Functionally localized representations contain distributed information: insight from simulations of deep convolutional neural networks

Nicholas Blauch (nblauch@umass.edu)
University of Massachusetts, Amherst
Center for the Neural Basis of Cognition, Pittsburgh PA

Elissa Aminoff
Department of Psychology, Fordham University
Center for the Neural Basis of Cognition, Pittsburgh PA

Michael J. Tarr
Department of Psychology, Carnegie Mellon University
Center for the Neural Basis of Cognition, Pittsburgh PA

Abstract

Preferential activation to faces in the brain's fusiform gyrus has led to the proposed existence of a face module termed the Fusiform Face Area (FFA) (Kanwisher et. al, 1997). However, arguments for distributed, topographical object-form representations in FFA and across visual cortex have been proposed to explain data showing that FFA activation patterns contain decodable information about non-face categories (Haxby et. al, 2001; Hanson & Schmidt, 2011). Using two deep convolutional neural network models able to perform human-level object and facial recognition, respectively, we demonstrate that both localized category representations (LCRs) and high-level face-specific representations allow for similar decoding accuracy between non-preferred visual categories as between a preferred and non-preferred category. Our results suggest that neuroimaging of a cortical "module" optimized for face processing should yield significant decodable information for non-face categories so long as representations within the module are activated by non-face stimuli.

Keywords: module, localized categorical representation, distributed object-form topography, deep convolutional neural network, virtual electrophysiology

Introduction

How are mental representations organized in the brain? Do certain brain regions contain functional modules, dedicated to representing and processing a very specific type of information? Or is neural real estate more generally involved in the processing of many different types of stimuli? Evidence from fMRI has been used to propose the existence of functional modules for the processing of certain classes of visual information within the brain. Cortical modularity was proposed first for the visual processing of faces in the so-called Fusiform Face Area (FFA) (Kanwisher et. al, 1997), then for the visual processing of scenes/places in the so-called Parahippocampal Place Area (PPA) (Epstein & Kanwisher, 1998), and then for the visual processing of body parts in the so-called Extrastriate Body Area (EBA) (Downing et. al, 2001). In each of these studies, preferential activation of a certain class of visual stimuli (e.g. faces) in a certain region of the brain (e.g. fusiform gyrus) was used as evidence for modular processing within that region, leading to the authors renaming the region in terms of the modular processing (e.g. Fusiform Face Area). Not all authors agreed that preferential activation of a cortical region by a certain stimulus class was convincing evidence of underlying modular processing. Haxby et. al (2001) used multi-variate pattern analysis (MVPA) to demonstrate that putative functional modules for processing of scenes in the PPA and faces in the FFA contain patterns of activation useful for decoding whether a subject is viewing one of two categories not thought to be processed within the module. These authors interpreted their findings in the context of an "object-form topography" model, in which the ventral temporal cortex possesses a distributed, topographical representation of object-form features which underlie all forms of visual recognition. In their account, the large responses found in proposed functional modules are complemented by small responses throughout ventral temporal cortex in computations underlying visual categorization and other aspects of visual cognition.

Later, Spiridon & Kanwisher (2002) ran a similar fMRI study incorporating greater variability across images within a category (e.g. different viewpoints, exemplars, and image formats) in order to determine whether decodable abstract category information was truly distributed equally throughout ventral temporal cortex, as was argued by Haxby et. al (2001), or whether there might be localized decoding advantages corresponding to the locations of proposed functional modules. This study demonstrated that some abstract categorical information was present for certain categories outside their region of maximal activation (i.e. the location of a proposed module), replicating a main finding of Haxby et. al (2001). However, controlling for the number of voxels used in decoding analysis, this study demonstrated strong advantages in decodable information relating to discrimination between a preferred category (e.g. faces) and a non-preferred category (e.g. houses) in the region of proposed modularity (e.g. FFA). Additionally, in PPA and FFA, distinct disadvantages were found for the decoding of two non-preferred categories (e.g. faces vs. objects and objects vs. houses, respectively). Thus, while abstract

categorical information of certain categories may exist outside the region where modular processing is proposed, such abstract categorical information is by no means equally distributed throughout ventral temporal cortex. The authors thus argued for a more modular account of PPA and FFA, whereby these regions are primarily involved in the processing of a single category of information (scenes/houses and faces, respectively).

To account for both sets of findings, Cowell & Cottrell (2013) performed multi-variate pattern analysis (MVPA) on a neurocomputational model capable of discriminating between the 6 visual categories used in the analyses of Haxby et. al (2001). The neurocomputational model first applies Gabor filtering of input images to obtain a perceptual representation; it then feeds the activations of many Gabor filters into a self-organizing Kohonen map, which utilizes unsupervised learning to cluster its inputs into a two-dimensional representation. While the neurocomputational model contained no modular mechanisms, through unsupervised learning it developed the types of functionally localized stimulus representations used to argue in favor of modular processing, whereby certain patches of the Kohonen grid contained both preferential activation and enhanced decodable information about some categories. The effects were greatest for faces. Because they were able to simulate the data used to argue both for localized and distributed topographical representation with a neurocomputational model of distributed topographical representations that is more parsimonious than one postulating the existence of functional modules, the authors rejected the interpretation of a functional module for face processing based on the data of Spiridon & Kanwisher (2002).

The result of Cowell & Cottrell (2013) demonstrates that the evidence used to postulate functional modules may be accounted for by a model employing a distributed representation. However, the model is unable to account for human-level behavioral performance, and rather was constrained only to perform 6-way visual categorization. In the age of biologically-inspired computational systems capable of human-level object categorization (e.g. Krizhevsky et. al, 2012) and face individuation (e.g. Parki, Vedaldi, and Zimmerman, 2015), such behavioral constraints should become standard practice for models of neural representation. The decision not to constrain the computational model to perform human level face individuation, for example, belies the need for a functional module for face processing. Thus, we examine two deep, convolutional neural networks (DCNNs), one trained for large-scale object categorization and one trained for expert face individuation. In these networks, we focus on two types of category-specific representations for analysis. The first type of representation is the localized categorical representation (LCR), found in the final hidden layer of AlexNet (Krizhevsky et. al, 2012), whereby a single unit represents the likelihood of a given category in an image shown to the network. Such a localized categorical representation differs from the topographical object-form representations proposed by Haxby et. al (2001), in that a single value represents the abstract category information. However, localized category representations receive input from a processing layer which is well-described as a topographical object-form representation; thus, they are not true "modules". The second type of representation is taken as a deep layer of face-specific representations within the face-individuation network of Parki, Vedaldi, and Zimmerman (2015), VGG-Face, which is optimized for facial recognition only. In our view, VGG-Face *in toto* is a face-dedicated module; that is, a system optimized on and dedicated to to the processing of faces, only. The deep layer was chosen as a layer with high-level, complex face-specific features useful for recognition, but not explicitly representing individuals. We think that such representations are a reasonable model for what is proposed to be encoded in FFA (see Kanwisher & Yovel, 2006). We perform "virtual electrophysiology," (Yamins & DiCarlo, 2016) on both systems in order to determine whether these two types of category-specific representations produce the characteristic signal used to argue for distributed category-general representations: decodable information for non-preferred categories.

Method

Model simulations were run in the MATLAB programming environment, using the MatConvNet toolbox (Vedaldi & Lenc, 2015). Both models used in this study are examples of deep convolutional neural networks (DCNNs). Such networks were developed by computer vision researchers as engineering solutions for problems of visual recognition (e.g. LeCun et. al, 1998; Krizhevsky et. al, 2012). DCNNs contain several layers of processing, each of which contains a set of mathematical filtering operations (units or filters) which are convolved across the input, usually followed by a set of fully-connected layers which contain units which apply simple weighted summations of the units at the layer before. In all DCNNs for visual categorization, there exists a final layer of processing containing a set of units whose size is equal to the number of categories to be tested from, where each unit's activation corresponds to the likelihood that a certain category is present in the image; this vector of information is typically transformed via a softmax operation into explicit probabilities that the image may be categorized into each possible category.

The first DCNN model used is AlexNet (Krizhevsky et. al, 2012), pre-trained and uploaded to MatConvNet by Vedaldi & Lenc (2015). AlexNet was trained to perform 1000-way categorization of visual images on the 2011 ImageNet training set, which contains 1.2 million images evenly distributed across 1000 categories. For simulations, a different set of images not used in training, the 2011 ImageNet validation set, was used as stimuli, containing 50 images for each of 1000 categories. First, we recorded the activation patterns of each unit within AlexNet to each image of the validation set. While the network is said to contain 5 convolutional layers and 3 fully-connected layers, additional intermediate operations (rectification, pooling,

normalization) result in 22 stages of processing with activation values, where the first stage is defined by the RGB coordinates of the image. Full details on AlexNet can be found in the original paper (Krizhevsky et. al, 2012).

In our initial analyses, we consider four representative layers within AlexNet (Conv1, Conv5, FC6, and FC8) to demonstrate how informational content changes with depth in the network (Figure 1). The activation patterns of Conv1 are those which are input directly to Conv2, thus occurring after rectification, max-pooling, and response normalization. The activation patterns of Conv5 are taken after the fifth convolution and rectification, and are the inputs to FC6. The patterns of fully-connected layers FC6 and FC8 are taken after rectification. For each unit of each layer considered, we compute a set of "categorical signal-to-noise ratios" (cSNRs), for each category of ImageNet. For a given unit and category, the cSNR is computed as the signal-to-noise ratio of the unit's activation across all exemplars of the category. To create populations of units sorted by their cSNR for a given discrimination task on a subset of categories, we first create a vector containing the maximum unit cSNRs across all categories in the subset. This vector is then sorted in three ways, keeping the indices of units available: increasing cSNR, decreasing cSNR, and random. For each layer and discrimination task, three populations of size *n* are created by selecting the first *n* units from each of these vectors, for several values of *n,* and bootstrapping is performed across random samples of categories. The activation patterns of each population serve as the set of predictors for the classification of the ImageNet validation set images for each category in the discrimination task. Multi-class classification is achieved with a classification-tree based system, using the *fitctree* function in the MATLAB Statistics and Machine-Learning toolbox. The classifier is cross-validated using 10-folds of 80% training, 20% testing samples (*crossval* function). Finally, the loss is computed across the several folds of the cross-validated model (*kfoldLoss* function). Across bootrsapped camples of category, accuracy is reported as 1 – mean loss, and error bars are the standard error of accuracy.

Next, AlexNet FC8 is examined in more detail (Figure 2). In analyses similar to those conducted by Haxby et. al (2001), we compare the 2-way classification of preferred and non-preferred categories. For a given unit, the preferred category is the category which is explicitly represented; all other categories are potential non-preferred categories. Starting with 20 randomly-drawn ImageNet categories, we generate 100 pairs of preferred/non-preferred categories, and 100 pairs of non-preferred/non-preferred categories, with 5 pairs of each type for every category. For 2-way classification, we use a support-vector machine classifier (*fitcsvm* function, MATLAB Statistics and Machine-Learning toolbox). The same cross-validation and bootstrapping methods described in preceding analyses are used to generate an estimate of mean and standard error of classification accuracy for each pair type.

We perform similar analyses on VGG-Face, a DCNN trained for face-individuation on over 2000 faces (Parki, Vedaldi, and Zimmerman, 2015). To achieve a representation to serve as a model of FFA representations, we examine the activations in layer 35, the final layer before activations are condensed to individual face-specific representations. Layer 35 contains 4096 nodes representing high-level, complex information optimized for face individuation. In performing virtual electrophysiology on VGG-Face, we use as stimuli the fMRI localizer stimulus sets for faces, body-parts, objects, and scenes, in addition to the 2011 ImageNet validation set. These localizer sets are used by TarrLab and many other laboratories in the Center for the Neural Basis of Cognition, Pittsburgh PA, for fMRI research in order to localize functionally-defined regions such as the Fusiform Face Area (FFA), Extrastriate Body Area (EBA), Lateral Occipital area (LO), and Parahippocampal Place Area (PPA). Each localizer set contains 80 images of the category used to localize a corresponding functional brain area. 2-way classification tasks are created using pairs of the categories defining each localizer set. Additionally, a sample of 45 pairs taken from 10 randomly drawn ImageNet categories are used to bootstrap an estimate of the ability to predict all pairs of ImageNet categories. All units in VGG-Face layer 35 are used as predictors in a 2-way SVM classification system akin to that used in Figure 2 and the results are shown in Figure 3.

Results

The results of initial system-wide analyses of AlexNet representations are shown in Figure 1. In nearly all cases, decoding accuracy increases with the number of units used as predictors in the classifier, and decreases with the number of categories required for discrimination. In Conv1, Conv5, and FC6, for all discrimination tasks, sorting units by their categorical signal-to-noise ratio (cSNR; see methods) allows for improvements in decoding accuracy given the same number of units. However, in FC8, for all discrimination tasks, sorting units by cSNR has no effect on decoding accuracy, suggesting that localized categorical representations in FC8 possess information relevant to decoding between non-preferred categories.

The results of detailed analyses of AlexNet fully-connected layer 8 (FC8) are shown in Figure 2, where categories are organized by whether they are preferred (explicitly represented) or non-preferred (not represented) by a given unit in FC8. The decoding accuracy between the preferred category and a randomly chosen non-preferred category (see Methods) is not significantly greater than the decoding accuracy between randomly-generated pairs of non-preferred categories ($p=0.57$); both values are significantly greater than chance ($p<0.001$).

The results of all analyses of VGG-Face are shown in Figure 3. All discriminations involving face as one of two categories yield perfect discrimination accuracy. 2 out of 4 discriminations involving pairs of non-face categories (scene vs. body part; scene vs. object) yield perfect discrimination accuracy. The remaining 2 discrimination tasks (body part vs. object; pairs of ImageNet categories) yield non-perfect but greater than chance discrimination accuracy ($p<0.001$).

Figures

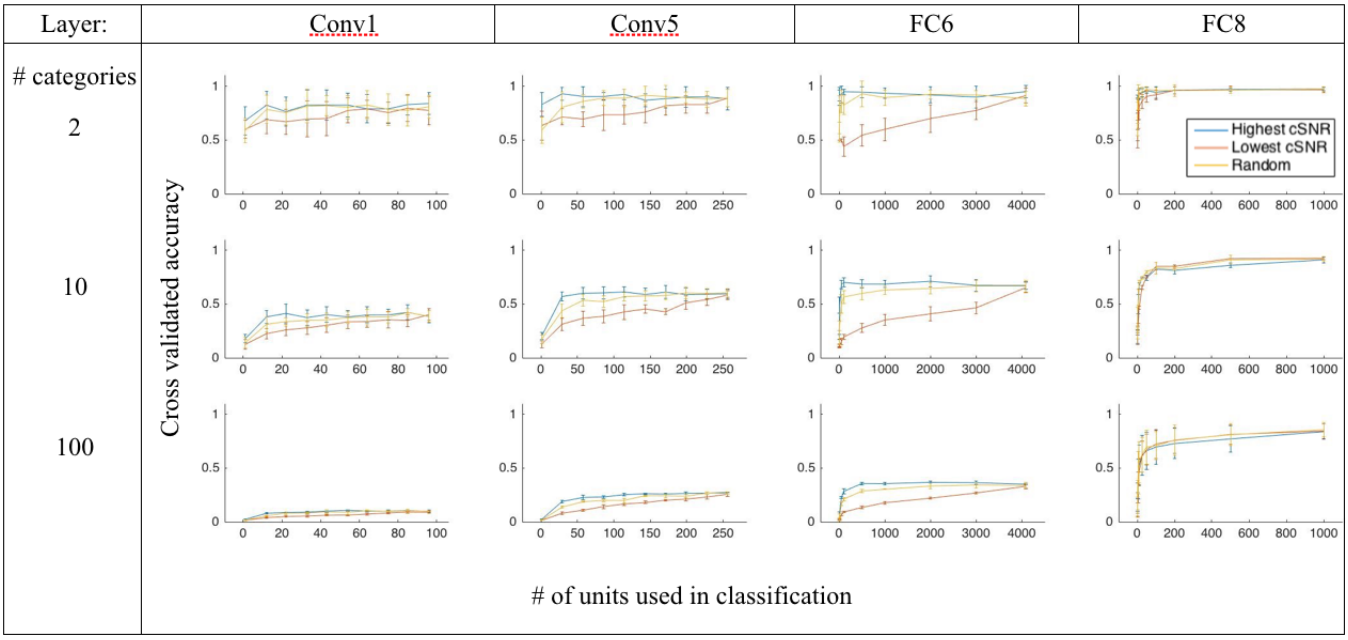

Figure 1: Population decoding from specified layers of the DCNN. Populations are selected in three ordering schemes, choosing the units with the highest cSNR (blue), the lowest cSNR (red), or at random (yellow). A large separation between curves indicates a local code, whereas a small separation indicates a distributed code.

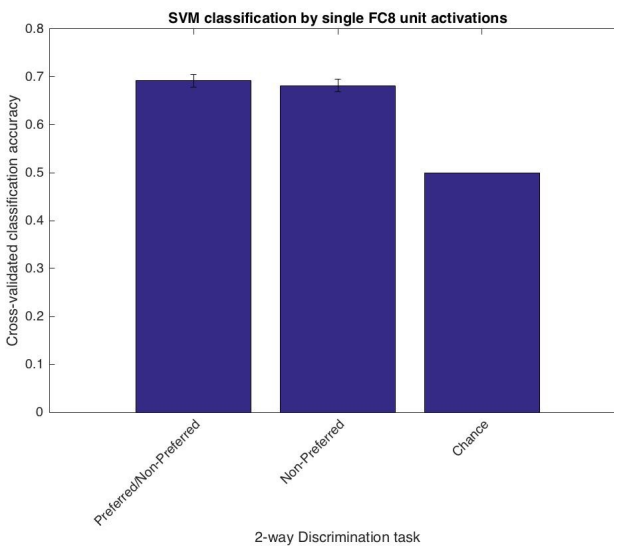

Figure 2: Single-unit decoding of FC8 in AlexNet. Each unit has a "preferred" category – the category it represents. All other categories are non-preferred categories. Bootstrapping was performed as described in Methods. Mean accuracies and standard errors across bootstrapping are shown. Accuracy for preferred/non-preferred is not significantly greater than accuracy for non-preferred only (p=0.57). Both values are significantly greater than chance (p<0.001).

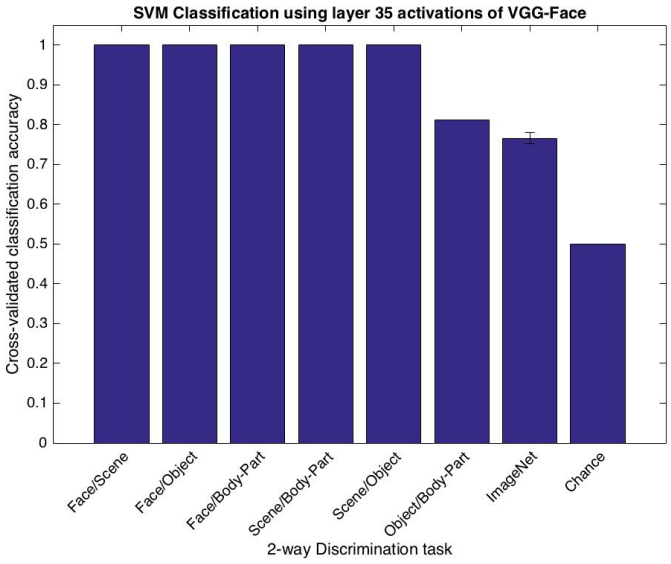

Figure 3: Support-vector machine (SVM) classification of layer 35 activations of VGG-Face network, shown for different pairs of categories. Each SVM is cross-validated with 10 folds of 80% training/20% test data, and the accuracy is reported as 1 – mean loss across folds. Error bars are shown as 1 – standard error of loss across folds. For ImageNet, standard error is computed across 45 bootstrapped pairs of 10 randomly chosen categories. All accuracies are significantly greater than chance (p < 0.001).

Discussion

First, using AlexNet, an arbitrary deep convolutional neural network (DCNN) for visual object categorization, we demonstrated that units explicitly representing a single visual category, what we deem localized categorical representations (LCRs), provide information allowing for the decoding of non-preferred categories, at a level equal to that for decoding between the represented category and a non-preferred category. Demonstrating that these localized categorical representations contain distributed information related to non-preferred categories suggests that the presence of domain-general decodable information in a putative domain-specific cortical region is not grounds to reject domain-specificity, as was done by Haxby et. al (2001), and Hanson & Schmidt (2011), in favor of distributed object-form topographical representations. While LCRs differ from distributed object-form topographical representations in that they represent abstract category information in a single value, the LCRs in AlexNet receive their input from a distributed object-form representation of the sort proposed by Haxby et. al (2001), and are by no means well-described as the sort of functional modules rejected by this study and proposed by the likes of Kanwisher et. al, (1997) for visual face processing, Epstein & Kanwisher (1998) for visual place/scene processing, or Downing et. al (2001), for visual body-part processing, whereby functional modules likely contain several processing stages for fine-grained analysis of exemplars of the preferred category. To ask whether such functional modules might also give rise to decodable information about stimuli outside the domain of modularity, we relied on a second DCNN specialized for face individuation, VGG-Face.

To acquire a representation of maximal similarity to the high-level face-optimized representations thought to be housed in the Fusiform Face Area, we took the layer 35 activations of VGG-Face, a set of 4096 nodes which project to the individual face probability nodes one layer later. We model these layer 35 nodes as a face module akin to the domain-specific interpretation of the Fusiform Face Area (e.g. Kanwisher et. al, 1997; Kanwisher et. al, 2006). We demonstrated that this "face module" contains patterns of activation capable of perfect discrimination between pairs of categories containing a face, and two of four pairs of categories not containing a face; the other two pairs yielded high discrimination significantly above chance. Thus, we find that domain-specific, face-optimized representations yield domain-general decodable information. This result provides support for the idea that activations within a cortical "module" might contain information relevant to decoding between categories for which that module is not specialized to process. This result strengthens our earlier result, demonstrating that it is improper to reject the possibility of a functional module associated with a given brain region on the grounds that the region's activation patterns allow for decoding between stimuli unrelated to the module's proposed primary function.

An important conceptual point is that the interpretation of our results – that modules should not be rejected on the grounds of producing domain-external information – rests on the assumption that a cortical module would be activated by domain-external information. In the case of localized categorical representations for object categories, it seems likely that all categories would be processed and that some activation might reach LCRs not representing the category of viewing. However, in the case of a cortical module for face processing, this point is less clear. Evidence of subcortical face detection mechanisms (for review, see Johnson, 2005) suggest that the brain may be capable of filtering out non-face information from higher processing (i.e., in FFA), via a fast detection process. Though, as we sometimes perceive faces on trees and in other places in which there are not faces, it is likely that non-face information does, on occasion, pass through the face-detector for further processing. It is possible that all information that arises for domain-external stimuli in FFA, for example, comes from images or image parts which contain something that looks enough like a face to pass through an early detection process, into higher regions of the face processing network. Once the visual information is allowed to pass, our results demonstrate that its processing within a face-optimized processor should give rise to decodable information.

Some authors have argued that the Fusiform Face Area (FFA) is better described as a mechanism for expert-level, fine-grained visual discriminations rather than a face-processing module, suggesting that neural substrate within FFA is specialized for visual categorization requiring repeated subordinate-level identification, a task which happens to occur most frequently in the context of face processing, thus resulting in the large preference for faces (e.g. Gauthier et. al, 1999; Tarr & Gauthier, 2000). Indeed, our results add an interesting point to this theoretical framework. Regardless of whether FFA is specialized for faces or expertise, if it develops representations useful for discriminating between individual faces, these representations are also likely to be useful for discriminating other visual objects. Thus, learning a new category of expertise (e.g. birds) might recruit a previously face-specific cortical region, on the basis of that region containing the most useful representations for the expert task, especially if exemplars of these categories have sufficient visual similarity to faces to pass through an early face-detection gate (if one exists). In this sense, FFA would not be a face module, but rather a brain area optimized most strongly for face-recognition, but also recruited for expert subordinate-level visual recognition.

Whether cortical modules exist in the sense motivated most strongly by Kanwisher (2010) remains an open debate. However, should such cortical modules exist, if their representations are activated by non-preferred categories, these "modules" are likely to produce activation patterns which allow for decoding between non-preferred categories, the characteristic result of studies which sometimes claim evidence of distributed, non-modular processing. As such, it

behooves the field to develop more sensitive and diagnostic measures to assess these critical questions regarding the fundamental nature of representation in the brain.

Acknowledgments

NB is grateful to his family, and especially to his parents Lisa and James, for their endless support; to the summer Undergraduate Program in Neural Computation at CNBC, Pittsburgh, where the majority of this research was performed; to Ying Yang for mentorship on machine learning techniques, and for helpful discussions of early pieces of this work throughout the summer; and finally to Rosie Cowell for introducing him to the modular debate, which spawned the theoretical interpretation of earlier results, as well as for helpful comments on an early version of this manuscript.

References

Cowell, R. A., & Cottrell, G. W. (2013). What Evidence Supports Special Processing for Faces? A Cautionary Tale for fMRI Interpretation. *Journal of Cognitive Neuroscience*, *25*(11), 1777–1793. https://doi.org/10.1162/jocn

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(September), 2470–2473.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expert ise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–73. Retrieved from http://www.biac.duke.edu/education/courses/spring03/cogdev/readings/I. Gauthier et al (1999).pdf

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. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hanson, S. J., & Schmidt, A. (2011). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. *NeuroImage*, *54*(2), 1715–1734. https://doi.org/10.1016/j.neuroimage.2010.08.028

Haxby, J. V, Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(September), 2425–2430. https://doi.org/10.1126/science.1063736

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–70. https://doi.org/10.1073/pnas.1005062107

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 : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–11. https://doi.org/10.1098/Rstb.2006.1934

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron*, *35*(6), 1157–1165. https://doi.org/10.1016/S0896-6273(02)00877-2

Tarr, M. J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764–769. https://doi.org/10.1038/77666

Vedaldi, A., & Lenc, K. (2015). MatConvNet. *Proceedings of the 23rd ACM International Conference on Multimedia - MM '15*, 689–692. https://doi.org/10.1145/2733373.2807412

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

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), 8619–24. https://doi.org/10.1073/pnas.1403112111

Parkhi, O. M., Vedaldi A.and Zisserman, A. (2015). Deep face recognition, *Proceedings of the British Machine Vision Conference (BMVC*).