OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images

Umut Güçlü*, Marcel A. J. van Gerven

Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

Abstract

Encoding and decoding in functional magnetic resonance imaging has recently emerged as an area of research to noninvasively characterize the relationship between stimulus features and human brain activity. To overcome the challenge of formalizing what stimulus features should modulate single voxel responses, we introduce a general approach for making directly testable predictions of single voxel responses to statistically adapted representations of ecologically valid stimuli. These representations are learned from unlabeled data without supervision. Our approach is validated using a parsimonious computational model of (i) how early visual cortical representations are adapted to statistical regularities in natural images and (ii) how populations of these representations are pooled by single voxels. This computational model is used to predict single voxel responses to natural images and identify natural images from stimulus-evoked multiple voxel responses. We show that statistically adapted low-level sparse and invariant representations of natural images better span the space of early visual cortical representations and can be more effectively exploited in stimulus identification than hand-designed Gabor wavelets. Our results demonstrate the potential of our approach to better probe unknown cortical representations.

Citation: Güçlü U, van Gerven MAJ (2014) Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. PLoS Comput Biol 10(8): e1003724. doi:10.1371/journal.pcbi.1003724

Editor: Nikolaus Kriegeskorte, Medical Research Council, United Kingdom

Received November 5, 2013; **Accepted** April 28, 2014; **Published** August 7, 2014

Funding: The first author was supported by the Huygens Scholarship Programme of the Netherlands organisation for international cooperation in higher education (http://www.nuffic.nl/) and the Academy Assistants Programme of the Royal Netherlands Academy of Arts and Sciences (http://www.knaw.nl/). No further funding was received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: u.guclu@donders.ru.nl

Introduction

An important goal of contemporary cognitive neuroscience is to characterize the relationship between stimulus features and human brain activity. This relationship can be studied from two distinct but complementary perspectives of encoding and decoding [1]. The encoding perspective is concerned with how certain aspects of the environment are stored in the brain and uses models that predict brain activity in response to certain stimulus features. Conversely, the decoding perspective uses models that predict specific stimulus features from stimulus-evoked brain activity and is concerned with how specific aspects of the environment are retrieved from the brain.

Stimulus-response relationships have been extensively studied in computational neuroscience to understand the information contained in individual or ensemble neuronal responses, based on different coding schemes [2]. The invasive nature of the measurement techniques of these studies has restricted human subjects to particular patient populations [3,4]. However, with the advent of functional magnetic resonance imaging (fMRI), encoding and decoding in fMRI has made it possible to noninvasively characterize the relationship between stimulus features and human brain activity via localized changes in blood-oxygen-level dependent (BOLD) hemodynamic responses to sensory or cognitive stimulation [5].

Encoding models that predict single voxel responses to certain stimulus features typically comprise two main components. The first component is a (non)linear transformation from a stimulus space to a feature space. The second component is a (non)linear transformation from the feature space to a voxel space. Encoding models can be used to test alternative hypotheses about what a voxel represents since any encoding model embodies a specific hypothesis about what stimulus features modulate the response of the voxel [5]. Furthermore, encoding models can be converted to decoding models that predict specific stimulus features from stimulus-evoked multiple voxel responses. In particular, decoding models can be used to determine the specific class from which the stimulus was drawn (i.e. classification) [6,7], identify the correct stimulus from a set of novel stimuli (i.e. identification) [8,9] or create a literal picture of the stimulus (i.e. reconstruction) [10–12].

The conventional approach to encoding and decoding makes use of feature spaces that are typically hand-designed by theorists or experimentalists [8,9,11,13–16]. However, this approach is prone to the influence of subjective biases and restricted to a priori hypotheses. As a result, it severely restricts the scope of alternative hypotheses that can be formulated about what a voxel represents. This restriction is evident by a paucity of models that adequately characterize extrastriate visual cortical voxels.

A recent trend in models of visual population codes has been the adoption of natural images for the characterization of voxels that respond to visual stimulation [8,13]. The motivation behind this trend is that natural images admit multiple feature spaces such as low-level edges, mid-level edge junctions, high-level object parts and complete objects that can modulate single voxel responses [5]. Implicit about this motivation is the assumption that the brain is adapted to the statistical regularities in the environment [17] such

Author Summary

An important but difficult problem in contemporary cognitive neuroscience is to find what stimulus features best drive responses in the human brain. The conventional approach to solve this problem is to use descriptive encoding models that predict responses to stimulus features that are known a priori. In this study, we introduce an alternative to this approach that is independent of a priori knowledge. Instead, we use a normative encoding model that predicts responses to stimulus features that are learned from unlabeled data. We show that this normative encoding model learns sparse, topographic and invariant stimulus features from tens of thousands of grayscale natural image patches without supervision, and reproduces the population behavior of simple and complex cells. We find that these stimulus features significantly better drive blood-oxygen-level dependent hemodynamic responses in early visual areas than Gabor wavelets–the fundamental building blocks of the conventional approach. Our approach will improve our understanding of how sensory information is represented beyond early visual areas since it can theoretically find what stimulus features best drive responses in other sensory areas.

as those in natural images [18,19]. At the same time, recent developments in theoretical neuroscience and machine learning have shown that normative and predictive models of natural image statistics learn statistically adapted representations of natural images. As a result, they predict statistically adapted visual cortical representations, based on different coding principles. Some of these predictions have been shown to be similar to what is found in the primary visual cortex such as topographically organized simple and complex cell receptive fields [20].

Building on previous studies of visual population codes and natural image statistics, we introduce a general approach for making directly testable predictions of single voxel responses to statistically adapted representations of ecologically valid stimuli. To validate our approach, we use a parsimonious computational model that comprises two main components (Figure 1). The first component is a nonlinear feature model that transforms raw stimuli to stimulus features. In particular, the feature model learns the transformation from unlabeled data without supervision. The second component is a linear voxel model that transforms the stimulus features to voxel responses. We use an fMRI data set of voxel responses to natural images that were acquired from the early visual areas (i.e. V1, V2 and V3) of two subjects (i.e. S1 and S2) [21]. We show that the encoding and decoding performance of this computational model is significantly better than that of a hand-designed Gabor wavelet pyramid (GWP) model of phase-invariant complex cells. The software that implements our approach is provided at http://www.ccnlab.net/research/.

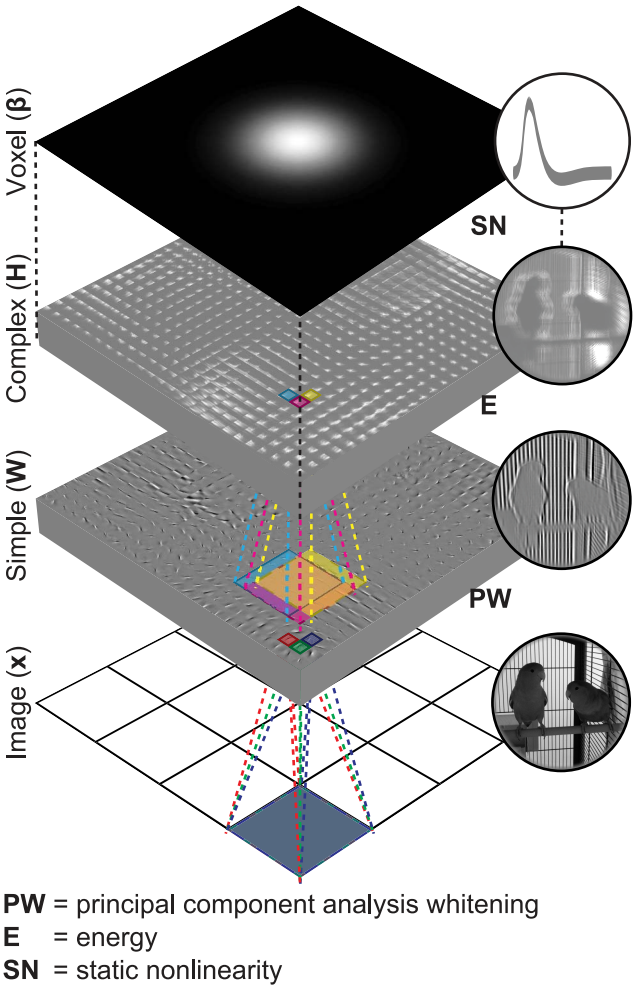

Figure 1. Encoding model. The encoding model predicts single voxel responses to images by nonlinearly transforming the images to complex cell responses and linearly transforming the complex cell responses to the single voxel responses. For example, the encoding model predicts a voxel response to a 128×128 image $\mathbf{x}$ as follows: Each of the 16 non-overlapping 32×32 patches of the image $\hat{\mathbf{z}}^{(\mathbf{i})}$ is first vectorized, preprocessed and linearly transformed to 625 simple cell responses, i.e. $\mathbf{W}\mathbf{z}^{(\mathbf{i})}$ where $\mathbf{z}^{(\mathbf{i})}$ is a vectorized and preprocessed patch. Energies of the simple cells that are in each of the 625 partially overlapping 5×5 neighborhoods are then locally pooled, i.e. $\mathbf{H}(\mathbf{W}\mathbf{z}^{(\mathbf{i})})^2$, and nonlinearly transformed to one complex cell response, i.e. $\log(1+\mathbf{H}(\mathbf{W}\mathbf{z}^{(\mathbf{i})})^2)$. Next, 10000 complex cell responses are linearly transformed to the voxel response, i.e. $\beta^\top \phi(\mathbf{x})$ where $\phi(\mathbf{x}) = ((\log(1+\mathbf{H}(\mathbf{W}\mathbf{z}^{(1)})^2)^\top, \ldots, (\log(1+\mathbf{H}(\mathbf{W}\mathbf{z}^{(16)})^2)^\top)^\top$. The feature transformations are learned from unlabeled data. The voxel transformations are learned from feature-transformed stimulus-response pairs.
doi:10.1371/journal.pcbi.1003724.g001

Results

Feature models

To learn the feature transformation, we used a two-layer sparse coding (SC) model of 625 simple (i.e. first layer) and 625 complex (i.e. second layer) cells [22]. Concretely, the simple cells were first arranged on a square grid graph that had circular boundary conditions. The weights between the simple and complex cells were then fixed such that each complex cell locally pooled the energies of 25 simple cells in a 5×5 neighborhood. There were a total of 625 partially overlapping neighborhoods that were centered around the 625 simple cells. Next, the weights between the input and the simple cells were estimated from 50000 patches of size 32×32 pixels by maximizing the sparseness of the locally pooled simple cell energies. Each simple cell was fully connected to the input (i.e. patch of size 32×32 pixels). The patches were randomly sampled from the 1750 images of size 128×128 pixels in the estimation set. To maximize the sparseness, the energy function (i.e. square nonlinearity) encourages the simple cell responses to be similar within the neighborhoods while the sparsity function (i.e. convex nonlinearity) encourages the locally pooled simple cell energies to be thinly dispersed across the neighborhoods. As a result, the simple cells that are in the same

neighborhood have simultaneous activation and similar preferred parameters. Since the neighborhoods overlap, the preferred parameters of the simple and complex cells change smoothly across the grid graph. Finally, the complex cell responses of the SC model were defined as a static nonlinear function of the locally pooled simple cell energies after model estimation (i.e. total of 625 complex cell responses per patch of size 32×32 pixels and 10000 complex cell responses per image of size 128×128 pixels). The SC model learned topographically organized, spatially localized, oriented and bandpass simple and complex cell receptive fields that were similar to those found in the primary visual cortex (Figure 2A) [23–26].

To establish a baseline, we used a GWP model [25,27,28] of 10921 phase-invariant complex cells [8]. Variants of this model were used in a series of seminal encoding and decoding studies [8,13,14,16]. Note that the fMRI data set was the same as that in [8,13]. Concretely, the GWP model was a hand-designed population of quadrature-phase Gabor wavelets that spanned a range of locations, orientations and spatial frequencies (Figure 2B). Each wavelet was fully connected to the input (i.e. image of size 128×128 pixels). The complex cell responses of the GWP model were defined as a static nonlinear function of the pooled energies of the quadrature-phase wavelets that had the same location, orientation and spatial frequency (i.e. total of 10921 complex cell responses per image of size 128×128 pixels).

Voxel models

To learn the voxel transformation, we used regularized linear regression. The voxel models were estimated from the 1750 feature-transformed stimulus-response pairs in the estimation set by minimizing the L^2 penalized least squares loss function. The combination of a voxel model with the complex cells of the SC and GWP models resulted in two encoding models (i.e. SC2 and GWP2 models). The SC2 model linearly pooled the 10000 complex cell responses of the SC model. The GWP2 model linearly pooled the 10921 complex cell responses of the GWP model.

Receptive fields

We first analyzed the receptive fields of the SC model (i.e. simple and complex cell receptive fields). The preferred phase, location, orientation and spatial frequency of the simple and complex cells were quantified as the corresponding parameters of Gabor wavelets that were fit to their receptive fields. The preferred parameter maps of the simple and complex cells were constructed by arranging their preferred parameters on the grid graph (Figure 3). Most adjacent simple and complex cells had similar location, orientation and spatial frequency preference, whereas they had different phase preference. In agreement with [22], the preferred phase, location and orientation maps reproduced some of the salient features of the columnar organization of the primary visual cortex such as lack of spatial structure [29], retinotopy [30] and pinwheels [31], respectively. In contrast to [22], the preferred spatial frequency maps failed to reproduce cytochrome oxidase blobs [32]. The preferred phase map of the simple cells suggests that the complex cells are more invariant to phase and location than the simple cells since the complex cells pooled the

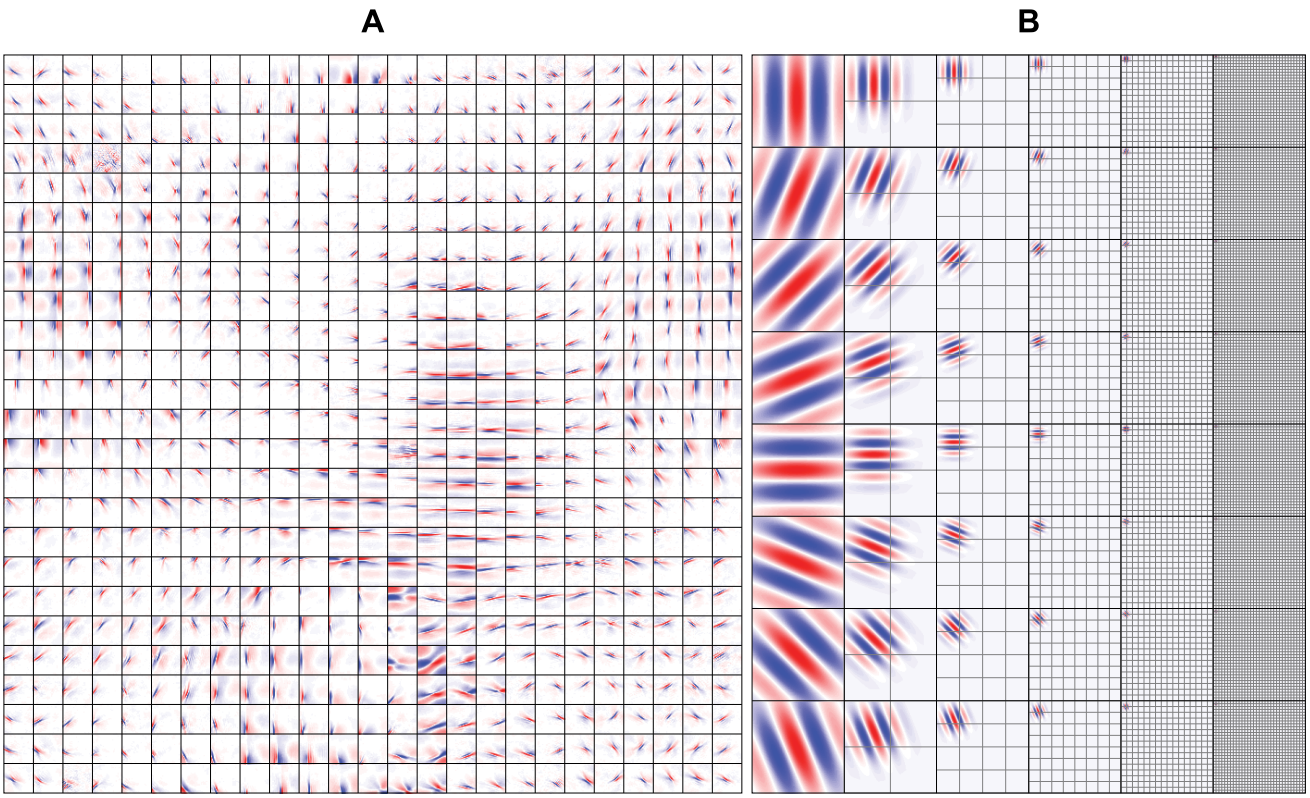

Figure 2. Simple cell receptive fields. (A) Simple cell receptive fields of the SC model. Each square is of size 32×32 pixels and shows the inverse weights between the input and a simple cell. The receptive fields were topographically organized, spatially localized, oriented and bandpass, similar to those found in the primary visual cortex. (B) Simple cell receptive fields of the GWP model. Each square is of size 128×128 pixels and shows an even-symmetric Gabor wavelet. The grids show the locations of the remaining Gabor wavelets that were used. The receptive fields spanned eight orientations and six spatial frequencies.
doi:10.1371/journal.pcbi.1003724.g002

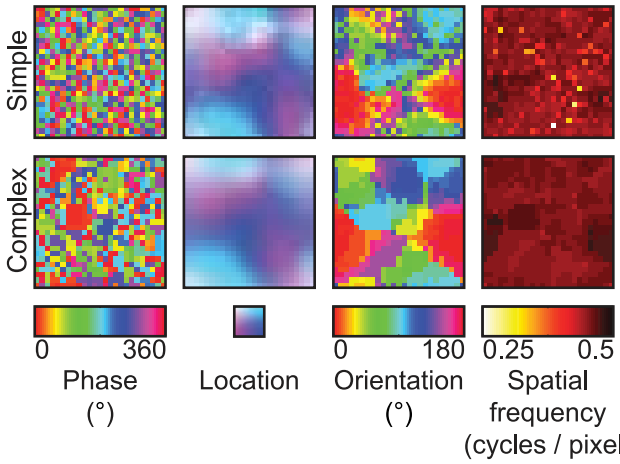

Figure 3. Preferred parameter maps of the SC model. The phase, location, orientation and spatial frequency preference of the simple and complex cells were quantified as the corresponding parameters of Gabor wavelets that were fit to their receptive fields. Each pixel in a parameter map shows the corresponding preferred parameter of a simple or complex cell. The adjacent simple and complex cells had similar location, orientation and spatial frequency preference but different phase preference.
doi:10.1371/journal.pcbi.1003724.g003

energies of the simple cells that had different phase preference. To verify the invariance that is suggested by the preferred phase map of the simple cells, the population parameter tuning curves of the simple and complex cells were constructed by fitting Gaussian functions to the median of their responses to Gabor wavelets that had different parameters (Figure 4). Like the simple cells, most complex cells were selective to orientation (i.e. standard deviation of 21.8° versus 22.9°) and spatial frequency (i.e. standard deviation of 0.52 versus 0.54 in normalized units). Unlike the simple cells, most complex cells were more invariant to phase (i.e. standard deviation of 50.0° versus 158.1°) and location (i.e. standard deviation of 3.70 pixels versus 5.86 pixels). Therefore, they optimally responded to Gabor wavelets that had a specific orientation and spatial frequency, regardless of their phase and exact position.

We then analyzed the receptive fields of the SC2 model (i.e. voxel receptive fields). The eccentricity and size of the receptive fields were quantified as the mean and standard deviation of two-dimensional Gaussian functions that were fit to the voxel responses to point stimuli at different locations, respectively. The orientation and spatial frequency tuning of the receptive fields were taken to be the voxel responses to sine-wave gratings that spanned a range of orientations and spatial frequencies. While the eccentricity, size and orientation tuning varied across voxels, most voxels were tuned to relatively high spatial frequencies (Figure 5A and Figure 5B). The mean predicted voxel responses to sine-wave gratings that had oblique orientations were higher than those that had cardinal orientations and this difference decreased with spatial frequency (Figure 5C). While this result is in contrast to those of the majority of previous single-unit recording and fMRI studies [33,34], it is in agreement with those of [35]. In line with [36,37], the receptive field size systematically increased from V1 to V3 and from low receptive field eccentricity to high receptive field eccentricity (Figure 6). The properties of the GWP2 model were similar to those in [8]. The relationship between the receptive field parameters (i.e. size, eccentricity, area) of the GWP2 model were the same as those of the SC2 model. However, the GWP2 model did not have a large orientation bias.

Encoding

The encoding performance of the SC2 and GWP2 models was defined as the coefficient of determination (R^2) between the observed and predicted voxel responses to the 120 images in the validation set across the two subjects. The performance of the SC2 model was found to be significantly higher than that of the GWP2 model (binomial test, $p \ll 0.05$). Figures 7A and 7B compare the performance of the models across the voxels that survived an R^2 threshold of 0.1. The mean R^2 of the SC2 model systematically decreased from 0.28 across 28% of the voxels in V1 to 0.21 across 11% of the voxels in V3. In contrast, the mean R^2 of the GWP2 model systematically decreased from 0.24 across 24% of the voxels in V1 to 0.16 across 6% of the voxels in V3. Figure 7C compares the performance of the models in each voxel. More than 71% of the voxels that did not survive the threshold in each area and more than 92% of the voxels that survived the threshold in each area were better predicted by the SC2 model than the GWP2 model. These results suggest that statistically adapted low-level sparse representations of natural images better span the space of early visual cortical representations than the Gabor wavelets.

Decoding

The decoding performance of the SC2 and GWP2 models was defined as the accuracy of identifying the 120 images in the validation set from a set of 9264 candidate images. The set of candidate images contained the 120 images in the validation set and the 9144 images in the Caltech 101 data set [38]. Note that the set of candidate images was ten- to hundred-fold larger than the sets in [8] but comparable to the largest set in [15]. The performance of the SC2 model was found to be significantly higher than that of the GWP2 model (binomial test, $p < 0.05$). Figure 8 compares the performance of the models. The mean accuracy of the SC2 model across the subjects was 61%. In contrast, the mean accuracy of the GWP2 model across the subjects was 49%. The chance-level accuracy was 0.01%. These results suggest that statistically adapted low-level sparse representations of natural images can be more effectively exploited in stimulus identification than the Gabor wavelets.

Spatial invariance

In principle, the SC2 and GWP2 models should have some degree of spatial invariance since they linearly pooled the responses of the complex cells that displayed insensitivity to local stimulus position. Spatial invariance is of particular importance for decoding since a reliable decoder should be able to identify a stimulus, regardless of its exact position. Furthermore, a difference between the degree of spatial invariance of the models can be a contributing factor to the difference between their performance. To analyze the spatial invariance of the models, we evaluated their encoding and decoding performance after translating the images in the validation set by 0.8° (i.e. approximately the standard deviation of the population location tuning curves of the complex cells of the SC model) in a random dimension (Figure 9). The encoding and decoding performance of the models was found to decrease after the translations. Unlike the encoding performance of the GWP2 model, that of the SC2 model decreased less in V3 than V1. This result suggests greater spatial invariance in V3 than V1. The difference between the mean R^2 of the models across the voxels that survived the threshold before the translations increased from 0.05 to 0.11. The difference between the mean accuracy of

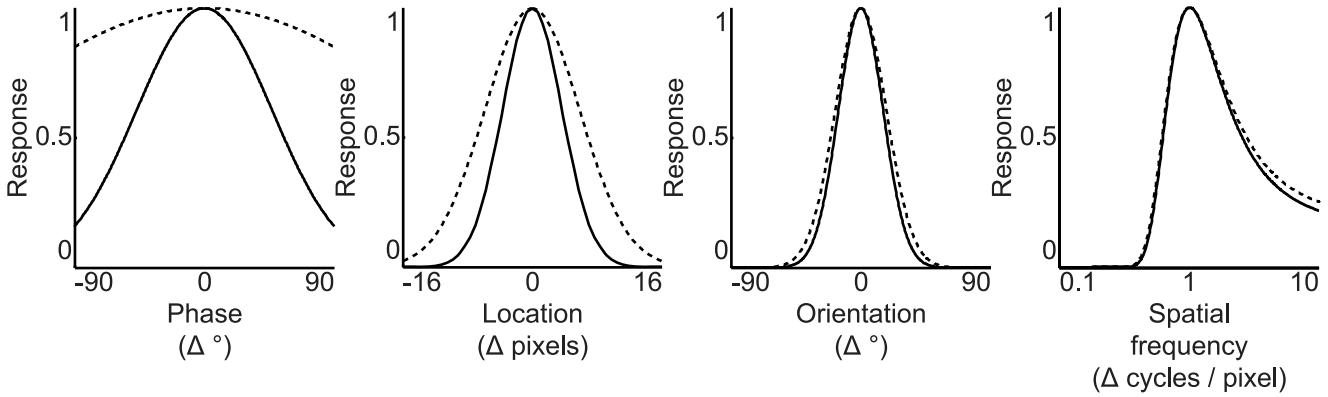

Figure 4. Population parameter tuning curves of the SC model. The population phase, location, orientation and spatial frequency tunings of the simple (solid lines) and complex cells (dashed lines) were quantified by fitting Gaussian functions to the median of their responses to Gabor wavelets that had different parameters. Each curve shows the median of their responses as a function of change in their preferred parameter. The complex cells were more invariant to phase and location than the simple cells.
doi:10.1371/journal.pcbi.1003724.g004

the models across the subjects increased from 12% to 24%. These results suggest that the SC2 model is more tolerant to local translations in stimulus position than the GWP2 model.

Control models

Since the SC2 and GWP2 models had different nonlinearities (i.e. pooling and static nonlinearity), a direct evaluation of the contribution of their components (i.e. representations and nonlinearities) to the difference between their encoding performance was not possible. Therefore, we estimated two control models that pooled the same static nonlinear function of the simple cell responses of the SC and GWP models. The static nonlinear function was a compressive nonlinearity (i.e. $\log(1+|s|)$ where s is a simple cell response). The compressive nonlinearity roughly accounts for insensitivities by increasing responses to a stimulus that is not entirely within a receptive field [39]. The simple cell responses were defined as the linear responses of the first layer of the SC model and the even-symmetric Gabor wavelets. While the performance of the compressive nonlinear SC model was significantly higher than that of the compressive nonlinear GWP model, the difference between the performance of the compressive nonlinear models was significantly lower than that of the SC2 and GWP2 models (Figure 10). This result suggests that both the representations and the nonlinearities of the SC2 model contribute to the difference between the encoding performance of the SC2 and GWP2 models.

To verify the contribution of the nonlinearities to the individual encoding performance of the SC2 and GWP2 models, we estimated two more control models that pooled a linear function of the simple cell responses of the SC and GWP models. We used linear models since they retain selectivities that are discarded by nonlinearities. We found that the performance of the linear models were significantly lower than that of the compressive nonlinear, SC2 and GWP2 models (Figure 10). This result confirms the contribution of the nonlinearities that introduced the insensitivities to the individual encoding performance of the SC2 and GWP2 models.

Discussion

This study addresses the question of how to model feature spaces to better predict brain activity. We introduced a general approach for making directly testable predictions of single voxel responses to statistically adapted representations of ecologically valid stimuli. Our approach relies on unsupervised learning of a feature model followed by supervised learning of a voxel model. To benchmark our approach against the conventional approach that makes use of predefined feature spaces, we compared a two-layer sparse coding model of simple and complex cells with a Gabor wavelet pyramid model of phase-invariant complex cells. While the GWP model is the fundamental building block of many state-of-the-art encoding and decoding models, the GWP2 model was found to be significantly outperformed by the SC2 model. We used control models to determine the contribution of the different components of the SC2 and GWP2 models to this performance difference. Analyses revealed that the SC2 model better accounts for both the representations and the nonlinearities of the voxels in the early visual areas than the GWP2 model. Given that the representations of the SC2 model are qualitatively similar to those of the GWP model, their contribution to this performance difference suggests that the SC model automatically learns an optimal set of spatially localized, oriented and bandpass representations that better span the space of early visual cortical representations since it adapts to the same statistical regularities in the environment as the brain is assumed to be adapted to [20].

Our approach eliminates the need for predefining feature spaces. However, the SC model does have a number of free parameters (e.g. patch size, number of simple and complex cells, etc.) that must either be specified by hand or using model selection methods such as cross-validation. Because of computational considerations, we used the same free parameters as those in [22]. While the choice of these free parameters can influence what the SC model can learn, the SC2 model was shown to outperform the GWP2 model even without cross-validation. Next to cross-validation, other methods that also infer these free parameters can further improve the performance of the SC2 model. One method is to first estimate voxel receptive fields using any approach and then use these estimates as free parameters (e.g. voxel receptive field eccentricity as patch size) of voxel-specific feature models. Another method is to use more sophisticated nonparametric Bayesian sparse factor models [40] that can simultaneously learn sparse representations while inferring their number. Furthermore, our approach included only feedforward projections such that representations and responses were solely determined by stimuli. However, taking top-down modulatory effects into account is

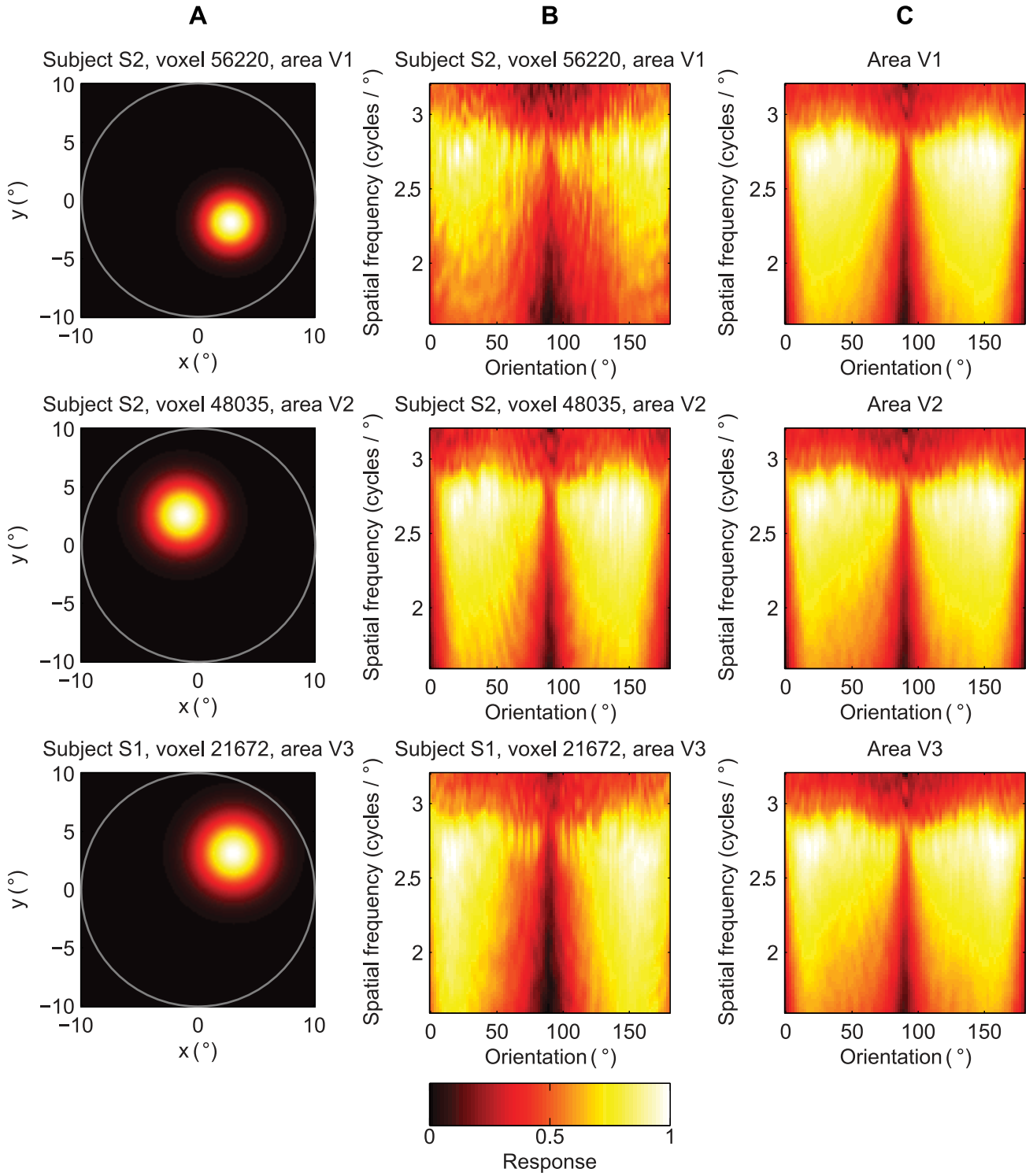

Figure 5. Receptive fields of the SC2 model. The parameter tuning varied across the voxels and had a bias for high spatial frequencies and oblique orientations. (A) Two-dimensional Gaussian functions that were fit to the responses of three representative voxels to point stimuli at different locations. (B) Responses of three representative voxels to sine-wave gratings that spanned a range of orientations and spatial frequencies. (C) Mean responses across the voxels to sine-wave gratings that spanned a range of orientations and spatial frequencies.
doi:10.1371/journal.pcbi.1003724.g005

essential to adequately characterize how sensory information is represented and processed in the brain. For example, attention has been shown to warp semantic representations across the human brain [41], and prior expectations have been shown to bias sensory representations in visual cortex [42]. Extensions of our approach that include feedback projections can be used to address the question of how representations and responses are influenced by top-down processes.

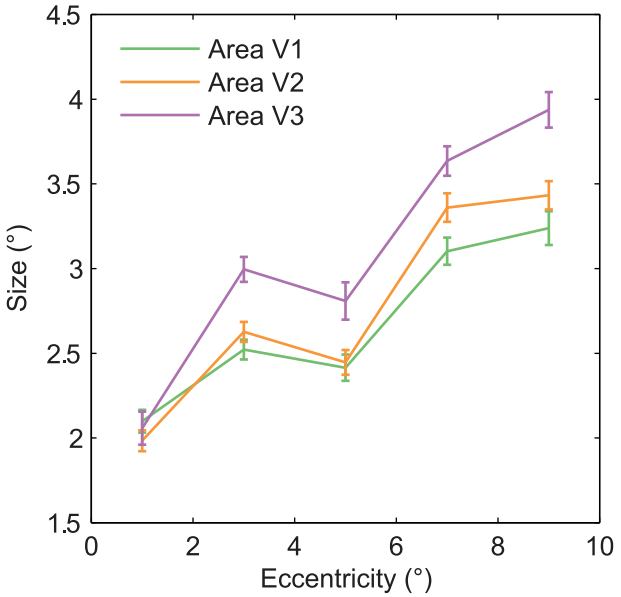

Figure 6. Receptive field size of the SC2 model as a function of receptive field eccentricity of the SC2 model and area. The eccentricity and size of the receptive fields were quantified as the mean and standard deviation of two-dimensional Gaussian functions that were fit to the voxel responses to point stimuli at different locations, respectively. The receptive field size systematically increased from low to high receptive field eccentricity and from area V1 to V3. Error bars show ± 1 SEM across the voxels (bootstrapping method).
doi:10.1371/journal.pcbi.1003724.g006

Further extensions of our approach can be used to probe mid- to high-level extrastriate visual cortical representations in a fully automated manner. In particular, the SC model can be replaced by highly nonlinear multi-layer statistical models of natural images that learn hierarchical feature spaces (i.e. deep learning [43]). Some of the feature spaces that are learned by these models such as mid-level edge junctions have been shown to match well with neural response functions in area V2 [44]. Models that learn even higher-level representations such as high-level object parts [45] or complete objects [46] can be used to probe extrastriate visual cortical representations. For example, heterogenous hierarchical convolutional neural networks have been shown to predict the representational dissimilarity matrices that characterize representations in human inferior temporal gyrus [47]. Similar models have been shown to learn feature spaces that are admitted by stimulus sets other than natural images, both within the visual modality (e.g. natural movies [48]) as well as in other modalities (e.g. auditory or somatosensory [49]). These models can be used to probe cortical representations in different sensory modalities.

One approach to estimate deep models is to maximize the likelihood of all layers at the same time. However, this approach is not scalable and requires the computation of intractable partition functions that are impossible to integrate analytically and computationally expensive to integrate numerically. Nevertheless, methods such as score-matching [50] and noise-contrastive estimation [51] have been used to estimate unnormalized nonlinear multi-layer statistical models of natural images [52,53]. An alternative approach is to use models such as deep belief networks that comprise multiple layers of restricted Boltzmann machines. These models can be scaled by convolution [45] and estimated by maximizing the likelihood of one layer at a time, using the output of each layer as input for the subsequent layer [54]. Importantly, generative models such as deep belief networks make it possible to sample stimuli based on internal network states. Conditioning these internal network states on stimulus-evoked brain activity results in a generative approach to decoding. For example, we have previously shown that a deep belief network that comprise multiple layers of conditional restricted Boltzmann machines can reconstruct handwritten digits by sampling from the model after conditioning it on stimulus-evoked multiple voxel responses [55].

While introducing a new approach to probe cortical representations, this study complements other developments in encoding and decoding. For example, encoding models that involve computations to account for contrast saturation or heterogeneous contrast energy were shown to improve prediction of single voxel responses to visual stimuli [16]. At the same time, these modeling efforts go hand in hand with developments in fMRI such as the improvements in contrast-to-noise ratio and spatial resolution that are facilitated by increases in magnetic field strength [56]. For example, spatial features of orientation-selective columns in humans were demonstrated by using high-field fMRI [57]. Jointly, such developments can provide novel insights into how cortical representations are learned, encoded and transformed.

In conclusion, we introduced a general approach that improves prediction of human brain activity in response to natural images. Our approach primarily relies on unsupervised learning of transformations of raw stimuli to representations that span the space of cortical representations. These representations can also be effectively exploited in stimulus classification, identification or reconstruction. Taken together, unsupervised feature learning heralds new ways to characterize the relationship between stimulus features and human brain activity.

Materials and Methods

Data

We used the fMRI data set [21] that was originally published in [8,13]. Briefly, the data set contained 1750 and 120 stimulus-response pairs of two subjects (i.e. S1 and S2) in the estimation and validation sets, respectively. The stimulus-response pairs consisted of grayscale natural images of size 128×128 pixels and stimulus-evoked peak BOLD hemodynamic responses of 5512 (S1) and 5275 (S2) voxels in the early visual areas (i.e. V1, V2 and V3). The details of the experimental procedures are presented in [8].

Problem statement

Encoding. Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^q$ be a stimulus-response pair where $\mathbf{x}$ is a vector of pixels in a grayscale natural image, and $\mathbf{y}$ is a vector of voxel responses. The parameters d and q denote the number of pixels and voxels, respectively. Given $\mathbf{x}$, we are interested in the problem of predicting $\mathbf{y}$:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\phi(\mathbf{x})) = \mathbf{B}^\top \phi(\mathbf{x}) \tag{1}$$

where $\hat{\mathbf{y}}$ is the predicted response to $\mathbf{x}$, and p is the encoding distribution of $\mathbf{y}$ given $\phi(\mathbf{x})$. The function ϕ nonlinearly transforms $\mathbf{x}$ from the stimulus space to the feature space, and $\mathbf{B}$ linearly transforms $\phi(\mathbf{x})$ from the feature space to the voxel space.

Decoding. Let $\mathbb{X}$ be a set of images that contains $\mathbf{x}$. Given $\mathbb{X}$ and $\mathbf{y}$, we are interested in the problem of identifying $\mathbf{x}$:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathbb{X}} \rho_{\mathbf{y}, \mathbf{B}^\top \phi(\mathbf{x})} \tag{2}$$

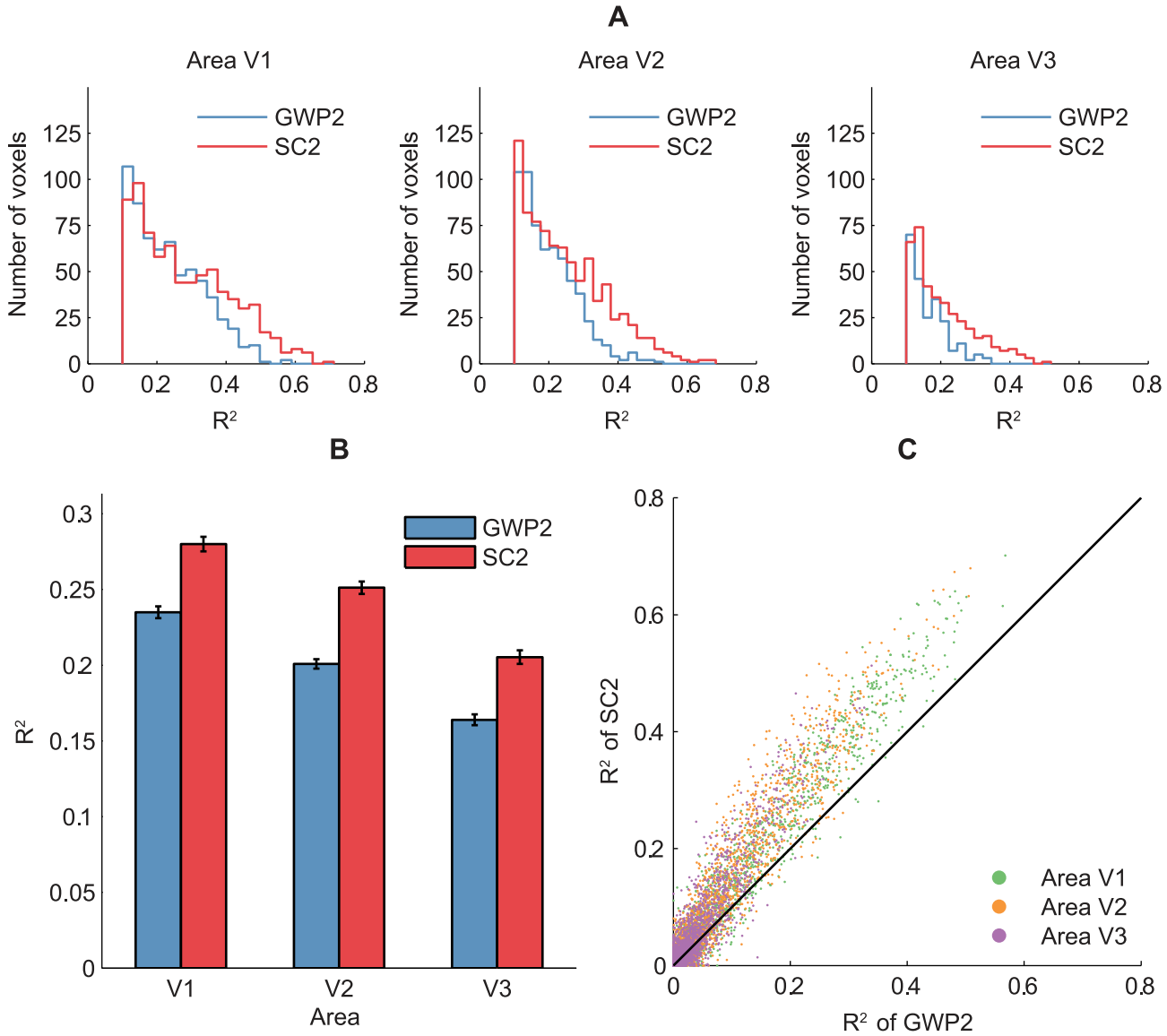

Figure 7. Encoding performance of the SC2 and GWP2 models. The encoding performance was defined as R^2 between the observed and predicted voxel responses to the 120 images in the validation set across the two subjects. The encoding performance of the SC2 model was significantly higher than that of the GWP2 model. (A) Prediction R^2 across the voxels that survived the R^2 threshold of 0.1. (B) Mean prediction R^2 across the voxels that survived the R^2 threshold of 0.1. Error bars show ± 1 SEM across the voxels (bootstrapping method). (C) Prediction R^2 in each voxel.

doi:10.1371/journal.pcbi.1003724.g007

where $\hat{\mathbf{x}}$ is the identified image from $\mathbf{y}$, and ρ is the Pearson product-moment correlation coefficient between $\mathbf{y}$ and $\mathbf{B}^\top \phi(\mathbf{x})$.

Solving the encoding and decoding problems requires the definition and estimation of a feature model ϕ followed by a voxel model $\mathbf{B}$.

Feature model

Model definition. Following [22], we summarize the definition of the SC model. We start by defining a single-layer statistical generative model of whitened grayscale natural image patches. Assuming that a patch is generated by a linear superposition of latent variables that are non-Gaussian (in particular, sparse) and mutually independent, we first use independent component analysis to define the model by a linear transformation of independent components of the patch:

$$\mathbf{z} = \mathbf{A}\mathbf{s} \tag{3}$$

where $\mathbf{z} \in \mathbb{R}^n$ is a vector of pixels in the patch, $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a mixing matrix, and $\mathbf{s} \in \mathbb{R}^m$ is a vector of the components of $\mathbf{z}$ such that $m \leq n$. The parameters n and m denote the number of pixels and components, respectively. We then define $\mathbf{s}$ by inverting the linear system that is defined by $\mathbf{A}$:

$$\mathbf{s} = \mathbf{W}\mathbf{z} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{m \times n}$ is an unmixing matrix such that $\mathbf{W} = \mathbf{A}^{-1}$. We constrain $\mathbf{W}$ to be orthonormal and s_i to have unit variance such that s_i are uncorrelated and unique, up to a multiplicative sign.

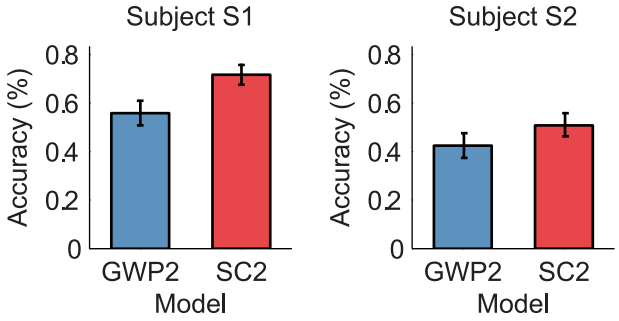

Figure 8. Decoding performance of the SC2 and GWP2 models. The decoding performance was defined as the accuracy of identifying the 120 images in the validation set from a set of 9264 candidate images. The decoding performance of the SC2 model was significantly higher than that of the GWP2 model. Error bars show ± 1 SEM across the images in the validation set (bootstrapping method). A more detailed figure that shows the identified images is provided at http://www.ccnlab.net/research/.
doi:10.1371/journal.pcbi.1003724.g008

Next, we define the joint probability of $\mathbf{s}$ by the product of the marginal probabilities of s_i since s_i are assumed to be independent:

$$p(\mathbf{s}) = \prod_{i=1}^{m} p(s_i) \tag{5}$$

where $p(s_i)$ are peaked at zero and have high kurtosis since s_i are assumed to be sparse.

While one of the assumptions of the model is that s_i are independent, their estimates are only maximally independent. As a result, residual dependencies remain between the estimates of s_i. We continue by modeling the nonlinear correlations of s_i since s_i are constrained to be linearly uncorrelated. In particular, we assume that the locally pooled energies of s_i are sparse. Without loss of generality, we first arrange s_i on a square grid graph that has circular boundary conditions. We then define the locally pooled energies of s_i by the sum of the energies of s_i that are in the same neighborhood:

$$\mathbf{c} = \mathbf{H}\mathbf{s}^2 \tag{6}$$

where $\mathbf{c} \in \mathbb{R}^m$ is a vector of the locally pooled energies of s_i and $\mathbf{H} \in \mathbb{R}^{m \times m}$ is a neighborhood matrix such that $h_{i,j} = 1$ if c_i pools the energy of s_j and $h_{i,j} = 0$ otherwise. Next, we redefine $\log p(\mathbf{s})$ in terms of $\mathbf{c}$ to model both layers:

$$\log p(\mathbf{s}) \approx \sum_{i=1}^{m} G(c_i) \tag{7}$$

where G is a convex function. Concretely, we use $G(c_i) = -\log(1 + c_i)$.

In a neural interpretation, simple and complex cell responses can be defined as $\mathbf{s}$ and a static nonlinear function of $\mathbf{c}$, respectively. Concretely, we use $\log(1+\mathbf{c})$ to define the complex cell responses after we estimate the model.

Model estimation. We use a modified gradient ascent method to estimate the model by maximizing the log-likelihood of $\mathbf{W}$ (equivalently, the sparseness of $\mathbf{c}$) given a set of patches:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z}) \tag{8}$$

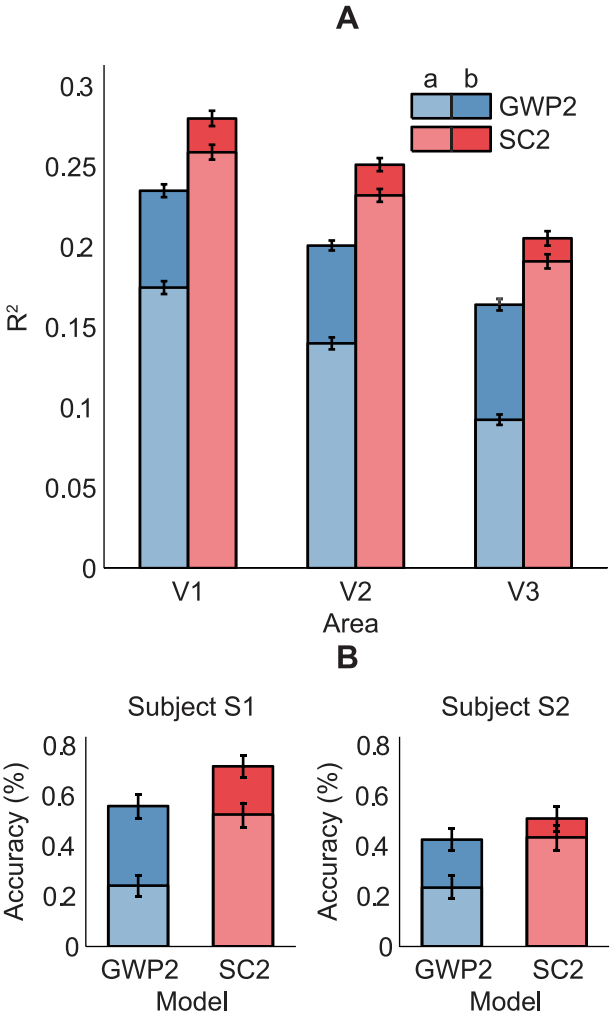

Figure 9. Mean prediction R^2 and identification accuracy of the SC2 and GWP2 models after (a) and before (b) translating the images in the validation set by 0.8° in a random dimension. The SC2 model was more invariant than the GWP2 model and its invariance increased from V1 to V3. (A) Mean prediction R^2 across the voxels that survived the R^2 threshold of 0.1 in the case of (b). Error bars show ± 1 SEM across the voxels (bootstrapping method). (B) Identification accuracy. Error bars show ± 1 SEM across the images in the validation set (bootstrapping method).
doi:10.1371/journal.pcbi.1003724.g009

where $\mathcal{L}(\mathbf{W}|\mathbf{Z}) = -\sum_{\mathbf{z}^{(i)}} \log p\left(\mathbf{H}\left(\mathbf{W}\mathbf{z}^{(i)}\right)^2\right)$ is an approximation of the log-likelihood of $\mathbf{W}$ and $\mathbf{Z} = \left(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots\right)$ is the set of patches. At each iteration, we first find the gradient of $\mathcal{L}(\mathbf{W}|\mathbf{Z})$:

$$\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z}) = -\mathbf{H}^{\top}\left(1 + \mathbf{H}(\mathbf{W}\mathbf{Z})^2\right)^{-1} \circ (2\mathbf{W}\mathbf{Z})\mathbf{Z}^{\top} \tag{9}$$

where $\circ$ is the Hadamard (element-wise) product. We then project it onto the tangent space of the constrained space [58]:

$$\overline{\nabla}_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z}) = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z}) - \mathbf{W}\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z})^{\top}\mathbf{W} \tag{10}$$

Next, we use backtracking line search to choose a step size by reducing it geometrically with a rate from (0,1) until the

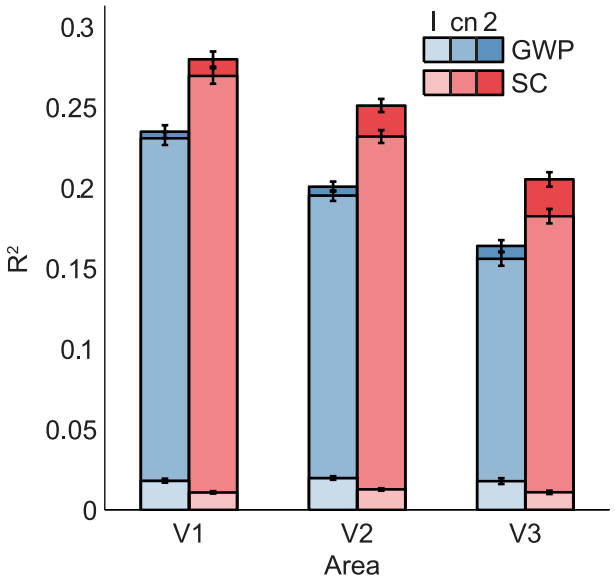

Figure 10. Mean prediction R^2 of the linear one-layer (l), compressive nonlinear one-layer (cn) and nonlinear two-layer (2) SC and GWP models across the voxels that survived the R^2 threshold of 0.1 in the case of (2). The mean prediction R^2 of the linear one-layer models were below the R^2 threshold of 0.1. The mean prediction R^2 of the nonlinear SC models were significantly better than those of the nonlinear GWP models. The compressive nonlinearity and the nonlinear second layer increased the mean prediction R^2 of the linear and compressive nonlinear models, respectively. The nonlinear second layer increased the mean prediction R^2 of the compressive nonlinear SC model more than it increased that of the compressive nonlinear GWP model. The error bars show ± 1 SEM across the voxels (bootstrapping method).
doi:10.1371/journal.pcbi.1003724.g010

Armijo-Goldstein condition holds [59]. Finally, we update $\mathbf{W}$ and find its nearest orthogonal matrix:

$$\mathbf{W} \leftarrow \mathbf{W} + \mu \overline{\nabla}_{\mathbf{W}} \mathcal{L}(\mathbf{W}|\mathbf{Z}) \tag{11}$$

$$\mathbf{W} \leftarrow \left(\mathbf{W}\mathbf{W}^{\top}\right)^{-\frac{1}{2}}\mathbf{W} \tag{12}$$

where μ is the step size.

Voxel model

Model definition. We start by defining a model for each voxel. Assuming that $p(\mathbf{y}|\phi(\mathbf{x})) \sim \mathcal{N}\left(\mathbf{B}^{\top}\phi(\mathbf{x}), \sum\right)$, where $\mathbf{B} = (\beta_1, \ldots, \beta_q) \in \mathbb{R}^{m \times q}$ and $\sum = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2) \in \mathbb{R}^{q \times q}$, we use linear regression to define the models by a weighted sum of $\phi(\mathbf{x})$:

$$y_i = \beta_i^{\top}\phi(\mathbf{x}) + \varepsilon_i \tag{13}$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$.

Model estimation. We estimate the model using ridge regression:

$$\hat{\beta}_i = \arg\min_{\beta_i} \frac{1}{N}\sum_{j=1}^{N}\left(y_i^{(j)} - \beta_i^{\top}\phi(\mathbf{x}^{(j)})\right) + \lambda_i ||\beta_{\mathbf{i}}||_2^2 \tag{14}$$

where $\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right)^{\top} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} = \left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}\right)^{\top} \in \mathbb{R}^{N \times q}$ is an estimation set, and $\lambda_i \geq 0$ is a complexity parameter that controls the amount of regularization. The parameter N denotes the number of stimulus-response pairs in the estimation set. We obtain $\hat{\beta}_{\mathbf{i}}$ as:

$$\hat{\beta}_i = (\lambda_i \mathbf{I}_m + \mathbf{\Phi}^{\top}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\top}\mathbf{Y}_i \tag{15}$$

where $\mathbf{\Phi} = \left(\phi\left(\mathbf{x}^{(1)}\right), \ldots, \phi\left(\mathbf{x}^{(N)}\right)\right)^{\mathrm{T}} \in \mathbb{R}^{N \times m}$ and $\mathbf{Y}_i = \left(y_i^{(1)}, \ldots, y_i^{(N)}\right)^{\mathrm{T}} \in \mathbb{R}^{N \times 1}$. Since $m \gg N$, we solve the problem in a rotated coordinate system in which only the first N coordinates of $\mathbf{\Phi}$ are nonzero [60,61]. We first factorize $\mathbf{\Phi}$ using the singular value decomposition:

$$\mathbf{\Phi} = \mathbf{U}\mathbf{S}\mathbf{V}^{\top} \tag{16}$$

where $\mathbf{U}\mathbf{U}^{\top} = \mathbf{U}^{\top}\mathbf{U} = \mathbf{I}_N, \mathbf{S} = \mathrm{diag}(\mathbf{s}) \in \mathbb{R}^{N \times N}$ and $\mathbf{V}^{\top}\mathbf{V} = \mathbf{I}_N$. The columns of $\mathbf{U}$, the diagonal entries of $\mathbf{S}$ and the columns of $\mathbf{V}$ are the left-singular vectors, the singular values and the right-singular vectors of $\mathbf{\Phi}$, respectively. We then reobtain $\hat{\beta}_i$ as:

$$\hat{\beta}_i = \mathbf{V}\,\mathrm{diag}\left(\frac{\mathbf{s}}{\mathbf{s} \circ \mathbf{s} + \lambda_i}\right)\mathbf{U}^{\top}\mathbf{Y}_i \tag{17}$$

where division is defined element-wise. The rotation reduces the complexity of the problem from $O(m^3)$ to $O(mN^2)$. To choose the optimal λ_i, we perform hyperparameter optimization using grid search guided by a generalized cross-validation approximation to leave-one-out cross-validation [60]. We define a grid by first sampling the effective degrees of freedom of the ridge regression fit from $[1, N]$ since its parameter space is bounded from above. The effective degrees of freedom of the ridge regression fit is defined as:

$$\mathrm{df}(\lambda_i) = \sum_{j=1}^{N}\frac{s_j^2}{s_j^2 + \lambda_i} \tag{18}$$

We then use Newton's method to solve df for λ_i. Once the grid is defined, we choose the optimal λ_i that minimizes the generalized cross-validation error:

$$\hat{\lambda}_i = \arg\min_{\lambda \in \Lambda}\left\{\sum_{j=1}^{N}\left[\frac{y_i^{(j)} - \hat{y}_i^{(j)}(\lambda)}{1 - \mathrm{df}(\lambda)/N}\right]^2\right\} \tag{19}$$

where Λ is the grid, and $\hat{y}_i^{(j)}(\lambda)$ is $\hat{y}_i^{(j)}$ given a particular λ.

Encoding and decoding

In the case of the SC model, each randomly sampled or non-overlapping patch was transformed to its principal components such that 625 components with the largest variance were retained and whitened prior to model estimation and validation. After the images were feature transformed, they were z-scored. The SC model of 625 simple and 625 complex cells was estimated from 50000 patches of size 32×32 pixels that were randomly sampled from the 1750 images of size 128×128 pixels in the estimation set. The details of the GWP model are presented in [8]. The SC2 and GWP2 models were estimated from the 1750 feature-transformed stimulus-response pairs in the estimation set.

Voxel responses to an image of size 128×128 pixels were predicted as follows. In the case of the SC model, each 16 non-overlapping patch of size 32×32 pixels of the image were first transformed to the complex cell responses of the SC model (i.e. total of 625 complex cell responses per patch and 10000 complex cell responses per image). The 10000 complex cell responses of the SC model were then transformed to the voxel responses of the SC2 model. In the case of the GWP model, the image was first transformed to the complex cell responses of the GWP model (i.e. total of 10921 complex cell responses per image). The 10921 complex cell responses of the GWP model were then transformed to the voxel responses of the GWP2 model. The encoding performance was defined as the coefficient of determination between the observed and predicted voxel responses to the 120 images in the validation set across the two subjects.

A target image was identified from a set of candidate images as follows. Prior to identification, 500 voxels were selected without using the target image. The selected voxels were those whose responses were predicted best. The target image was identified as the candidate image such that the observed voxel responses to the target image were most correlated with the predicted voxel responses to the candidate image (i.e. highest Pearson product-moment correlation coefficient between observed and predicted voxel responses). The decoding performance was defined as the accuracy of identifying the 120 images in the validation set from the set of 9264 candidate images. The set of candidate images contained the 120 images in the validation set and the 9144 images in the Caltech 101 data set [38].

Author Contributions

Conceived and designed the experiments: UG MAJvG. Performed the experiments: UG. Analyzed the data: UG. Contributed reagents/materials/analysis tools: UG MAJvG. Wrote the paper: UG MAJvG.

References

1. Dayan P, Abbott LF (2005) Theoretical Neuroscience: Computational And Mathematical Modeling of Neural Systems. Cambridge: MIT Press.
2. Brown EN, Kass RE, Mitra PP (2004) Multiple neural spike train data analysis: State-of-the-art and future challenges. Nat Neurosci 7: 456–461.
3. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435: 1102–1107.
4. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, et al. (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10: e1001251.
5. Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56: 400–410.
6. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293: 2425–2430.
7. Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8: 679–685.
8. Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452: 352–355.
9. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, et al. (2008) Predicting human brain activity associated with the meanings of nouns. Science 320: 1191–1195.
10. Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, et al. (2006) Inverse retinotopy: Inferring the visual content of images from brain activation patterns. Neuroimage 33: 1104–1116.
11. Miyawaki Y, Uchida H, Yamashita O, Sato Ma, Morito Y, et al. (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60: 915–929.
12. Schoenmakers S, Barth M, Heskes T, van Gerven M (2013) Linear reconstruction of perceived images from human brain activity. Neuroimage 83: 951–961.
13. Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63: 902–915.
14. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, et al. (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21: 1641–1646.
15. Vu VQ, Ravikumar P, Naselaris T, Kay KN, Gallant JL, et al. (2011) Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. Ann Appl Stat 5: 1159–1182.
16. Kay KN, Winawer J, Rokem A, Mezer A, Wandell BA (2013) A two-stage cascade model of BOLD responses in human visual cortex. PLoS Comput Biol 9: e1003079.
17. Barlow HW (1961) Possible principles underlying the transformations of sensory messages. In: Rosenblith WA, editor, Sensory communication, Cambridge: MIT Press. pp. 217–234.
18. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.
19. Bell AJ, Sejnowski TJ (1997) The "independent components" of natural scenes are edge filters. Vision Res 37: 3327–3338.
20. Hyvärinen A (2010) Statistical models of natural images and cortical visual representation. Top Cogn Sci 2: 251–264.
21. Kay KN, Naselaris T, Gallant JL (2011). fMRI of human visual areas in response to natural images. CRCNS.org.
22. Hyvärinen A, Hoyer PO (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Res 41: 2413–2423.
23. Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195: 215–243.
24. De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res 22: 545–559.
25. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58: 1233–1258.
26. Parker AJ, Hawken MJ (1988) Two-dimensional spatial structure of receptive fields in monkey striate cortex. J Opt Soc Am A Opt Image Sci Vis 5: 598–605.
27. Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A 2: 1160–1169.
28. Lee TS (1996) Image representation using 2D Gabor wavelets. IEEE Trans Pattern Anal Mach Intell 18: 959–971.
29. DeAngelis GC, Ghose GM, Ohzawa I, Freeman RD (1999) Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. J Neurosci 19: 4046–4064.
30. Hubel DH, Wiesel TN (1977) Ferrier lecture: Functional architecture of macaque monkey visual cortex. Proc R Soc Lond B Biol Sci 198: 1–59.
31. Blasdel G (1992) Orientation selectivity, preference, and continuity in monkey striate cortex. J Neurosci 12: 3139–3161.
32. Tootell R, Silverman M, Hamilton S, Switkes E, De Valois R (1988) Functional anatomy of macaque striate cortex. V. Spatial frequency. J Neurosci 8: 1610–1624.
33. Mansfield RJW (1974) Neural basis of orientation perception in primate vision. Science 186: 1133–1135.
34. Furmanski CS, Engel SA (2000) An oblique effect in human primary visual cortex. Nat Neurosci 3: 535–536.
35. Swisher JD, Gatenby JC, Gore JC, Wolfe BA, Moon CH, et al. (2010) Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. J Neurosci 30: 325–330.
36. Dumoulin SO, Wandell BA (2008) Population receptive field estimates in human visual cortex. Neuroimage 39: 647–660.
37. Smith A, Singh K, Williams A, Greenlee M (2001) Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. Cereb Cortex 11: 1182–1190.
38. Fei-Fei L, Fergus R, Perona P (2007) Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Comput Vis Image Underst 106: 59–70.
39. Kay KN, Winawer J, Mezer A,Wandell BA (2013) Compressive spatial summation in human visual cortex. J Neurophysiol 110: 481–494.
40. Knowles D, Ghahramani Z (2011) Nonparametric Bayesian sparse factor models. Ann Appl Stat 5: 1534–1552.
41. Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16: 763–770.
42. Kok P, Brouwer GJ, van Gerven MA, de Lange FP (2013) Prior expectations bias sensory representations in visual cortex. J Neurosci 33: 16275–16284.
43. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35: 1798–1828.
44. Lee H, Ekanadham C, Ng A (2007) Sparse deep belief net model for visual area V2. In: Neural Information Processing Systems.
45. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: International Conference on Machine Learning.
46. Le Q, Ranzato M, Monga R, Devin M, Chen K, et al. (2012) Building high-level features using large scale unsupervised learning. In: International Conference on Machine Learning.
47. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A.

48. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Conference on Computer Vision and Pattern Recognition.
49. Saxe AM, Bhand M, Mudur R, Suresh B, Ng AY (2011) Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In: Neural Information Processing Systems.
50. Hyvärinen A (2005) Estimation of non-normalized statistical models by score matching. J Mach Learn Res 6: 695–709.
51. Gutmann MU, Hyvärinen A (2012) Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. J Mach Learn Res 13: 307–361.
52. Köster U, Hyvärinen A (2010) A two-layer model of natural stimuli estimated with score matching. Neural Comput 22: 2308–2333.
53. Gutmann MU, Hyvärinen A (2013) A three-layer model of natural image statistics. J Physiol Paris 107: 369–398.
54. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18: 1527–1554.
55. van Gerven MAJ, de Lange FP, Heskes T (2010) Neural decoding with hierarchical generative models. Neural Comput 22: 3127–3142.
56. Duyn JH (2012) The future of ultra-high field MRI and fMRI for study of the human brain. Neuroimage 62: 1241–1248.
57. Yacoub E, Harel N, Uğurbil K (2008) High-field fMRI unveils orientation columns in humans. Proc Natl Acad Sci U S A 105: 10607–10612.
58. Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal A 20: 303–353.
59. Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge: Cambridge University Press.
60. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.
61. Murphy KP (2012) Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press.