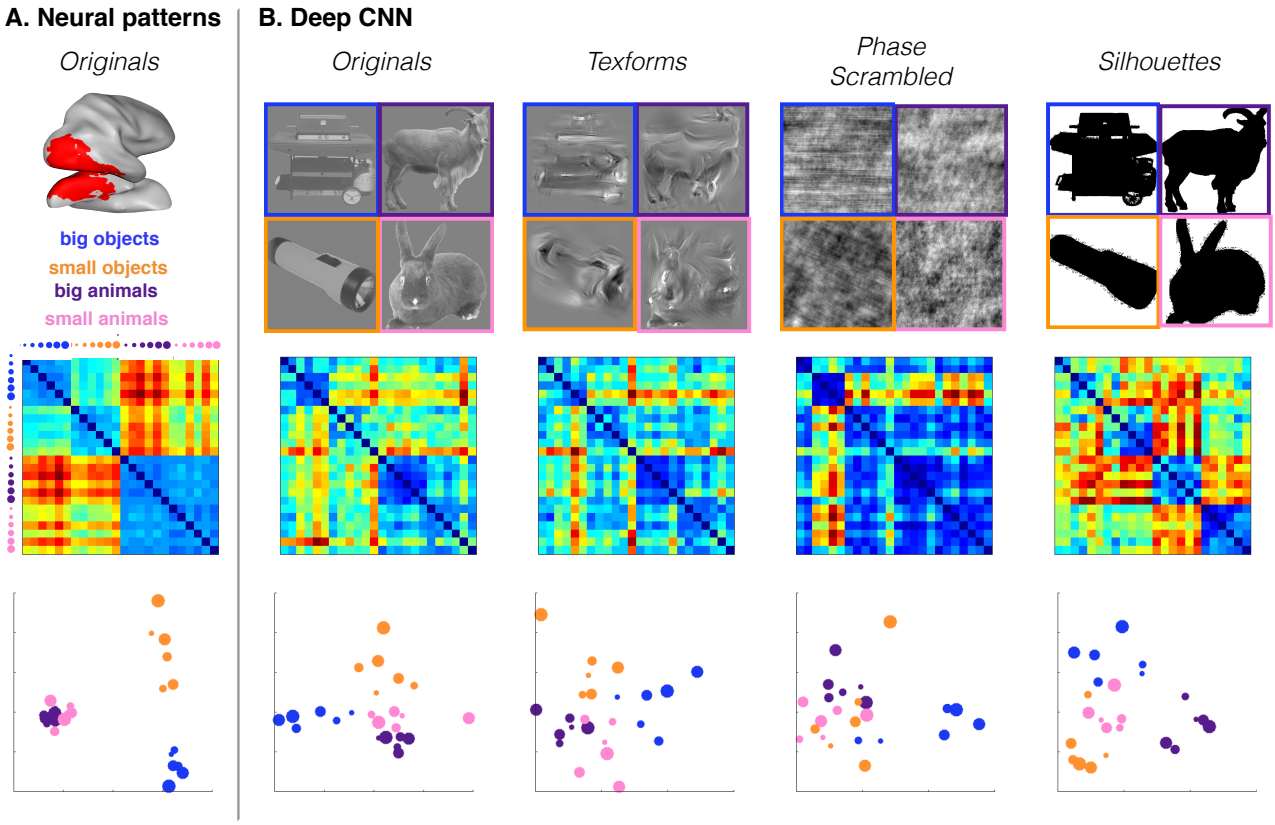

Figure 1. (Left panel). Neural patterns in occipitotemporal cortex to images of animals and objects of different sizes, and a multi-dimensional scaling visualization of this representational layout. (Right panel). The top row shows example stimuli from the four different image sets. The middle row shows the representational dissimilarity matrix in response to these stimuli sets in the last convolutional layer (Conv5) of AlexNet; all RDMs are scaled. The bottom rows shows multidimensional scaling plots of these RDMs; dots are scaled according to how well the texforms in each group were classifiable by their animacy and object size.

objects tended to cluster together, likely owing to greater power at vertical/horizontal spatial frequencies. In contrast, silhouettes of big animals strongly clustered together, suggesting that big animals may be likely to have more similar contours than small animals, even though this distinction is not strongly evident in the neural similarity in OTC.

Next, we examined how activations from these different image sets explained the predictive power of the original-CNN features in each layer of the model (see Figure 2). Texforms, which consist of both texture and coarse form information present in original images, accounted for a substantial part of the predictive power. Interestingly, even more primitive phase-scrambled images accounted for much of the predictive power in early layers. In contrast, silhouette content had relatively little predictive power in early layers, reaching an equivalent proportion by the later layers. Finally, these results should also be interpreted keeping in mind that the overall amount of variance accounted for by the original CNN features is different for each layer (see light gray numbers under each plot). Though earlier layer features have non-zero predictive power in occipitotemporal cortex, later convolutional layer features are better predictors.

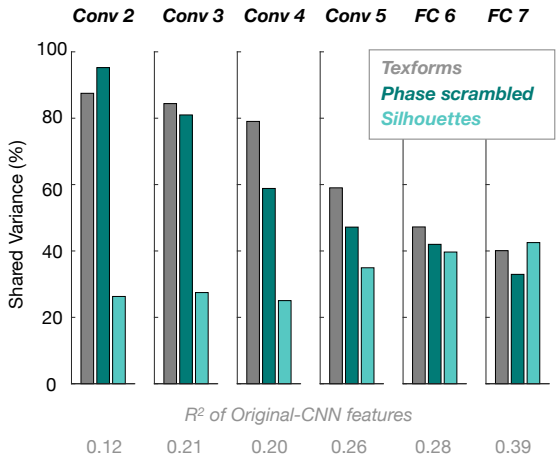

Figure 2. The shared variance between activations to each image set and originals in explaining OTC patterns is plotted for convolutional layers 2-5 and the last two fully connected layers. The degree to which original CNN features from each layer predicted the neural patterns is shown in grey below each layer plot; note that conv 1 features are not analyzed here as they yielded an r-squared < .01.

Discussion

Overall, we found that CNN features derived from texture statistics explained a substantial portion of the predictive power in OTC of the features derived from recognizable images, especially in early and intermediate layers. In contrast, silhouettes held relatively less predictive power, only becoming more predictive than features from phase-scrambled images in the latest layer of the network. In general, silhouettes seemed capture image features that do not seem to be differentiate neural patterns in occipitotemporal cortex, despite being relatively recognizable.

Thus, these results suggest a relatively strong contribution of textural statistics in explaining the predictive power of deep CNN features—here with respect to the large-scale distinctions by animacy and object size. An important future avenue for this work is to understand how well these results generalize when observers are viewing different kinds of images (e.g., scenes or videos) or when observers are performing more complex tasks. We propose that comparing the relative predictive power of deep CNN activations to various kinds of transformed image sets may help us understand when and why deep CNN features predict patterns of neural activity or visual behavior.

Acknowledgments

This work was funded by a Harvard Star Family grant to T.K., an NSF SPRF-FR Grant #1714726 to B.L, and a NIH Shared Instrumentation Grant (S10OD020039) to the Harvard Center for Brain Science.

All code and stimuli for these analyses are available at: www.github.com/brialorelle/cnn-features

References

Bonner M.F. & Epstein R.A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology* 14(4): e1006111

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915.

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242.

Krizhevsky, A., Sulskever, Ii., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 1–9.

Long, B., Yu., C.P., & Konkle, T. (2017). A mid-level organization of the ventral stream. *bioRxiv* doi: 10.1101/213934

Nimon K.F., Oswald F.L. (2013) Understanding the Results of Multiple Linear Regression. *Organizational Research Methods.*

Yamins, D. L. K., 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 of the United States of America*, *111*(23).