Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations

Kamila Maria Jozwik (kmjozwik@mit.edu)
Massachusetts Institute of Technology and University of Cambridge, McGovern Institute for Brain Research, 43 Vassar St
Cambridge, MA 02139 United States

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Columbia University, Zuckerman Institute, 3227 Broadway
New York, NY 10027 United States

Radoslaw Martin Cichy (rmcichy@zedat.fu-berlin.de)
Free University Berlin, Department of Education and Psychology, 45 Habelschwerdter Allee
Berlin, 14195 Germany

Marieke Mur (marieke.mur@mrc-cbu.cam.ac.uk)
University of Cambridge, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road
Cambridge, CB2 7EF United Kingdom

Abstract:

Deep convolutional neural networks (DNNs) are currently the best computational model for explaining image representations across the visual cortical hierarchy. However, it is unclear how the representations in DNNs relate to those of simpler "oracle" models of features and categories. We obtained DNN (AlexNet) representations for a set of 92 real-world object images. Human observers generated category and feature labels for the images. Category labels included subordinate, basic and superordinate categories; feature labels included object parts, colors, textures, and contours. We used the AlexNet representations and labels to explain brain representations of the images, measured with fMRI in humans and cell recordings in monkeys. For both human and monkey inferior temporal (IT) cortex, late AlexNet layers perform similarly to basic categories and object parts. Furthermore, late AlexNet layers can account for more than half of the variance that these labels explain in IT. Finally, while feature and category models predominantly explain image representations in high-level visual cortex, AlexNet layers explain representations across the entire visual cortical hierarchy. DNNs may provide a computationally explicit model of how features and categories are computed by the brain.

Keywords: object vision; primate; features; categories; DNN

Introduction

The best models for explaining responses in primate high-level visual cortex have for long been "oracle" models. Oracle models consist of object labels generated by human observers, and thus leave open how the visual system computes the labels. However, in recent years, deep convolutional neural networks (DNNs) have revolutionized computer vision, reaching human-level performance on object classification. Like the visual system, DNNs learn representations of rich inputs, such as colored real-world object images. DNNs predict representations of object images in visual cortex, as measured in humans via fMRI (Khaligh-Razavi & Kriegeskorte, 2014) and in monkeys via electrophysiology (Yamins et al., 2014). These findings suggest that there are considerable similarities between DNN and brain representations of objects. However, even DNNs capable of near-human-level object classification performance classify certain images in highly counterintuitive ways (e.g. a leopard pattern sofa as a leopard), calling into question their commonalities with brain representations.

To investigate the commonalities of DNNs with brain representations, we use a well-known DNN (AlexNet) to predict brain representations of object images and compare its performance to that of oracle models consisting of feature and category labels. Does AlexNet explain the brain data as well as the oracle models do? Do the two types of models explain the same variance?

Methods

Stimuli

Stimuli were 92 colored images of real-world objects spanning a range of categories, including humans, non-human animals, natural objects, and artificial objects. Objects were segmented from their backgrounds and presented on a gray background.

Monkey Single-Unit Recordings

Macaque monkeys (n=2) viewed the images while single-unit responses were recorded from anterior IT cortex (674 neurons, data from Kiani, Esteky, Mirpour, & Tanaka, 2007). Monkeys performed a fixation task. Images were presented at the center of fixation (size: 7° visual angle, stimulus duration: 105 ms). Spike rates were averaged within a 140 ms window (71-210 ms after stimulus onset).

Human fMRI

Subjects (n=15) viewed the images while their brain activity was measured with a 3T fMRI scanner (GE EPI, TR: 2s, voxel resolution: 2 mm^3, data from Cichy, Pantazis, & Oliva, 2014). Subjects performed a fixation task. Images were presented at the center of fixation (size: 2.9° visual angle, stimulus duration: 500 ms). Regions of interest (ROIs) were defined in each subject. V1 was defined using an anatomical eccentricity template (361 voxels on average), IT was defined using an anatomical mask of bilateral fusiform and inferior temporal cortex (361 most strongly activated voxels). We also performed a volume-based searchlight analysis in each subject (radius of 4 voxels, Spearman's r, two-sided Wilcoxon signed-rank test, FDR controlled at 0.05).

Oracle Models

Human observers generated visual feature labels (e.g., "eye") and category labels (e.g., "animal") for the images (Jozwik, Kriegeskorte, & Mur, 2016). Feature labels were divided into parts, colors, textures and contours, while category labels were divided into subordinate categories, basic categories and superordinate categories. The final full feature and category models consisted of 119 and 110 labels, respectively.

Deep Neural Network (AlexNet)

We computed activations for the images in each layer of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). AlexNet was trained on the ImageNet database to classify images into 1,000 categories. We used convolutional (conv) and fully-connected (fc) layers in our analyses.

Comparing Model Performance

We computed response patterns (across neurons, voxels, features, categories, and units within AlexNet layers) for each image. We then computed response-pattern dissimilarities between images and placed these in a representational dissimilarity matrix (RDM). An RDM captures which distinctions among stimuli are emphasized and which are de-emphasized by a particular model or brain region. We estimated model performance by correlating model and data RDMs using Kendall's rank correlation coefficient tau a. We determined whether each of the model RDMs was significantly related to the data RDMs using a stimulus-label randomization test (10,000 randomizations) for the monkey data and a subject-as-random-effect analysis for the human data (one-sided Wilcoxon signed-rank test). We subsequently tested for differences in model performance between each pair of models using bootstrap resampling of the stimuli (1,000 resamplings) for the monkey data and a subject-as-random-effect analysis for the human data (two-sided Wilcoxon signed-rank test). For each analysis, we accounted for multiple comparisons by controlling the FDR at 0.05.

Results

Comparison of Oracle and AlexNet Model Performance at Explaining Primate IT

We have previously shown that both features and categories explain the human IT object representation well (Jozwik et al., 2016). Do late AlexNet layers explain human IT as well as these models derived from human perception? To evaluate this, we compared performance at explaining human IT between AlexNet layers and feature and category models. We also tested a model that combined all feature and category labels together ("features and categories"). Results indicate that late AlexNet layers (layers 6 and 7) perform at a level similar to basic and superordinate categories and object parts, but outperform the other feature models (color, contour, texture) (Figure 1BC). Performance of AlexNet layer 1 is in a similar range as that of the color, contour and texture models. Results for monkey IT (Figure 1AC) are largely consistent with those for human IT. In contrast, for human V1, earlier AlexNet layers 2 and 3 perform best, outperforming all feature and category models.

Similarity of Object Representations in Oracle Models and AlexNet Layers

Similar model performance does not necessarily indicate that models explain the same variance. To address this issue, we first examined the degree of similarity between the object representations in the two types of models by correlating every oracle model with every AlexNet layer (Figure 2). Given the proximity of late AlexNet layers to the final category readout, we might expect the object representation in late layers to match the object representation in category models. Indeed, late AlexNet layers (layers 6-8) appear to correlate more strongly with the category models than earlier AlexNet layers (layers 1-3) do. This progression from early to late layers is weaker but also visible for the object-part and all-features models. These observations suggest that late AlexNet layers should be able to account for a substantial proportion of the variance that oracle models explain in IT. Indeed, we found that late AlexNet layers 6 and 7 each can account for approximately half of the IT variance explained by object parts, and for approximately sixty percent of the IT variance explained by categories (this holds for all categories except subordinate categories).

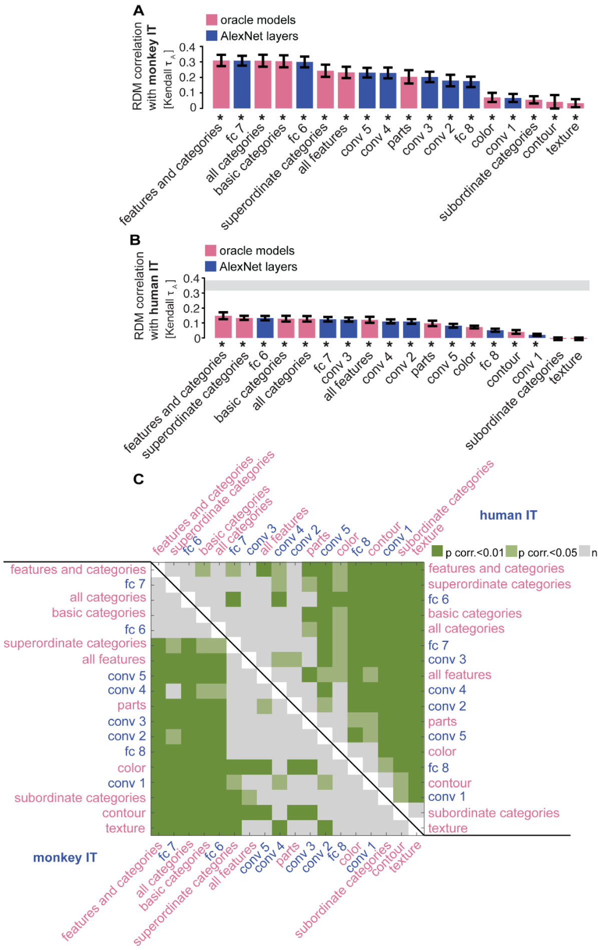

Figure 1. AlexNet and oracle model performance at explaining monkey and human IT representations. (A) Bars show the correlation between the monkey IT RDM and each model RDM. A significant correlation between a model RDM and the IT RDM is indicated by an asterisk (stimulus-label randomization test, FDR controlled at 0.05). Error bars are based on bootstrap resampling of the stimuli. "conv" indicates a convolutional layer and "fc" indicates a fully-connected layer. (B) Bars show the correlation between the human IT RDMs and each model RDM, using the same conventions as in Figure 1A, but using subject-as-random-effect analyses for inference and error bars. The gray bar represents the noise ceiling, indicating the expected performance of the true model given the noise in the data. (C) Pairwise differences between model performance for monkey and human IT. Green color indicates significant pairwise differences based on bootstrap resampling of the stimuli (monkey) or subject-as-random-effect analysis (human) (dark green: FDR < 0.01, light green: FDR < 0.05).

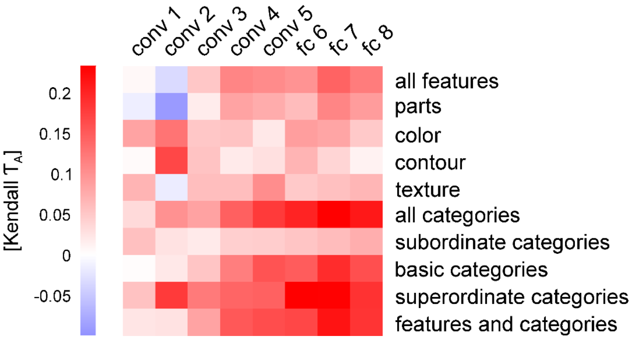

Figure 2. Correlations between AlexNet and oracle model RDMs.

Comparison of Oracle and AlexNet Model Searchlight Maps

To complement our ROI analysis, we performed a searchlight analysis on the human fMRI data, testing where else in the brain AlexNet layers and oracle models explain image representations (Spearman's r, two-sided Wilcoxon signed-rank test, FDR controlled at 0.05). Results are shown in Figure 3. Visual inspection of the results reveals that there is little correspondence between the layer 1 object representation and the brain representations; the layer 2 and 3 object representations explain the brain representation in similar locations as colors; the layer 6 and 7 representations explain the brain representation in similar locations as superordinate categories. High-level visual cortex representations that are explained well by both the category and feature models were best captured by layers 7 and 8. One difference between the representations of oracle and AlexNet models is that for most feature and category models (except color and superordinate categories) the location of the representations is confined to high-level visual cortex, suggesting that oracle model representations are already quite complex. For almost all AlexNet layers (with the exception of layer 1 that has almost no signal) representations match brain representations in both early and high-level visual cortex. Therefore, AlexNet seems to better capture the entire visual hierarchy including early, intermediate and high-level representations.

In summary, late AlexNet layers outperform most feature models (color, texture, contour, but not object parts), but not category models, in explaining primate IT. Object representations in late AlexNet layers correlate with categories and object parts but less so with lower-level visual features. Furthermore, late AlexNet layers can account for more than half of the variance that categories and object parts explain in IT. Searchlight analysis shows that AlexNet captures object representations across the visual hierarchy, whereas oracle models correlate mostly with representations in high-level visual cortex.

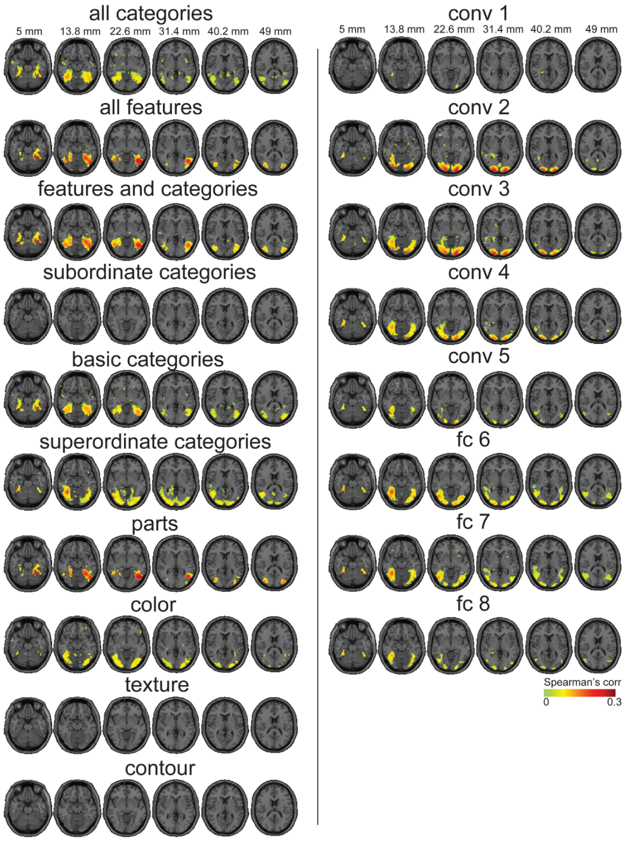

Figure 3. Searchlight analysis results, showing where in the brain oracle models and AlexNet layers explain image representations (Spearman's r between model and brain representations, two-sided Wilcoxon signed-rank test, FDR controlled at 0.05).

Discussion

DNNs perform considerably well at predicting object representations across the primate visual system (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). To better understand the similarities and differences between intuitive oracle models and complex DNNs, we compared DNN (AlexNet) representations to those of oracle models. Late DNN layers explain similar amounts of variance in primate IT as category and object-part models. The categorical representations in late DNN layers might contribute to their explanatory power. Consistent with this hypothesis, we show that late DNN layers can account for more than half of the variance that categories can explain in IT. Late DNN layers explain more variance than lower-level feature models, which further suggests that they can model variance that cannot be explained by lower-level visual features alone.

Searchlight analysis revealed a progression from early to high-level visual cortex with increasing DNN layer number. However, this progression is relative: even late DNN layers can explain representations in early visual cortex. This result indicates that object features may be overrepresented at late stages of processing in DNNs. This phenomenon could potentially contribute to adversarial examples or misclassification of objects. For example, a common misclassification of a sofa with a leopard pattern as an actual leopard, might result from a DNN relying too much on texture information. It is possible that some of the features that humans extract for categorization are different from those that DNNs extract. Nevertheless, DNNs are better than oracle models at explaining the object representation in early and intermediate visual cortex, capturing the entire visual hierarchy. The correspondence between representations in oracle models and DNN layers suggests that DNNs may provide a computationally explicit model of how features and categories are computed in the primate brain.

Acknowledgments

This work was funded by a Sir Henry Wellcome Postdoctoral Fellowship (206521/Z/17/Z) to KMJ, a DFG grant CI 241-1/1 (Emmy Noether Programme) to RMC and a British Academy Postdoctoral Fellowship (PS 140117) to MM.

References

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462.

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, 1–29.

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

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6), 4296–4309.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9.

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, 8619–8624.