Artificial neural networks reveal multivariate integration of information across different brain regions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unlike deep artificial neural networks for object recognition, the human brain is organized into category-selective regions that respond preferentially to images of faces, bodies, artifacts and scenes. However, when people make inferences about the world, they need to integrate information about objects in different categories. Therefore, the architecture of the human brain raises the question of how information in different category-selective regions is integrated. In this paper, we propose a framework to systematically detect hubs in the human brain that integrate information from multiple category-selective regions. For each location in the brain, neural responses are predicted timepoint-by-timepoint from the response patterns in category-selective regions using the Multivariate Pattern Dependence Network (MVPN): a feed-forward neural network extended from multivariate pattern dependence (MVPD). Brain regions whose responses are predicted significantly better using a combination of multiple category-selective regions than by the best of category-selective regions taken individually are identified as hubs. We applied MVPN to fMRI responses recorded during the presentation of complex dynamic stimuli (the movie Forrest Gump), and identified two integration hubs: the middle cingulate gyrus (MC), and the posterior medial thalamus, that were robust to different choices of the artificial neural networks used in MVPN. These results indicate that in the human brain information about objects in different categories is integrated by specialized, highly localized regions. More broadly, MVPN paves the way for the investigation of the fusion of information across different processing streams in the human brain.

1 Introduction

Extensive evidence from neuropsychology [1, 2], neurophysiology [3, 4], and functional brain imaging studies [5, 6, 7, 8] shows that distinct brain regions selectively respond to objects in different categories. In functional magnetic resonance imaging (fMRI), preferential responses to human faces are observed in the occipital face area (OFA), the fusiform face area (FFA) and face-selective posterior superior temporal sulcus (pSTS). Preferential responses to human bodies are observed in the extrastriate body area (EBA) and fusiform body area (FBA). Preferential responses to artifacts are observed in the medial fusiform gyrus (mFus) and middle temporal gyrus (MTG). Finally, preferntial responses to scenes are observed in the transverse occipital sulcus (TOS), parahippocampal place area (PPA), and retrosplenial cortex (RSC).

In contrast with the category-selective organization of visual cortex, solving the daily life problems people encounter requires to integrate perceptual information of objects across multiple categories. Imagine observing a person picking up an apple from a bowl with many different kinds of fruits.

Looking at her/him happily eating the apple, you might conclude that she/he really likes apples. To make such a seemingly simple inference, an observer needs to integrate information about the person and the kind of fruit they picked.

The mechanisms by which information about objects from different categories is integrated in the human brain remain unknown. According to one hypothesis, information across multiple regions is integrated via synchrony in their responses over multiple frequency bands [9]. According to an alternative account, integration of information from different category-selective regions might occur in specialized, highly localized hubs [10].

In this study, we used fMRI and an extension of multivariate pattern dependence [11] based on artificial neural networks (multivariate pattern dependence network - MVPN) to identify potential hubs for the integration of information across multiple categories. Specifically, we sought to detect brain regions where fMRI responses are predicted significantly better by the response patterns across all category-selective regions combined than by the regions selective for the single category with the highest prediction accuracy.

2 Methods

2.1 Data

High resolution BOLD fMRI responses to the movie Forrest Gump were obtained from the publicly available *studyforrest* dataset (`http://studyforrest.org`). In addition to the fMRI responses to the movie, the dataset includes an independent functional localizer that was used to identify category-selective regions. fMRI responses (acquired with a T2*-weighted echo-planar imaging sequence) were collected on a whole-body 3 Tesla Philips Achieva dStream MRI scanner equipped with a 32 channel head coil (see (Hanke et al., 2016) for more details). Fifteen right-handed participants took part in the study (6 females; age range 21-39, mean=9.4).

During the category localizer session, participants were shown 24 unique gray-scale images from each of six stimulus categories: human faces, human bodies without heads, small artifacts, houses, outdoor scenes, and phase scrambled images. They were presented with four block-design runs and a one-back matching task. During the movie localizer session, the movie stimulus ‘Forrest Gump’ was cut into eight segments, approximately 15 min long each. All eight movie segments were presented individually to participants in a chronological order in eight separate functional runs.

2.2 Preprocessing and de-noising

Data were preprocessed and analyzed using fMRIPrep (`https://fmriprep.readthedocs.io/en/latest/index.html`) in combination with custom MATLAB software. Specifically, we corrected all images for head movement with FSL MCFLIRT (`https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT`). Functional data of each participant were coregistered to her/his anatomical scan. Brain segmentation and normalization were performed with ANTs (`http://stnava.github.io/ANTs/`). The raw data of one participant could not pass the fMRIPrep processing pipeline. Of the remaining 14 participants, we removed noise from data by jointly using CompCorr and the removal of global signal, which has proven to be the optimal de-noising combination approach in a recent study [12].

2.3 ROI definition

Four sets of category-selective brain regions were identified using the first block-design run in the category localizer session (Fig. 1): face-selective regions (including the occipital face area - OFA, fusiform face area - FFA, and face-selective superior temporal sulcus - face STS), body-selective regions (including the extrastriate body area - EBA, fusiform body area - FBA, and body-selective superior temporal sulcus - body STS), artifact-selective regions (including the medial fusiform gyrus - mFus, and middle temporal gyrus - MTG), and scene-selective regions (including the transverse occipital sulcus - TOS, parahippocampal place area - PPA, and retrosplenial cortex - RSC). Data were modeled with a standard GLM using SPM12, and each seed ROI was defined as a 9mm radius sphere centered in the peak for its corresponding contrast (e,g. face-selective contrast: faces > bodys, artifacts, scenes and scrambled images). We combined data from both left and right hemisphere for

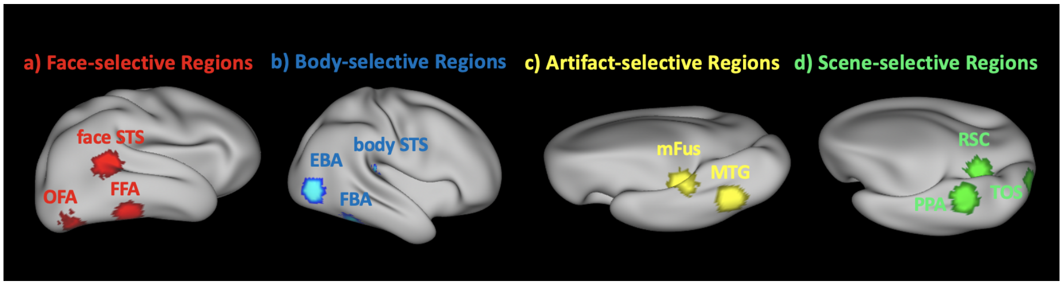
a) Face-selective Regions
b) Body-selective Regions
c) Artifact-selective Regions
d) Scene-selective Regions
face STS
OFA
FFA
body STS
EBA
FBA
mFus
MTG
RSC
TOS
PPA

Figure 1: **Regions of interest of one example participant shown on an inflated cortical surface: a.** face-selective regions are shown in red (fusiform face area - FFA, occipital face area - OFA, face-selective superior temporal sulcus - face STS); **b.** body-selective regions are shown in blue (fusiform body area - FBA, extrastriate body area - EBA, body-selective superior temporal sulcus - body STS); **c.** artifact-selective regions are shown in yellow (middle temporal gyrus - MTG, medial fusiform area - mFus); **d.** scene-selective regions are shown in green (parahippocampal place area - PPA, transverse occipital sulcus - TOS, retrosplenial cortex - RSC).

each ROI and then selected the 80 voxels which showed highest t-values for the contrast between the preferred category and other categories.

2.4 MVPN: Multivariate pattern dependence network

Most research on the interactions between brain regions has focused on the mean responses across voxels in different regions. However, fine grained patterns of response encode important information that could be lost by spatially-averaging. Over the past 15 years, multivariate pattern analysis (MVPA) [13] of fMRI data has achieved great success in the investigation of neural coding at a level of specificity that was unthinkable with simpler univariate analyses [14, 15, 16, 17, 18]. Despite this, relatively few attempts have been made to transport the potential of multivariate analyses to study brain connectivity. To fill this gap, a novel technique called multivariate pattern dependence (MVPD) [11] is recently developed to investigate the functional interactions between brain regions in terms of the multivariate relationship between their response patterns. MVPN has proven to drastically boost sensitivity than univariate connectivity methods [11], and is more robust to noise with independent training and testing data.

In this work, we extended MVPD to artificial neural networks (Multivariate Pattern Dependence Network, or MVPN) to investigate how information is integrated across multiple category-selective regions. In this approach, the multivariate patterns of response in one or more brain regions are used as the input of a feed-forward neural network trained to generate the patterns of response in another brain region (or in the case of this study, in the whole brain).

Let's consider an fMRI scan with m experimental runs. We can denote with $X_1, ..., X_m$ the multivariate timecourses in the predictor regions. Each matrix X_i is of size $n_X \times T_i$, where n_X is the number of voxels in the input regions, and T_i is the number of timepoints in the experimental run i. Analogously, we can denote with $Y_1, ..., Y_m$ the multivariate timecourses in the region that is the target of prediction, where each matrix Y_i is of size $n_Y \times T_i$, and n_Y is the number of voxels in the target region.

MVPN is trained with a leave-one-run-out procedure to learn a function f such that

$$Y_i(t) = f(X_i(t)) + \epsilon(t).$$

Specifically, for each choice of an experimental run i, data in the remaining runs are concatenated as the training dataset

$$\{(X_1, Y_1), ..., (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), ..., (X_m, Y_m)\},$$

while data $\{(X_i, Y_i)\}$ in the left-out run i is used as a testing dataset.

We use the voxelwise proportion of variance explained as a measurement of the multivariate statistical dependence between two brain regions. For each voxel j in the ROI, variance explained in run i is calculated as

$$\text{varExpl}_i(j) = 1 - \frac{SS(Y_i(j) - f(X_i(j)))}{SS(Y_i(j))},$$

where$f(X_i(j))$ is the voxel-wise prediction by MVPN. The values $\text{varExpl}_i(j)$ obtained for the different runs $i = 1, ..., m$ are averaged yielding $\overline{\text{varExpl}}(j)$.

2.5 Exploring integration hubs

To identify brain hubs that integrate information from multiple category-selective regions, we used 8 experimental runs during which participants watched the movie 'Forrest Gump'. The runs were used as separate folds for cross validation. fMRI segmentation based on ANTs (Advanced Normalization Tools) was used to generate an average gray matter mask comprising 53539 voxels. In a first analysis, we used MVPN to calculate the variance explained in each gray matter voxel using the predictor from each of the four category-selective regions (face, body, artifact, and scene) individually. In a second analysis, we combined all the category-selective regions jointly as inputs of MVPN to predict the brain responses of each voxel in the gray matter mask.

If the neural responses in a brain region are predicted significantly better by a model including all category-selective regions combined, than by the best of the category-selective regions in isolation, we can conclude that the region is receiving information from multiple category-selective regions. Otherwise, the brain region could be predicted at most as good as one of the category-selective regions taken individually if it only contains information from one category of objects. To make things more precise, we can denote with varExpl_{all} the variance explained by MVPN using as input the responses in all category-selective regions, and with varExpl_{max} the variance explained by MVPN using as input the responses in regions corresponding to the single best-predicting category. We then calculated for each voxel the difference

$$\Delta\,\text{varExpl} = \text{varExpl}_{all} - \text{varExpl}_{max},$$

which can be used as a metric to evaluate the potential role of integration across object categories. A positive $\Delta\,\text{varExpl}$ value indicates that the corresponding brain region is processing information from multiple category-selective regions, and is very likely to serve as a integration role of objects from different categories.

Critically, the best-predicting category was chosen separately for each voxel. Since the best performance among the predictions of each individual category-selective regions can be affected by the noise in different experimental runs, for each voxel we used 7 of the 8 runs to determine which category yielded the highest variance explained, and then selected the corresponding variance explained in the left-out run as the max value. This procedure was iterated across all 8 choices of the left-out run. Finally, the variance explained of the single best-predicting category varExpl_{max} is obtained by averaging the max values across all iterations, while varExpl_{all} is simply computed as the average across all experimental runs.

For each subject, we calculated the difference $\Delta\,\text{varExpl}$ in the whole brain, and statistical significance across participants was assessed with statistical nonparametric mapping using the SnPM extension for SPM (`http://warwick.ac.uk/snpm`).

2.6 Computational architectures

We trained MVPNs testing a variety of computational architectures (Fig. 2) to compare the predictive power of different models and to assess the robustness of our findings. Within each network architecture, we tested different parameter choices, such as how many layers are in the network, how many units are allocated to each hidden layer, and how the information across different layers are connected. All the architectures and parameter choices we have tested are reported in full in this article.

One-Layer MVPN As an initial exploration, we started from a simple linear fully connected network with one hidden layer (Fig. 2a). We tested this architecture in two variants, one with 20 hidden units and another with 100 hidden units.

A large number of parameters as compared to the number of datapoints can lead networks to overfit the data. To mitigate this issue, we tested an additional variant of the one-layer network with 100

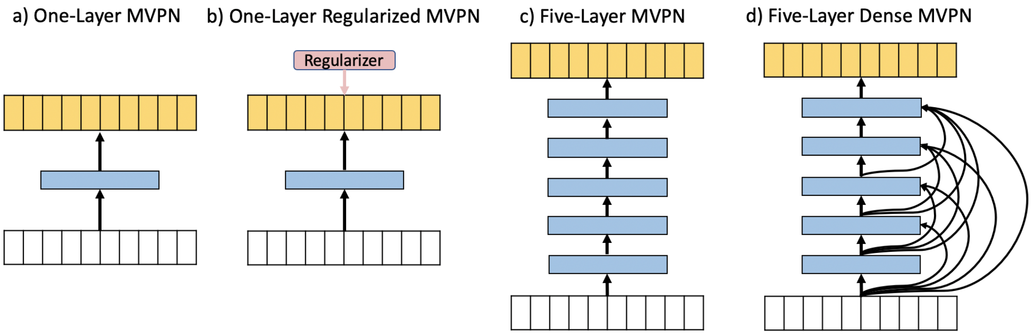

Figure 2: **Different MVPN architectures tested.**

hidden units and the L2 regularization (Fig. 2b). We chose the regularization parameter λ from values of 0, 0.001, 0.01, and 0.1 using nested cross-validation. Given time limitations, we adopted a simplified procedure that reduced the high computational demands of full nested cross-validation. Specifically, we split the movie data into 8 folds according to their experimental runs. The testing fold was chosen randomly within the inner cross validation for parameter tuning, and the evaluation was then conducted within a complete outer cross validation.

Five-Layer MVPN Previous studies have shown that deeper networks can approximate the same classes of functions as shallower networks using fewer parameters [19]. To test the performance of deeper network architectures, we therefore constructed a MVPN with five fully connected linear hidden layers (Fig. 2c).

Five-Layer Dense MVPN A challenge encountered in training deep neural networks is the vanishing-gradient problem. Recent architectures designed to mitigate this problem (i.e. DenseNet, [20]) have been extremely effective on a variety of datasts. To test the performance of DenseNet architecture in MVPN, we built a five-layer dense MVPN (Fig. 2d) which connects all layers directly with each other in a feed-forward manner. In this model, the outputs of all preceding layers are used as inputs for each layer, and its own outputs are used as inputs into all subsequent layers.

In this study, all different architectures of MVPN were trained over 5000 epochs using stochastic gradient descent (SGD) on a mean square error (MSE) loss, with a learning rate of 0.001 and a momentum of 0.9. We used a batch size of 32, and batch normalization was applied to the inputs of each layer. The original code implemented in pytorch is available on `https://github.com/pandamt/MVPN`.

3 Results

To explore how and where the information from multiple object categories is integrated in the brain, we trained three different MVPN architectures in a total of six conditions (one-layer MVPN with 20 hidden units, one-layer MVPN with 100 hidden units, one-layer regularized MVPN with 100 hidden units, five-layer MVPN with 20 hidden units, five-layer MVPN with 100 hidden units, five-layer dense MVPN with 100 hidden units). After training for 5000 epochs, all MVPN architectures we tested (with the exception of the 5 layer model with 20 hidden units) using left-out data successfully identified two highly localized brain regions: **the middle cingulate gyrus (MC)** and **the posterior medial thalamus**, where the difference Δ varExpl showed strongly and significantly positive values across all 14 participants ($p < 0.05$, corrected for multiple comparisons using SnPM, Fig. 3). Both of the two brain regions revealed a more accurate prediction with all category-selective regions jointly than the best prediction made with regions selective for a single category, and thus indicated the integration role of information about objects in different categories.

In comparison of the predictive power of different network architectures, MVPNs with a single hidden layer outperformed models with multiple layers given the same number of hidden units. However, this difference does not seem merely due to the overfitting problem as models with more hidden

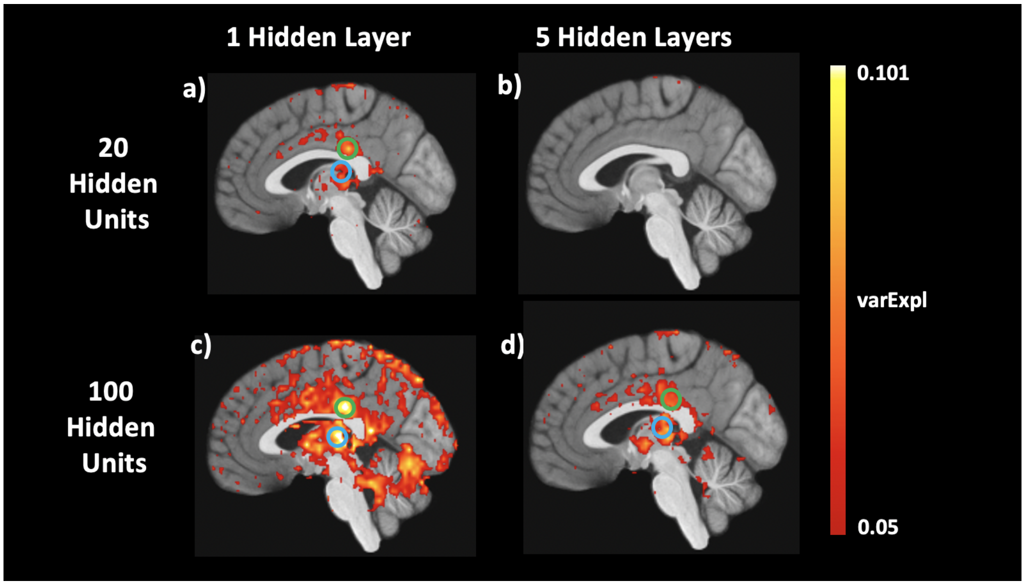

Figure 3: **Brain regions showing the integration of information from multiple categories in different MVPN architectures: a.** one-layer MVPN with 20 hidden units; **b.** five-layer MVPN with 20 hidden units in each hidden layer; **c.** one-layer MVPN with 100 hidden units; **d.** five-layer MVPN with 100 hidden units in each hidden layer. The middle cingulate gyrus (MC) was outline in green, and the posterior medial thalamus was outlined in blue. Difference Δ varExpl between the variance explained by MVPN using the inputs of all category-selective regions and of the single best-predicting category was averaged across all 14 participants and was thresholded at 5%.

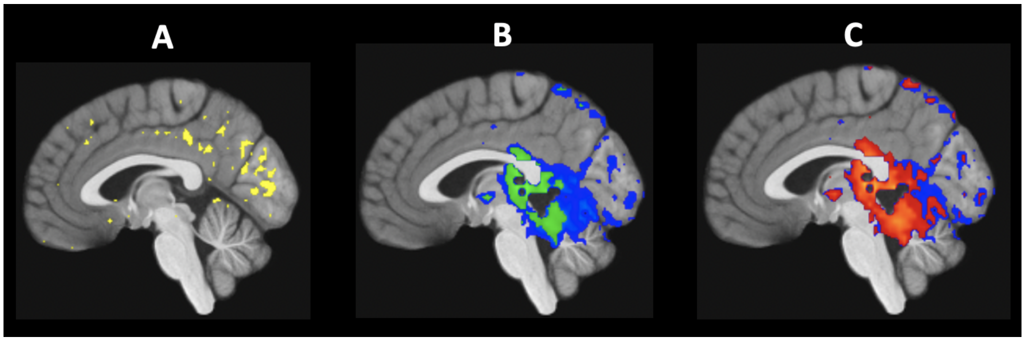

Figure 4: **Comparison of the predictive power of different MVPN architectures: A)** regions that are better predicted by the one-layer MVPN with 100 hidden units after adding the L2 regularization term are shown in yellow; Using all the category-selective regions combined as MVPN inputs, **B)** regions with variance explained higher than 20 % from the five-layer dense MVPN are shown in blue; regions with variance explained higher than 20 % from the five-layer MVPN are shown in green; **C)** regions with variance explained higher than 20 % from the five-layer dense MVPN are shown in blue; regions with variance explained higher than 20 % from the one-layer MVPN with 100 hidden units are shown in red.

layers having a larger number of parameters. In fact, models using layers with 100 hidden units outperformed models using layers with 20 hidden units (Fig. 3 bottom vs top). Accordingly, adding L2 regularization to the one-layer MVPN with 100 hidden units improved the prediction performance only mildly (Fig. 4A).

Given this observation, we hypothesized that models with five layers might be affected by the vanishing gradient problem. To test this possibility, we used a five-layer dense MVPN with 100 hidden units in each layer. The five-layer dense MVPN (Fig. 4B and 4C, blue) outperformed the five-layer feedforward network with 100 hidden units without dense connections (Fig. 4B, green) and even the one-layer network with 100 hidden units (Fig. 4C, red).

4 Discussion

Across multiple neural network architectures, we have identified two highly localized brain regions where responses are better predicted by response patterns across all category-selective regions combined than by the regions selective for the single category with the highest prediction accuracy: the middle cingulate gyrus (MC), and the posterior medial thalamus.

Previous research has identified regions in thalamus responding to multiple different sensory modalities [21, 22]. Given the results in this article, thalamus might be involved in integrating information not only across sensory modalities, but also across categories. Furthermore, an analysis of activation co-occurrence in neurosynth revealed that reported activation peaks in middle cingulate cortex correlate with reported activation peaks in the regions reported in [21] as part of the multimodal 'semantic system'. Future work will be needed to investigate further the relationship between integration across modalities and across categories in the human brain, and to test whether integration of information across modalities and across categories within the same regions is dictated by constraints of computational efficiency.

While our results identify two regions that integrate information across multiple category-selective networks, other regions beyond the ones we identified could contribute to integration between categories as well. For example, the artificial neural networks we have used in this article are completely linear. It is possible that nonlinear integration might occur in other brain regions that we could not detect with the approach used in the current study. Nonlinear networks benefit vastly from deeper architectures - the promising results we have obtained with the five-layer Dense MVPN pave the way for the exploration of deep nonlinear dense connectivity models.

A variety of MVPNs we tested in this study predicted multivariate brain responses in a timepoint by timepoint fashion. However, this one-to-one real time prediction could also be a possibility to constrict our model detecting other integration hubs. Instead of using data from one single timepoint as the predictor, additional future research could involve more temporal information into the model due to the potential time delay during the process of integration. For example, a temporal window that contains data at one timepoint and the one or two occurring just before can be added to model more complicated relationships between brain regions.

More generally, we have introduced an analysis technique that leverages artificial neural networks to investigate how information is integrated across multiple processing streams in the human brain. While we have applied this technique to the study of how information is integrated across regions selective for different categories, the same approach can be used to investigate other cases of information fusion (i.e. multimodal integration). Fusion of information processed along multiple streams is a problem that needs to be solved not only by the brain, but also by artificial neural networks [23]. The architecture used for fusion by the human brain could inspire the design of stream fusion in deep network algorithms.

Acknowledgments

We would like to thank the researchers who contributed to the *studyforrest* project (Hanke et al., 2016; Sengupta et al., 2016) for sharing their data, and the developers of fmriprep (Esteban et al., 2018) for their assistance with the fmriprep preprocessing pipeline.

References

[1] Antonio R Damasio, Daniel Tranel, and Hanna Damasio. Face agnosia and the neural substrates of memory. *Annual review of neuroscience*, 13(1):89–109, 1990.

[2] Marlene Behrmann, Gordon Winocur, and Morris Moscovitch. Dissociation between mental imagery and object recognition in a brain-damaged patient. *Nature*, 359(6396):636, 1992.

[3] Robert Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3(1):1–8, 1991.

[4] David Ian Perrett, Jari K Hietanen, Michael W Oram, and Philip J Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical transactions of the royal society of London. Series B: Biological sciences*, 335(1273):23–30, 1992.

[5] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

[6] Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[7] Alex Martin and Linda L Chao. Semantic memory and the brain: structure and processes. *Current opinion in neurobiology*, 11(2):194–201, 2001.

[8] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598, 1998.

[9] Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie. The brainweb: phase synchronization and large-scale integration. *Nature reviews neuroscience*, 2(4):229, 2001.

[10] Antonio R Damasio. The brain binds entities and events by multiregional activation from convergence zones. *Neural computation*, 1(1):123–132, 1989.

[11] Stefano Anzellotti, Alfonso Caramazza, and Rebecca Saxe. Multivariate pattern dependence. *PLoS computational biology*, 13(11):e1005799, 2017.

[12] Yichen Li, Rebecca Saxe, and Stefano Anzellotti. Intersubject mvpd: Empirical comparison of fmri denoising methods for connectivity analysis. *bioRxiv*, 2018.

[13] James V. Haxby, Maria Gobbini, Maura Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293:2425–30, 10 2001.

[14] Nikolaus Kriegeskorte, Elia Formisano, Bettina Sorger, and Rainer Goebel. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51):20600–20605, 2007.

[15] Adrian Nestor, David C Plaut, and Marlene Behrmann. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24):9998–10003, 2011.

[16] Stefano Anzellotti, Scott L Fairhall, and Alfonso Caramazza. Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, 24(8):1988–1995, 2013.

[17] Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543, 2008.

[18] Jorie Koster-Hale, Rebecca Saxe, James Dungan, and Liane L Young. Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14):5648–5653, 2013.

[19] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[21] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796, 2009.

[22] Sascha Tyll, Eike Budinger, and Toemme Noesselt. Thalamic influences on multisensory integration. *Communicative & integrative biology*, 4(4):378–381, 2011.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.