Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks

Katharina Dobs (kdobs@mit.edu)
CBMM; Dept. of Brain and Cognitive Sciences; McGovern Institute for Brain Research;
Massachusetts Institute of Technology; 43 Vassar Street; Cambridge, MA 02139, USA

Alexander Kell (alex.kell@columbia.edu)
Zuckerman Mind Brain Behavior Institute; Columbia University
3227 Broadway, New York, NY 10027, USA

Ian Palmer (iapalm@mit.edu)
Dept. of Electrical Engineering and Computer Science; Massachusetts Institute of Technology
43 Vassar Street, Cambridge, MA 02139, USA

Michael Cohen (mcohen99@mit.edu)
CBMM; Dept. of Brain and Cognitive Sciences; McGovern Institute for Brain Research;
Massachusetts Institute of Technology; 43 Vassar Street; Cambridge, MA 02139, USA

Nancy Kanwisher (ngk@mit.edu)
CBMM; Dept. of Brain and Cognitive Sciences; McGovern Institute for Brain Research;
Massachusetts Institute of Technology; 43 Vassar Street; Cambridge, MA 02139, USA

Abstract:

Why does the human brain contain cortical regions specialized for the perception of some stimulus categories (e.g., faces), but not others (e.g., cars)? And why might functional specialization be a good design strategy for brains in the first place? Here, we used deep convolutional neural networks (CNNs) to test whether models optimized to recognize faces and objects require functional segregation for each task. First, we trained two separate CNNs with the same architecture to categorize either faces or objects. Unsurprisingly, the face-trained CNN performed worse on object categorization than the object-trained CNN and vice versa, demonstrating that the features optimized for each task differ from one another. Second, following the method of Kell et al (2018), we trained a family of dual-task CNNs on both tasks, asking how many layers can be shared before performance declines. Somewhat surprisingly, even the dual-task CNN that shared all layers performed nearly as well as the separate networks. This result is consistent with two hypotheses: 1) face and object recognition may be performed well by using a shared pool of common features or 2) the shared network has learned "hidden" functional specialization. In ongoing work, we are seeking to disambiguate these two hypotheses.

Keywords: functional specificity; object processing; face processing; deep neural networks; dual-task training

Introduction

Over the last 25 years, multiple regions of the human cortex have been identified that are engaged in specific components of perception and cognition. For example, the fusiform face area (FFA; Kanwisher, McDermott, & Chun, 1997) responds selectively to faces, the parahippocampal place area to scenes (PPA; Epstein & Kanwisher, 1998) and the extrastriate body area to images of bodies (EBA; Downing, Jiang, Shuman, & Kanwisher, 2001). The existence of these specialized regions raises two fundamental questions: i) Why might functional specialization be a good design strategy for brains?, and ii) Why are some perceptual and cognitive functions processed by specialized cortical modules while others apparently are not (e.g., cars or spiders; Downing, Chan, Peelen, Dodds, & Kanwisher, 2006)?

Deep convolutional neural networks (CNNs) offer a new approach for addressing these longstanding questions (Kell & McDermott, 2019). CNNs have been successfully used as models for visual processing in monkeys and humans (e.g., Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014), as well as for auditory processing in humans (Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018). In this latter study, the authors introduced a novel method to assess the extent to which representations at different layers in

CNNs can be shared across multiple tasks. The intuition behind this approach is that different tasks may employ an initial set of common, domain-general features, followed by branching into subsequent task-specific pathways. This approach was applied to two auditory tasks: music genre classification and word recognition. First, for each task, the authors trained a separate CNN —one on only the music task, and another on only the word task—to measure the performance attainable by networks that were free to learn task-specific features at all stages of processing. Second, they trained a single network on both tasks to measure how much task performance was impaired by being forced to share features across tasks. Third, they asked how many layers could be shared before task performance declined. They then trained "branched" networks, which shared the initial layers across both tasks before branching into two task-specific pathways, at all possible branch points (see Fig. 1 for examples). For musical genre and word recognition, the authors found that the dual-task network could share early (but not late) layers without impairing performance in each task. Here, we use this approach on category-specific visual processing to ask whether and at which branching point face and object processing performance declines relative to the fully separate networks that are each trained on only one task.

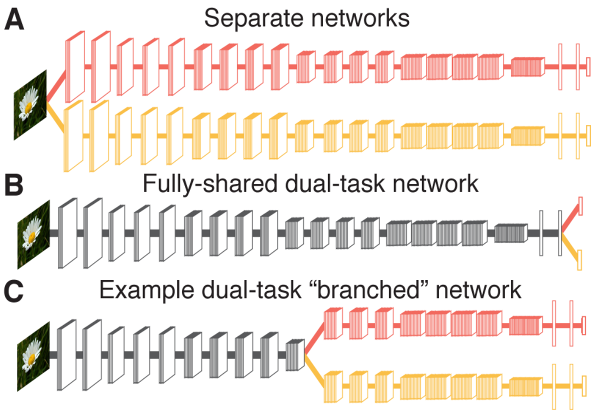
A
Separate networks
B
Fully-shared dual-task network
C
Example dual-task "branched" network

Figure 1: Examples of (A) separate networks trained on one task each, and (B) fully-shared and (C) "branched" dual-task architectures trained on two tasks simultaneously.

Methods

To address these questions, we trained CNNs based on the VGG16 architecture (Simonyan & Zisserman, 2015). This architecture has been successfully trained on object categorization and face identity recognition (Parkhi, Vedaldi, & Zisserman, 2015), respectively. It further has been shown to explain a significant amount of variance of human neural data (Schrimpf et al., 2018), suggesting that the CNNs learned visual features and representations similar to human neural visual representations. Here, we trained VGG16 networks separately on face and object categorization, and on both tasks simultaneously (i.e., dual-task).

Network Training and Testing

Separate face network: To measure face identity recognition performance in a fully separate network, we trained a randomly initialized network on face identity recognition only. The CNN was trained on a randomly sampled set of 1,000 identities (500 female) from the VGGFace2 database (Cao, Shen, Xie, Parkhi, & Zisserman, 2018). For each identity, we selected 300 images for training, and 50 images for validation. We used SGD with momentum (initial learning rate: 10^{-2}) and reduced the learning rate twice to 10^{-3} and 10^{-4} after 30 training epochs (i.e., full passes over the training set), respectively. All training parameters were selected in pilot experiments. The resulting classification accuracy on the validation set provides a performance measure of a network that is free to learn face-specific features at all stages of processing.

Separate object network: We measured object recognition performance achievable by a network trained on object recognition only. To keep the performance between the face and object networks comparable, we trained the same architecture (i.e., a randomly initialized VGG16 network) on 600 randomly sampled categories of the ILSVRC-2012 database (Deng et al., 2009) and used 500 images per category for training and 50 for validation. All other learning parameters were identical to the separate face network. The resulting classification accuracy on the validation set served as a measure for unconstrained object categorization performance.

Face and object decoding: To test whether the separate networks learned similar or distinct features, we decoded exemplars from an independent set of face identities and object categories in each network. Specifically, we extracted the activation in the penultimate layer of each network to 100 face and object images respectively (10 categories with 10 exemplars each). We then trained and tested a support vector machine on these activation patterns using a 10-fold cross-validation scheme.

Dual-task networks: To test how well a network trained on both tasks would perform, we used a dual-task architecture (see Fig. 1) and trained the same network on object categorization and face identity recognition simultaneously. We varied the number of shared layers between the two tasks by separating the network after each pooling (5 total) and fully connected layer (2 total) in the network, resulting in seven branch points. Each of these seven networks was randomly initialized, presented with batches (64 images) of face and object images interleaved, and otherwise trained with the same images and parameters as the separate networks. We assessed the classification accuracy for each task relative to the separate networks. If the cost of sharing representations across tasks is sufficiently high, we expect the performance to significantly drop relative to the separate networks. Meanwhile, a drop in performance at an early branching point would suggest that processing needs to be separated early to avoid an impairment in performance, while a drop at a later stage would suggest that the network can process both tasks simultaneously and still achieve performance similar to the separate networks.

Significance Testing: We obtained SEMs for all networks by bootstrapping across classes and images 10,000 times. Significance of comparison between the dual-task and the separate networks was obtained by using direct bootstrap tests and FDR-correction.

Results

Face and Object CNNs Learn Distinct Features

We were able to decode novel (i.e., untrained) object categories and face identities above chance (10%) from each network (Fig. 2). However, there was a large drop in decoding accuracy from the trained compared to the untrained stimulus type (faces versus objects). The features extracted from the network trained only on face identification were less useful to decode object categories (face CNN: 50% correct versus object CNN: 88%) and vice versa for face decoding (object CNN: 65% correct versus face CNN: 100%). These results suggest that the features optimized for face and object categorization differ from one another.

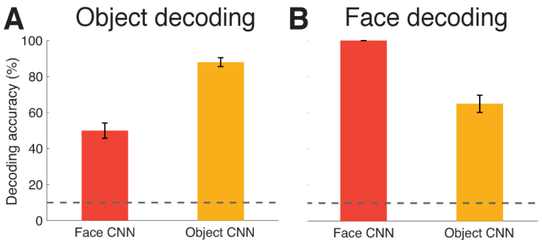

Figure 2: Decoding accuracy of decoding (A) object categories and (B) face identities from the penultimate layer of the network trained only on face identification (red) or only on object categorization (orange). Error bars indicate SEM across classification folds. The dashed grey line indicates chance level (10%).

Late-branching dual-task networks see modest, but significant detriment

Both separate networks achieve high performance on the validation set (face network: 1000-way classification: Top-1 94%, Top-5 98%; object network: 600-way classification: Top-1 57%, Top-5 81%), comparable to performance reported previously (Parkhi et al., 2015). The dual-task networks successfully learn to perform both tasks simultaneously (Fig. 3). While the performance of the dual-task networks that branched after the fifth pooling layer dropped significantly below the performance achieved by the separate networks, the differences were very small (object branch after pool 5: 1.49%; fc6: 1.34%; fc7: 1.84%; face branch after pool 5: 0.53%; fc6: 0.49%; fc7: 0.61%; all $p < 10^{-2}$, bootstrap test, FDR-corrected). In fact, even the network that does not branch until the last layer performed almost as well as the separate networks (face branch: Top-1 93.99% versus 94.60%; object branch: Top-1 55.37% versus 57.22%). At first glance, this result seems to suggest that shared processing does not substantially impair performance on face and object processing. But another possibility is that it is not necessary to impose any branching structure on the network before training, because the network discovers functional segregation itself. To address this question we are currently testing whether covert functional segregation is evident within the shared layers of the dual-task network, or whether in fact the two tasks are processed within largely overlapping units within these layers.

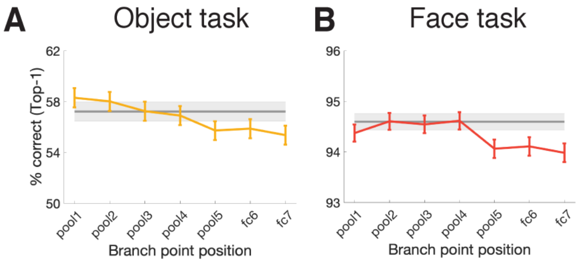

Figure 3: Performance of the dual-task networks as a function of location of branching point. Grey areas indicate performance obtained by the separate networks. Error bars plot SEM, bootstrapped over stimuli and classes.

Discussion

Our study tests the hypothesis that the functional segregation of visual category processing evident in the human brain reflects an efficient strategy for the problem of visual recognition. First, we showed that a network trained only on object recognition does not perform well at face recognition and vice versa, indicating that the feature spaces optimized for these two tasks differ from each other. Ongoing analyses are investigating the nature of these feature spaces at each layer, asking how early they diverge, and characterizing how they differ from one another. Next, we used a recently introduced method to test whether performance on the two tasks requires some segregation of processing in CNNs (Kell et al., 2018). To our surprise, we found that the performance of the fully shared dual-task network performed nearly as well on object and face recognition as the two separate networks did. At first glance, this result seems to argue against our hypothesis because we did not have to impose segregation between face and object processing in the network to maintain performance comparable to the separate networks. However, it is possible that even though we did not impose separate branches for face and object processing in the shared network, the network "discovered" the potential computational advantages of functional segregation on its own. Ongoing analyses are testing whether the shared network contains "hidden specialization", with separate processing of faces and objects within the layers. We further plan to compare the results for faces and objects (which are segregated in the brain) with findings for other pairs of tasks (e.g., cars and objects), for which functional specialization has not been observed in the human visual processing system.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, NIH grant Grant DP1HD091947 to N.K. and a Feodor-Lynen fellowship from the A. v. Humboldt foundation to K.D..

References

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *Intl. Conf. on Automatic Face and Gesture Recognition*, 1, 6.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2019 IEEE Conference on Computer Vision and Pattern Recognition,* 248–255.

Downing, P. E., Chan, A. W. Y., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex*, *16*(10), 1453–1461.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630–644.e16.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915–29.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*, *1*, 1–6.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007, 1–9.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv*, 1409.1556, 1-14.

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*, *111*(23), 8619–8624.