Inference via sparse coding in a hierarchical vision model

Joshua Bowren Department of Computer Science, University of Miami, Coral Gables, FL, USA

Luis Sanchez-Giraldo Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, USA

Odelia Schwartz Department of Computer Science, University of Miami, Coral Gables, FL, USA

Sparse coding has been incorporated in models of the visual cortex for its computational advantages and connection to biology. But how the level of sparsity contributes to performance on visual tasks is not well understood. In this work, sparse coding has been integrated into an existing hierarchical V2 model (Hosoya & Hyvärinen, 2015), but replacing its independent component analysis (ICA) with an explicit sparse coding in which the degree of sparsity can be controlled. After training, the sparse coding basis functions with a higher degree of sparsity resembled qualitatively different structures, such as curves and corners. The contributions of the models were assessed with image classification tasks, specifically tasks associated with mid-level vision including figure–ground classification, texture classification, and angle prediction between two line stimuli. In addition, the models were assessed in comparison with a texture sensitivity measure that has been reported in V2 (Freeman et al., 2013) and a deleted-region inference task. The results from the experiments show that although sparse coding performed worse than ICA at classifying images, only sparse coding was able to better match the texture sensitivity level of V2 and infer deleted image regions, both by increasing the degree of sparsity in sparse coding. Greater degrees of sparsity allowed for inference over larger deleted image regions. The mechanism that allows for this inference capability in sparse coding is described in this article.

Introduction

Computational models of the visual cortex have progressed significantly over the past few decades. One approach to modeling cortical neurons, denoted as goal-oriented (or supervised) learning, is based on optimizing model goals such as image classification (see e.g., review papers, Geisler, 2008; Yamins & DiCarlo, 2016; Turner et al., 2019). In recent years, deep neural network models optimized for image classification (e.g., Krizhevsky et al., 2012; Dapello et al., 2020) have captured neural processing in cortical visual areas (Kriegeskorte, 2015; Yamins & DiCarlo, 2016), including low and mid-level visual cortex (e.g., Pospisil et al., 2018; Cadena et al., 2019; Kindel et al., 2019; Laskar et al., 2020).

Another approach for modeling cortical visual neurons which is the focus here is denoted as stimulus-oriented (or unsupervised) learning. In particular, it has been hypothesized that neurons are matched to the statistical properties of images in the environment (Attneave, 1954; Barlow, 1961; Simoncelli & Olshausen, 2001) by optimizing statistical constraints such as sparsity or coding efficiency. For instance, the sparse coding model of Olshausen and Field (1996), and models of independent component analysis (ICA; Bell & Sejnowski, 1995; Hyvärinen & Oja, 1997), offered a principled mechanism for the derivation of oriented filters qualitatively similar to simple cells in the primary visual cortex (area V1). Other investigators have proposed methods of deriving models of V1 complex cell responses (Hyvärinen & Hoyer, 2001; Berkes & Wiskott, 2005; Karklin & Lewicki, 2009), deriving V2 model responses from V1 responses (Lee et al., 2007; Coen-Cagli & Schwartz, 2013; Shan & Cottrell, 2013; Hosoya & Hyvärinen, 2015), and hierarchical nonlinear models that learn patterns of statistical dependencies (Karklin & Lewicki, 2005). Stimulus-oriented approaches have also been adapted to deep neural networks with success in capturing aspects of the ventral visual cortex (Zhuang et al., 2021).

In addition to bottom-up approaches for optimizing statistical constraints, stimulus-oriented approaches can be closely tied to top-down generative approaches describing the process by which the signals are generated (Rao et al., 2002). For instance, sparse coding can be seen both from the perspective of optimizing sparseness

Citation: Bowren, J., Sanchez-Giraldo, L., & Schwartz, O. (2022). Inference via sparse coding in a hierarchical vision model. *Journal of Vision*, *22*(2):19, 1–19, https://doi.org/10.1167/jov.22.2.19.

https://doi.org/10.1167/jov.22.2.19 Received February 28, 2021; published February 25, 2022 ISSN 1534-7362

and as a generative model of images (Olshausen & Field, 2006). Generating and inferring image structure is also an important aspect of vision (Yuille & Kersten, 2006) in addition to classifying images. Although a major emphasis in computer vision has been on image classification, other ideas exist for how inference and other capabilities may be achieved in image models (Pei & Zeng, 2006; Zhaoping & Jingling, 2008; Goodfellow et al., 2014; Luo et al., 2015; Radford et al., 2015; Svanera et al., 2021).

The aim of this work is to investigate how sparse coding can be explicitly integrated into a V2 model (Hosoya & Hyvärinen, 2015) to introduce an inference mechanism (discussed elsewhere in this article); test its performance on an inference task; test its performance on image classification tasks spanning line combinations, figure–ground classification, and texture classification; and compare the model's texture sensitivity with the texture sensitivity of V2 as reported in the (fMRI) results of Freeman et al. (2013). This approach was taken, as opposed to a complete vision model capable of inference like a generative adversarial network, because it allowed for the principle of a sparse prior to be studied in a model V2 stage when holding lower-level stages (V1 and V1 complex) constant with respect to the effect of the sparse prior.

The work in Hosoya and Hyvärinen (2015) includes a model of V1 complex cell responses and a dimensionality reduction stage, followed by a version of independent component analysis for overcomplete codes and rectification to form V2-like model neurons. Although ICA results in filter responses with high kurtosis, it does not explicitly optimize for sparsity. In addition, studies comparing ICA and sparse coding in the overcomplete case in a single-layer model have found differences (Livezey et al., 2019). Therefore, the purpose of this work was to understand the implications of incorporating an explicit sparse coding at the last stage of the model, for which the sparsity level could also be controlled.

The current understanding of V2 characterizes its receptive fields in terms of its response properties, some of which include cross-orientation suppression (Rowekamp & Sharpee, 2017), selectivity for angles (e.g., Ito & Komatsu, 2004), selectivity for figure and ground (von der Heydt & Peterhans, 1989; Peterhans & von der Heydt, 1989; Zhaoping, 2005), and selectivity for texture (Freeman et al., 2013; Ziemba et al., 2016; Kohler et al., 2016). However, the receptive fields of V2 neurons are not fully understood, and there is no consensus that any one model best explains V2 units. The model of Hosoya and Hyvärinen (2015) can capture some properties of V2 neurons, but in this article, our primary goal was to highlight practical vision capabilities rather than to compare with neural data. The sparse coding model works by finding a dictionary of basis functions such that only a few are needed to reconstruct any given image. Sparse coding was successful for modeling V1, so some investigators continued to perform sparse coding twice to model V2 (e.g., Lee et al., 2007). Others have looked at hierarchical nonlinear generative sparse coding models (Karklin & Lewicki, 2005). Here, traditional sparse coding (Olshausen & Field, 1996) was performed in the V2-stage of a V2-like model.

The focus of this work was on a hierarchical visual cortical model with sparse coding because sparse coding has various computational advantages (Willshaw et al., 1969; Kanerva, 1992), is biologically plausible (Field, 1994; Olshausen et al., 2003; Olshausen & Field, 2004; Rozell et al., 2008), and sparse firing has been observed in visual cortical neurons in response to images (e.g., see Willmore et al., 2011; Yoshida & Ohki, 2020), although see also the discussion in the work by Berkes et al. (2009). Although the sparse coding model of Olshausen and Field (1996) is not the only method of achieving a sparse neural representation, its underlying generative model provides a coding strategy that models neuron responses as contributions of basis functions that sum to reconstruct the input image rather than linear filter responses to that image. This process allows for inference via the mechanism discussed next.

The original approach by Hosoya and Hyvärinen (2015) performed a variation of ICA for overcomplete codes (here referred to as overcomplete ICA) as its final V2-stage computation to derive an overcomplete sparse representation, but sparse coding provides several appealing computational differences. First, although the generative model of sparse coding is linear, its forward transformation is nonlinear (the solution to an optimization problem). By comparison, the forward transform of ICA is linear (multiplication by a filter matrix). Second, unlike ICA, the sparse coding algorithm allows for explicit control of the degree of sparsity. Third, sparse coding explicitly learns a dictionary of basis functions for which each of the model's responses is interpreted as the contribution of a single basis function to the image.

The degree of sparsity enforced by the L1 regularization coefficient of sparse coding allows the model to focus more on either a faithful reconstruction with low values or structure inference with high values. With low values (low sparsity), many basis functions are available and the image is reconstructed almost exactly. With high values (high sparsity), only a few basis functions constitute the image reconstruction, and each individual basis function must do more to explain the image (minimize reconstruction error). In practice, reconstruction error increases with fewer basis functions, but more latent information is introduced into the reconstruction. The idea, although counterintuitive, is that higher error may be advantageous. The error allows for missing image information owing to events such as occlusion to be

discarded in the model's representation of the image, and the model instead explains the image from an incomplete set of input responses. For this reason, this mechanism is referred to here as the model's *inductive inference mechanism*.

Both the overcomplete ICA and non-negative sparse coding based models were tested on three classification tasks: a figure–ground detection task, a texture classification task, and an angle discrimination task (see the Methods for details). Classification was performed by training a linear support vector machine (SVM) on the V2-stage responses generated by both models to give a sense of how linearly separable the image classes were in the V2-stage representation space. Although a good performance on these tasks is probably characteristic of a good initial (low-level) vision model, vision models should be expected to perform a large range of functions necessary for understanding the world. One might expect tasks like noise removal, image completion, and content generation to be necessary to compete with the wide range of tasks the human visual system can perform. One such task was explored here: the ability of both models to fill in missing image information when deleted midway through the visual processing pipeline. Good performance on this task would suggest an image understanding beyond an association with labels and provide evidence that the inductive inference mechanism postulated here might benefit other vision models.

The novel contribution of this work is an understanding of the importance of vision tasks such as image classification, image inference, and texture sensitivity, and their implications for model performance. Non-negative sparse coding was found to perform worse on the popular computer vision metric of image classification, but was more closely matched with V2 in terms of texture sensitivity. Non-negative sparse coding also better inferred deleted image regions in the image inference experiment with the proper value of the regularization coefficient. These results highlight some of the tradeoffs of sparse coding with different sparsity levels for the range of tasks. Also, sparse coding with a larger regularization coefficient (i.e., a larger sparsity level) is viewed here as providing an enhancement rather than only a degradation of the model. Although reconstruction error becomes higher with a larger regularization coefficient and may seem undesirable, it is proposed here that such a strategy may be useful for vision.

Methods

This work builds on the hierarchical unsupervised learning V2 model of Hosoya and Hyvärinen (2015). Non-negative sparse coding similar to that of Hoyer (2002) was incorporated in place of overcomplete ICA in order to maintain the non-negative response property of the original model. Also, the results were compared with the original overcomplete ICA based model. Hosoya and Hyvärinen (2015) made ICA overcomplete by increasing the number of independent components in the loss function beyond the number of inputs, then estimating the components with score matching according to Hyvärinen (2005). The same loss function in Hyvärinen (2005) was minimized here for the original model.

The generative model of sparse coding models images as sparse linear combinations of a set of basis functions given by the matrix $\mathbf{\Phi}$ called a dictionary:

$$\mathbf{x} = \mathbf{\Phi a} \quad (1)$$

where the vector $\mathbf{x}$ is the image and the vector $\mathbf{a}$ combines columns of $\mathbf{\Phi}$ and contains mostly zeros (sparse). Because there are a few variations of sparse coding, we define our precise method here. Non-negative sparse coding was performed with scikit-learn (Pedregosa et al., 2011, version 0.20.3) by inferring the basis function matrix $\mathbf{\Phi}$ such that

$$\mathbf{\Phi} = \arg\min_{\mathbf{\Phi}} \|\mathbf{X} - \mathbf{\Phi A}\|_{\mathrm{F}}^2 \mid \|\mathbf{\Phi}_i\|_2 = 1, \forall i \quad (2)$$

where $\mathbf{X}$ is a matrix with column input image vectors $\mathbf{x}_i$, $\mathbf{A}$ is a matrix of sparse coefficient column vectors $\mathbf{a}_i$, which are functions of $\mathbf{\Phi}$ and $\mathbf{x}_i$, and the operation $\| \ \|_F$ is the Forbenius norm. During each training step, first the sparse coefficient vector $\mathbf{a}_i$ for each input vector $\mathbf{x}_i$ is inferred via LASSO by choosing $\mathbf{a}_i$ such that

$$\mathbf{a}_i = \arg\min_{\mathbf{a}_i} \|\mathbf{x}_i - \mathbf{\Phi a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1 \mid a_{i,j} > 0, \forall i, j, \quad (3)$$

where the hyperparameter λ determines the level of sparsity in the non-negative sparse coding representation. The objective in Equation 3 is minimized via coordinate descent with the current value of the basis function matrix $\mathbf{\Phi}$. After inferring all $\mathbf{a}_i$, the basis function matrix is updated with one step of coordinate descent according to its objective in Equation 2. This process was repeated until there was little change in the appearance of the basis function visualizations (discussed elsewhere in this article). This was similar to the method of Olshausen and Field (1996), but with a non-negativity constraint.

In the original model, overcomplete ICA was followed by rectification to constrain the model V2 responses to be non-negative. The same could be done for sparse coding, but a more natural approach was available via non-negative sparse coding, a method that constrains the responses of sparse coding to be either zero or positive. Non-negative sparse coding

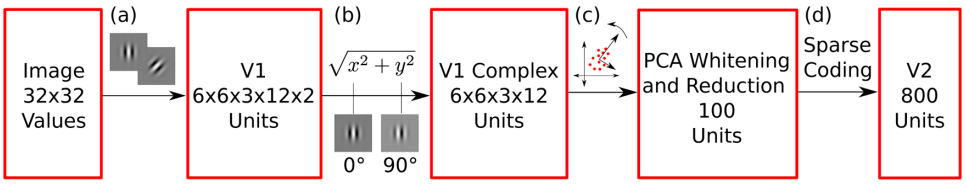

Figure 1. Hierarchical V2 model with sparse coding. (a) The model begins by computing the responses of the Gabor filters (3 frequencies, 12 orientations, and 2 phases) over the 6×6 central spatial locations. (b) Next, the V1 responses are pooled by taking the square root of the energy of each pair of filters that are 90° out of phase. (c) The pooling is followed by PCA whitening and reduction down to 100 components. (d) Finally, the representation is expanded by a factor of eight times with non-negative sparse coding.

is not equivalent to sparse coding with rectification, but constrains the model to find basis functions that combine without inverting contrast (with negative coefficients). The introduction of sparse coding into the overall V2 model also introduces an additional hyperparameter: the L1 regularization coefficient. The L1 regularization coefficient controls the degree of sparsity in the sparse coding responses. Larger values result in fewer active (non-zero) units for reconstructions. Several values for the L1 regularization coefficient in the range of [0.1, 4.0] were explored, including values common for discovering basis functions that are similar to Gabor wavelets in traditional sparse coding as well as much larger values. Values of 0.5 and 4.0 were of particular interest because they maximized the performance on the later classification tasks and forced the model to usually recruit only a few basis functions, respectively. The latter approximately maximizes prior information in each individual basis function. This approach is compared with the original ICA model qualitatively with V2 unit visualizations and quantitatively with their performances on several vision tasks. The overcomplete ICA model was less sparse than non-negative sparse coding with a regularization coefficient of 0.5, but we noticed that an approximate ceiling was reached. As the regularization coefficient decreased, the classification accuracy increased until a value of about 0.5.

The models (see Figure 1 for an illustration with sparse coding) were trained on 400,000 32×32 image patches from ImageNet ILSVRC12 (Russakovsky et al., 2015). The patches were randomly sampled from images after subtracting the mean and normalizing the variance of the images. Low-contrast patches were not included (variance of less than 0.32), as was done in Hosoya and Hyvärinen (2015). The mean of each patch was also subtracted and its variance normalized. The overcomplete ICA and non-negative sparse coding models were trained for 16 epochs (presentations of the whole training set). The model hyperparameters were matched to that of Hosoya and Hyvärinen (2015). The log of the probability density function (the function "G") of the input under the overcomplete ICA model was the negative log of the hyperbolic cosine function. The model V1 simple cell responses were computed with Gabor filters along the 6×6 center locations of each 32×32 image patch. This is equivalent to the two-dimensional convolution of the Gabor filters with the image with a stride of 4 and no padding around the edges of the image. There were 3 frequencies (1.25 cycles, 1.5 cycles, 1.75 cycles), 12 orientations (increments of 15° from 0° to 165°), and 2 phases (0° and 90°). The filters had a receptive field size of approximately 12×12 pixels. The resulting set of model V1 simple cell responses for the location and parameter choices had a dimension of (6, 6, 3, 12, 2) responses. The model V1 complex cell responses were computed by taking the square root of the sum of the squares of each quadrature (90° out of phase) pair of Gabor functions to model phase invariance. The resulting model V1 complex cell responses had a dimension of (6, 6, 3, 12) because the last dimension of the model V1 simple cell responses corresponded to the quadrature pair. Before computing the model V2 responses, the model V1 complex cell responses were pooled with principal component analysis (PCA) by maintaining only the 100 components with the largest eigenvalues. Finally, the V2 responses were computed with overcomplete ICA or non-negative sparse coding with 800 filters or basis functions. The source code for the complete V2 model has been made available at https://notabug.org/jbowren/hv2model.

The model is illustrated in Figure 1. The number of components and V2 units matches that of Hosoya and Hyvärinen (2015). A second configuration with 11×11 spatial locations and 350 principal components (for the increase in Gabors) was also explored for the inference experiment in order to reconstruct entire patches. The number of V2 units chosen was 2,800 to keep the representation 8 times overcomplete.

The V2 model neurons were visualized in a similar fashion to that of Hosoya and Hyvärinen (2015). First, a 1-of-K representation was inserted into the model as the V2-stage responses where the unit to be visualized is set to 1 and every other unit is set to 0. Next, the model proceeded backward until the corresponding V1 complex responses were obtained. This representation was then plotted in the input 32×32 image space

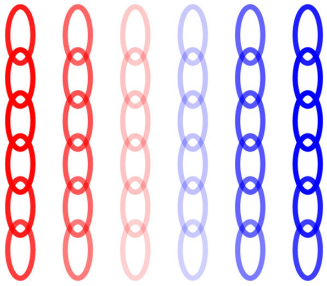

Figure 2. V2 model unit visualization. There are 6×6 spatial locations, 3 frequencies, and 12 orientations. Opacity reflects the strength of the response and color the sign with red being excitatory and blue inhibitory. Location represents the location of where the Gabor was applied in the image. Orientation represents the orientation of the Gabor. Size represents the frequency of the Gabor (larger size smaller frequency).

with ovals drawn over the 6×6 center locations in the 32×32 image space. The opacity of the ovals represents the strength of the responses, the color indicates the sign of the response (red for excitatory and blue inhibitory), the size of the ovals reflect the frequency of the Gabors, and the orientation of the ovals represents the orientation of the Gabors. Excitatory (red) Gabors signify the presence of stimuli with the same orientation and frequency (size). Inhibitory (blue) Gabors signify that stimuli with the same orientation and frequency should be absent to help excite the unit. An example of a V2 model unit is shown in Figure 2. In addition to this visualization, the six 32×32 image patches that maximally activate each unit are shown in the Results to provide insight into the representation of the units.

The resulting V2 models with non-negative sparse coding and overcomplete ICA were run on three different image datasets. The datasets were selected because they are considered to capture mid-level visual tasks that could be appropriate at the V2 level, namely to perform figure–ground classification, texture classification, and to predict the angle between two lines connected at one end–point. Classification was performed by training a linear SVM (with the SVC classifier of Scikit-Learn; Pedregosa et al., 2011) on the model responses of each model configuration. The choice of the regularization coefficient could influence the results, so a few values of the regularization coefficient for the linear SVM were tested as well as logistic regression model with the same values of the regularization coefficient (and without a regularization coefficient), but ICA consistently performed the best outside of the error standard deviation bars across all configurations and datasets. We include the classification results for these models and all the values of the regularization coefficient we tested in Appendix A. Because the regularization coefficient did not change the model that performed best, we simply report the result of the SVM classification with

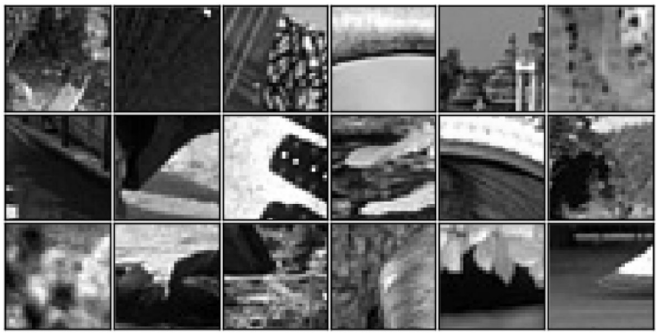

Figure 3. The 32×32 figure–ground image patches. Example 32×32 image patches sampled from the Berkeley Segmentation Database (BSDS300; Fowlkes et al., 2007). The label for each image patch indicates whether the figure of the image falls primarily on the right side or left side of the image. The labels were determined from the corresponding 32×32 region of the human-drawn contour line map for each image.

a regularization coefficient of 1. For each dataset, five-fold cross-validation was performed and the average accuracy and standard deviation were recorded.

For the figure–ground experiment, the images and figure–ground labelings were obtained from the Berkeley Segmentation DataSet (BSDS300; Fowlkes et al., 2007). Figure and ground refer to regions of images separated by some contour in the image that determines the main region of focus for an observer. The region denoted as the figure is the main region of focus that might grab the foveal attention of an observer, whereas the ground is considered to provide context to the figure. Neural processing is thought to have a mechanism of distinguishing between figure and ground regions (see Coen-Cagli & Schwartz, 2013, for a more in-depth description). Fowlkes et al. (2007) showed that the regions of images with figure rather than ground were usually smaller and more convex, so these convex regions probably require more advanced features for image classification. A total of 20,000 32×32 image patches (see Figure 3) were randomly sampled from the dataset with labels (figure or ground) assigned based on which side of the image (left or right) the human-labeled contour primarily fell. The second experiment included synthetically generated images according to Portilla and Simoncelli (2000). The images were generated to match the low order statistics of different classes of real texture images from the Brodatz dataset (Brodatz, 1966). The models were tested on 30,000 32×32 patches sampled from 15 texture categories; an example of a few texture patch families are shown in the left column of Figure 4. The last classification experiment tested the models on line segments joined at one endpoint with varying lengths, locations, rotations, and angles between the two lines. There are a total of 3 lengths (10, 15, and 20 pixels), 9 locations (the 3×3 center locations in the image), 12

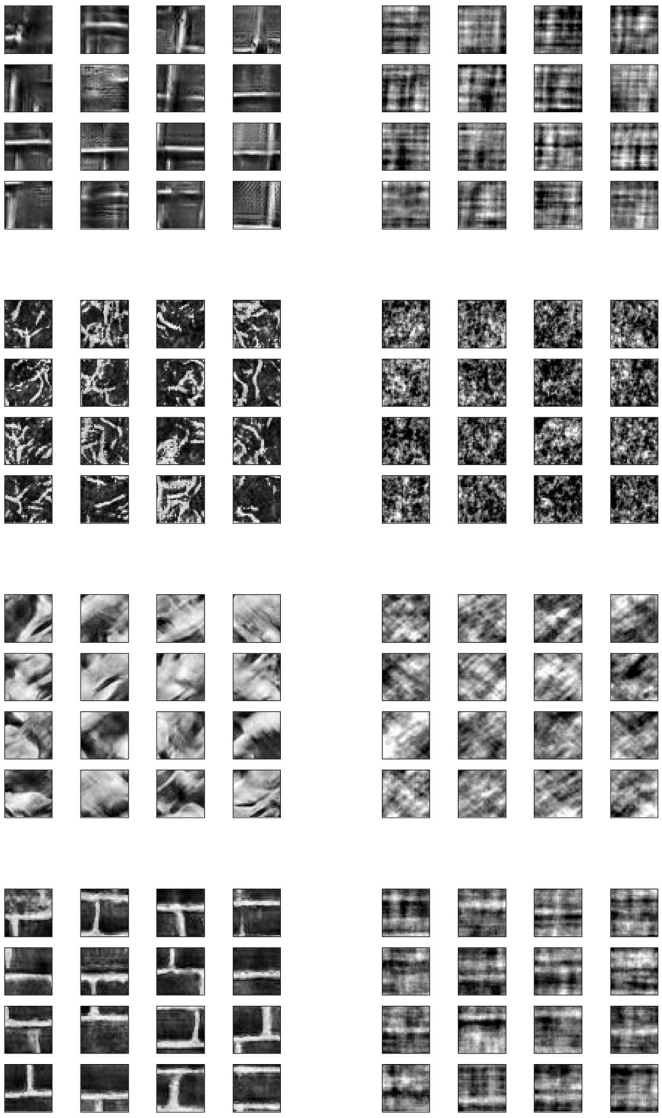

Figure 4. The 32×32 texture and noise patches extracted from 256×256 generated texture and noise images. These 32×32 patches were extracted from 256×256 Brodatz texture images generated with the code made available by Portilla and Simoncelli (2000). The column on the left shows 4×4 grids of 32×32 patches where each grid corresponds with a different texture. The column on the right shows the corresponding spectrally matched noise versions of the texture patches extracted at the same locations within the spectrally matched noise versions of the texture images. A complete set of the full size 256×256 texture and noise images are shown in Appendix B.

rotations (0° to 330° with an interval of 30°), and 6 angles (30° to 180° with an interval of 30°). Examples of the line stimuli are shown in Figure 5.

Next, the texture sensitivity of the models was measured via the texture modulation index (Freeman et al., 2013) computed with 30,000 32×32 texture patches and 30,000 32×32 spectrally matched noise

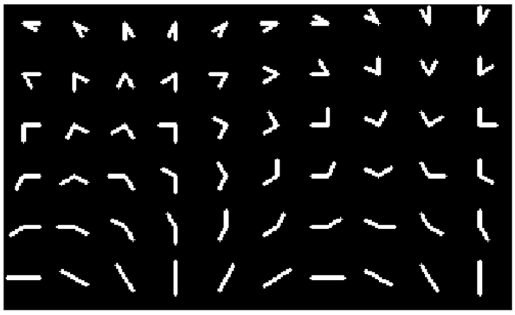

Figure 5. The 32×32 line stimuli. Two lines connected at one endpoint with 3×3 center locations, 6 angles, 12 rotations, and 3 lengths. Each line image is a 32×32 image.

versions (see end of paragraph) of the same patches. The texture modulation index is a measure of texture sensitivity in the range of [−1, 1] where 1 indicates the maximal sensitivity for texture and −1 the opposite. The texture modulation index is calculated by taking the difference of the responses of a model (or brain region) to texture stimuli and noise stimuli, then normalizing by the sum of the two. Here, the index was averaged over all of the model neurons. The equation for calculating the modulation index M is given by

$$M = \frac{r_{tex} - r_{noise}}{r_{tex} + r_{noise}}, \quad (4)$$

where r_{tex} is the response to a texture stimuli and r_{noise} is the response to a spectrally matched noise version of the texture patch. The 30,000 texture patches were taken from the same textures in the classification experiment. The corresponding 30,000 spectrally matched noise patches were obtained by taking 32×32 patches at the same locations of the texture patches from spectrally matched noise versions of the original 256×256 texture stimuli. The spectrally matched noise versions of the original texture stimuli were generated by first computing the magnitude and phase of a fast-Fourier transform (FFT) of each texture and a corresponding randomly generated Gaussian white noise image. Next, the phase component of the original texture image was replaced with the phase component of the Gaussian white noise image. Finally, the spectrally matched noise images were obtained by performing the inverse FFT on the new magnitude and phase representation. This ensures that the magnitude of the FFT of the spectrally matched noise image is the same as the magnitude of the FFT of the synthesized texture image with uniform random phase (see Galerne et al., 2010).

Next, the ability of the models to fill in missing information was tested by deleting part of the image representation within the model before non-negative sparse coding or overcomplete ICA, then going backward to reconstruct the image from the models'

responses. The backward computation proceeded by multiplying the V2 responses by the sparse coding dictionary (or ICA mixing matrix for overcomplete ICA) and the inverse PCA whitening matrix to recover the V1 complex responses. Next, the V1 simple responses were calculated from the V1 complex responses and the angles between each pair of responses (in polar coordinates) that were saved in the forward computation; this practice does not affect the relative reconstruction accuracy of the two models. Finally, the image was reconstructed by convolving the V1 simple responses with the transpose of the original Gabor filters. This is an approximation, but computationally practical to undo the forward Gabor filter transform. A total of 1,000 image patches were sampled from ImageNet and fed forward through the model until model V1 complex responses were obtained. Deletion was then performed by setting either a 1×1 or 2×2 region of the V1 complex responses to the minimum value of the responses minus 1. This corresponded with approximately a 2×2 or 6×6 region, respectively, in the original image space. Next the responses were filtered with the PCA whitening matrix, then fed through either non-negative sparse coding (with a regularization coefficient of 2.0 or 4.0) or overcomplete independent component analysis. A variety of values of the sparse coding regularization coefficient in the range [0.1, 4.0] were tested, and the pair 2.0 and 4.0 was shown instead of 0.5 and 4.0, because a regularization coefficient of 0.5 was not large enough to significantly change the input representation. Finally, the transformations were undone as detailed above to get the models' representation in the original image space. The reconstructions were visually inspected side-by-side to the original image patches, and the average mean-squared error (MSE) of the representations and the original image patches was computed for both models and compared.

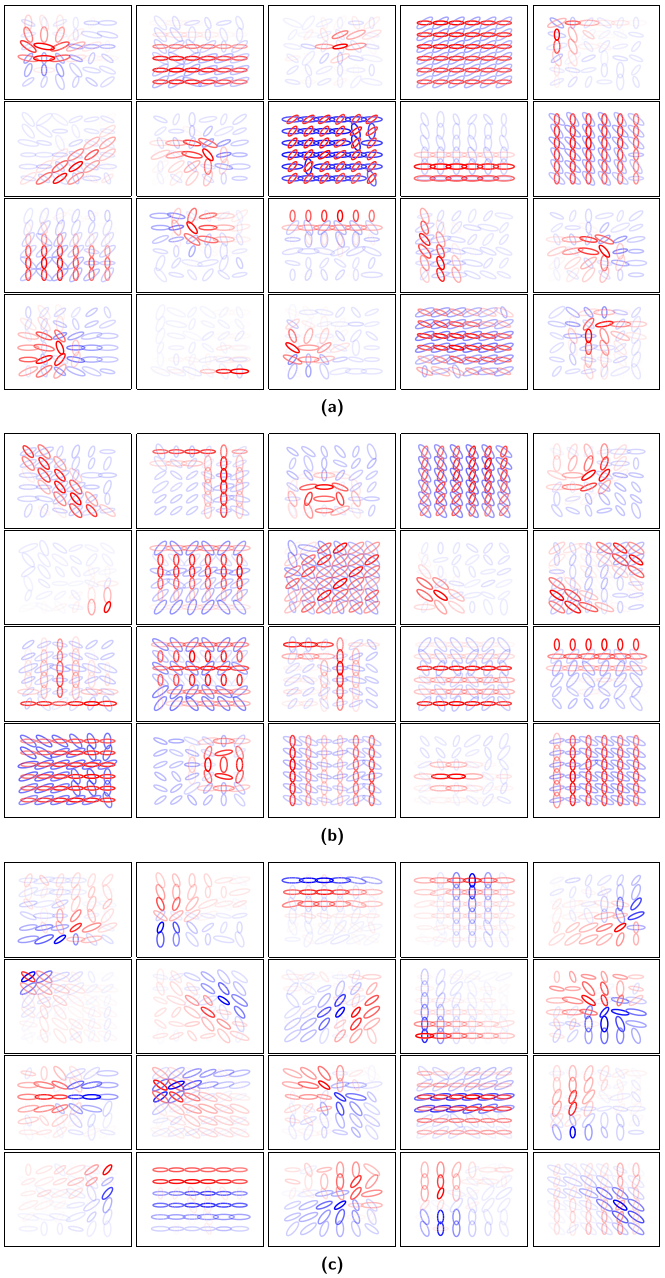

Figure 6. Visualization of V2 model units with 6×6 spatial locations and 100 principal components. (a) Sparse coding with a regularization coefficient of 0.5. (b) Sparse coding with a regularization coefficient of 4.0. (c) ICA. Values outside the central 6×6 region do not have a response. Opacity reflects the response intensity, color reflects the sign of the response (red for positive and blue for negative), and size reflects frequency.

Results

Unit properties

The visualizations for non-negative sparse coding with regularization coefficients of 0.5 and 4.0 and overcomplete ICA when run with 6×6 spatial locations are shown in Figures 6a, 6b, and 6c respectively. The two models discovered qualitatively different units. The non-negative sparse coding units contain corners, curves, circles, lines, parallel lines, and other structures. The overcomplete ICA units contain iso-oriented excitation with broad, side, cross, and end inhibition units and orientation-convergent excitation with end inhibition units as defined by Hosoya and Hyvärinen (2015). Each non-negative sparse coding unit also recruited more excitatory values than inhibitory values while overcomplete ICA had more balance between excitation and inhibition. However, it is important to remind the reader that, unlike overcomplete ICA, in non-negative sparse coding there is no explicit form of forward computation, so positive and negative values do not bear the same meaning. An excitatory value in

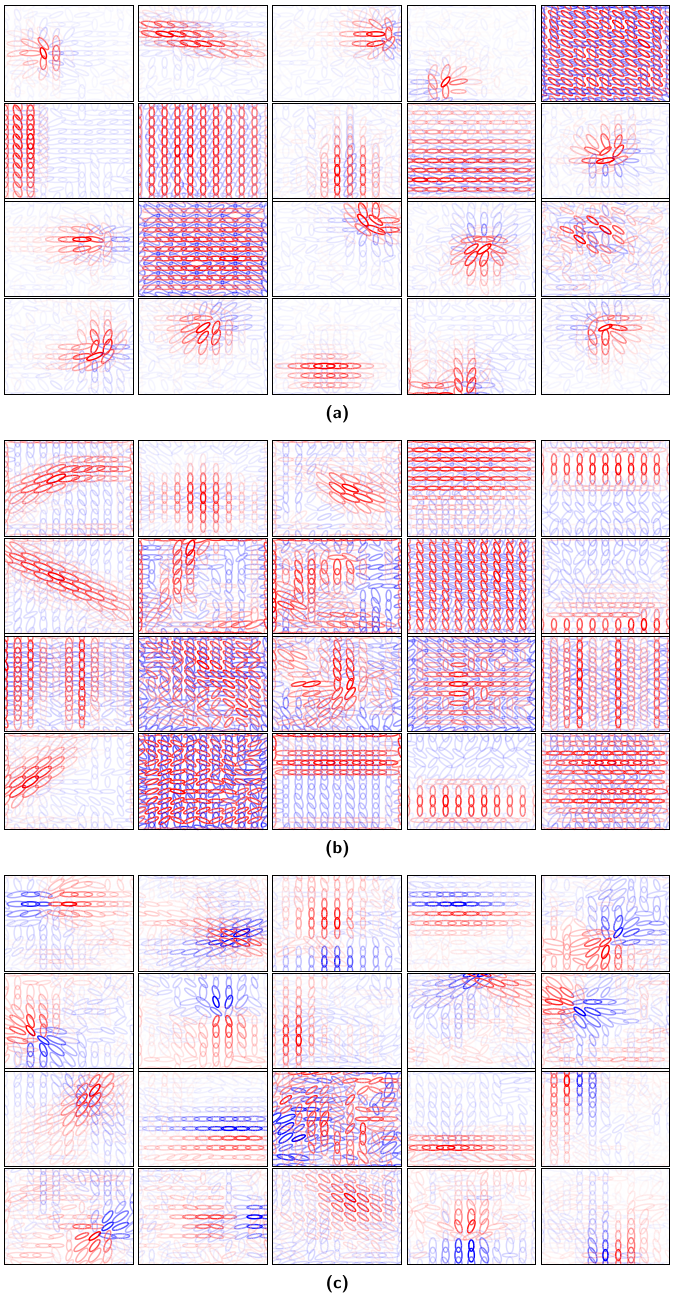

Figure 7. Visualization of sparse coding model V2 units with 11×11 spatial locations and 350 principal components. (a) Sparse coding with a regularization coefficient of 2.0. (b) Sparse coding with a regularization coefficient of 4.0. (c) ICA. Opacity reflects the response intensity, color reflects the sign of the response (red for positive and blue for negative), and size reflects frequency.

an overcomplete ICA unit simply means that stimulus was present with the orientation and frequency depicted by the Gabor plot and an inhibitory value the opposite, but the same stimulus can be described by negating the sign of the unit and unit response. By contrast, an excitatory value in a non-negative sparse coding unit means that a stimulus with the given orientation

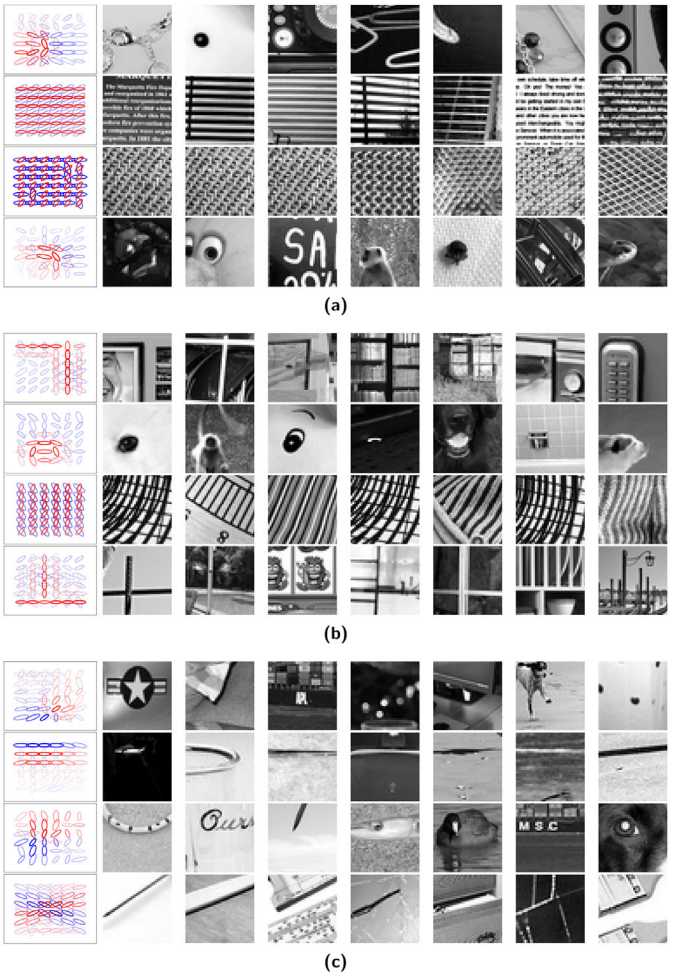

Figure 8. Maximum response patches. (a) Sparse coding with a regularization coefficient of 0.5. (b) Sparse coding with a regularization coefficient of 4.0. (c) ICA. The patches that maximally activated each V2 unit are shown to the right of its visualization. The response strength decreases from left to right.

and frequency was useful for reconstructing the input, but the sign of the unit cannot be flipped because the model is non-negative. The results for 11×11 spatial locations with regularization coefficients of 2.0 and 4.0 are shown in Figures 7a and 7b. The units discovered by non-negative sparse coding and overcomplete ICA (Figure 7c) with 11×11 spatial locations were similar to that for 6×6 spatial locations.

The patches that maximally excited selected units for non-negative sparse coding and overcomplete ICA with 6×6 spatial locations are shown in Figure 8. The patches for non-negative sparse coding (for both values of the regularization coefficient) reveal texture-like selectively in certain units (the second and third in Figure 8a and the third in Figure 8b) that are not easily described by common geometric primitives. The second unit in Figure 8a could be described as horizontal lines with gaps in between, although it was also activated by

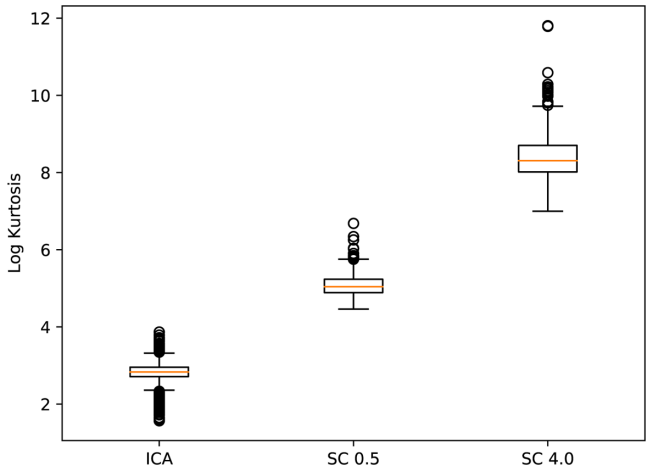

Figure 9. Box and whisker plot of kurtosis for all models. Plots are generated from the kurtosis over all 400,000 ImageNet patches for each unit. Circles represent outliers. Non-negative sparse coding with a regularization coefficient of 4.0 (SC 4.0) had the highest overall kurtosis, then non-negative sparse coding with a regularization coefficient of 0.5 (SC 0.5), then ICA. Each model finds a different sparse representation.

images of text. The third unit in Figure 8a appears as repeating small circles, and the third unit in Figure 8b appears as repeating curved lines. Corners appeared as lines connected at 90° angles, sometimes with other geometries nearby. One corner unit continued in both directions and was activated by crosses. Curves appeared mostly in circles. For the overcomplete ICA units, iso-oriented excitation units with side- and cross-inhibition appeared as lines, and iso-oriented excitation units with end inhibition appeared as lines stopping at a point. Orientation-convergent units with end inhibition appeared as blobs stopping at a point. Iso-oriented excitation with broad inhibition units varied, but often appeared as lines.

The kurtosis values were much larger for non-negative sparse coding with a regularization coefficient of 4.0 than for a coefficient of 0.5 or overcomplete ICA (see Figure 9). Exemplary units for each model are shown in Figures 10a, 10b, and 10c. Compared with other units, texture units had a large contribution to individual images (high kurtosis) in non-negative sparse coding, whereas overcomplete ICA relied often on iso-oriented excitation with broad inhibition units when assigning the largest coefficients. The distributions of the responses for each model (see Figure 10d) were similar to a mixture between an exponential distribution and a delta at zero, reflecting the rectification operation. The mixing proportion for the delta component is higher for the models with larger average kurtosis. Non-negative sparse coding distributions were more sparse and had higher kurtosis, with a regularization coefficient of 4.0 being the most sparse.

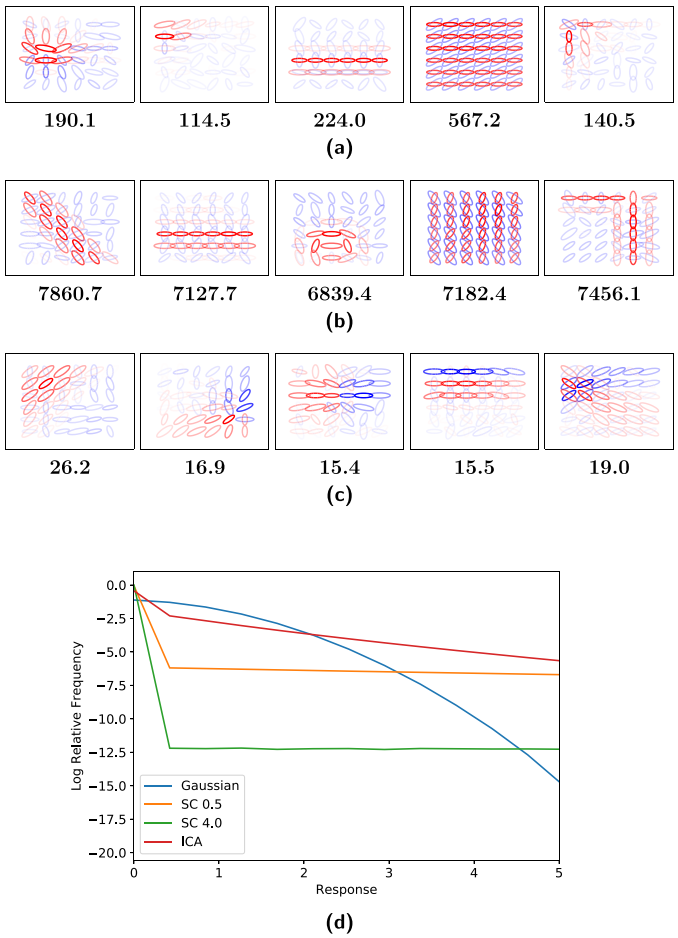

Figure 10. Model response properties. (a–c) Kurtosis of exemplary units of (a) sparse coding with a regularization coefficient of 0.5, (b) sparse coding with a regularization coefficient of 4.0, and (c) ICA. High kurtosis indicates more involvement of a unit in reconstructing particular images. (d) Histogram in the log domain of the responses to all 400,000 image patches for each of the three models.

Image classification

A common metric of vision models is performance on image classification tasks. A few classification tasks were explored here which test the ability to distinguish between figure and ground, multiple texture classes, and the angles between line segments connected at one point. The results for these experiments are shown in Figure 11. Overcomplete ICA performed the best. Non-negative sparse coding with a regularization coefficient of 0.5 was competitive with overcomplete ICA on the figure–ground and texture classification tasks, but non-negative sparse coding with a regularization coefficient of 4.0 was only competitive on the figure–ground task. Non-negative sparse coding with a regularization coefficient of 4.0 performed the worst on all tasks. For the figure–ground, texture, and

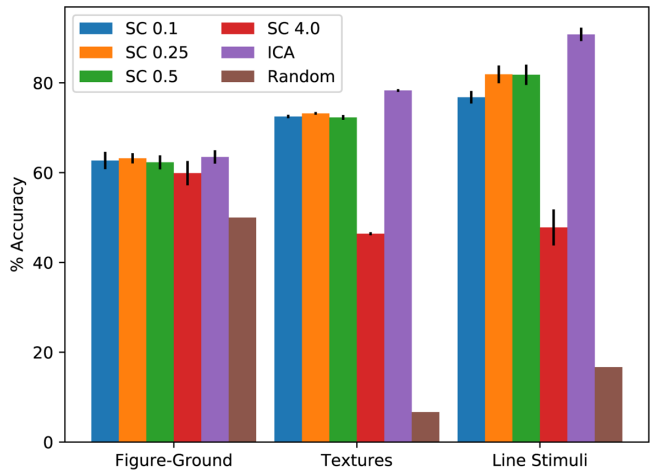

Figure 11. Classification accuracies. Average accuracy over 5-fold cross-validation for non-negative sparse coding with regularization coefficients of 0.1, 0.25, 0.5 and 4.0 (listed as SC followed by the regularization coefficient) and overcomplete ICA (listed as ICA). Error bars reflect standard deviation over the 5 folds. Random denotes the result of guessing (expectation computed for the number of labels).

	Figure-Ground	Texture	Line Stimuli
SC 0.1	5642.0	981.5	226.5
SC 0.25	5594.0	969.9	224.8
SC 0.5	5669.5	976.0	221.2
SC 4.0	6039.5	1232.3	222.3
ICA	5520.5	775.1	193.5

Figure 12. Average number of support vectors. Average number of support vectors for each experiment over all classes (2 for figure–ground, 15 for texture, and 6 for line stimuli).

line stimuli tasks, non-negative sparse coding with a regularization coefficient of 0.5 had percent accuracies of 62.3%, 72.3%, 81.8%, respectively; non-negative sparse coding with a regularization coefficient of 4.0 had percent accuracies of 59.9%, 46.4%, and 47.8%, respectively; and overcomplete ICA had percent accuracies of 63.5%, 78.3%, and 90.8%, respectively (all shown to one decimal place). We also report the average number of support vectors used for each manipulation of each experiment in Figure 12.

Texture sensitivity

A comparison with human vision can be made by analyzing the responses of these models to textures of varying classes, such as the textures of the second classification experiment and their spectrally matched noise versions that preserve the amplitude spectrums of the original textures but have randomized phase. Secondary visual cortex shows sensitivity to texture

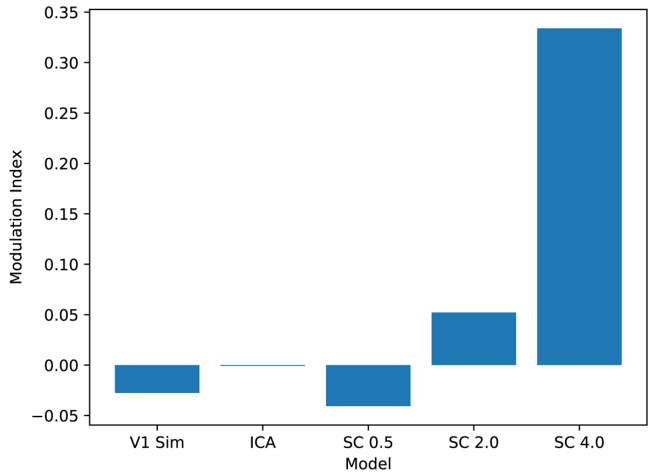

Figure 13. Texture modulation indices for vision models. The models include the initial simulated V1 stage via Gabor filters (V1 Sim), overcomplete ICA (ICA), and non-negative sparse coding with a regularization coefficient of 0.5 (SC 0.5), 2.0 (SC 2.0), and 4.0 (SC 4.0).

that is absent in V1 (Freeman et al., 2013; Kohler et al., 2016; Ziemba et al., 2016; Laskar et al., 2020). For instance, in an fMRI experiment, the modulation index (see Methods) for textures versus noise was much larger in V2 than in V1 with an average modulation index of about 0.13 across subjects for V2 (Freeman et al., 2013). We used the same texture and noise stimuli as in the texture classification experiment. For the models studied here, a similar difference in modulation index (between the V1 stage and V2 stage) would suggest that the trend of texture sensitivity in the primary and secondary visual cortex is also present in these models. The texture modulation indices for all models were computed by taking the responses to the 30,000 texture patches along with the responses to 30,000 spectrally matched noise versions of the texture patches, taking the difference between each, and normalizing via the sum of each (see Methods) to yield 30,000 modulation indices. Modulation indices for texture–noise pairs that both yielded 0 response were discarded because they did not provide any response information. The modulation indices for overcomplete ICA and non-negative sparse coding with regularization coefficients of 0.5, 2.0, and 4.0 are shown in Figure 13.

The large kurtosis of non-negative sparse coding resulted in many more texture-noise pairs with no response (0 response to the texture and noise image) as the regularization coefficient increased. The percentage of texture–noise pairs with a response for overcomplete ICA was 73.9% and for non-negative sparse coding with regularization coefficients of 0.1, 0.25, 0.5, 1.0, 1.5, 2.0. 2.5, 3.0, 3.5, and 4.0 were 21.7%, 18.9%, 15.2%, 9.73%, 6.16%, 3.84%, 2.38%, 1.46%, 0.925%, and 0.594% respectively. However, overcomplete ICA

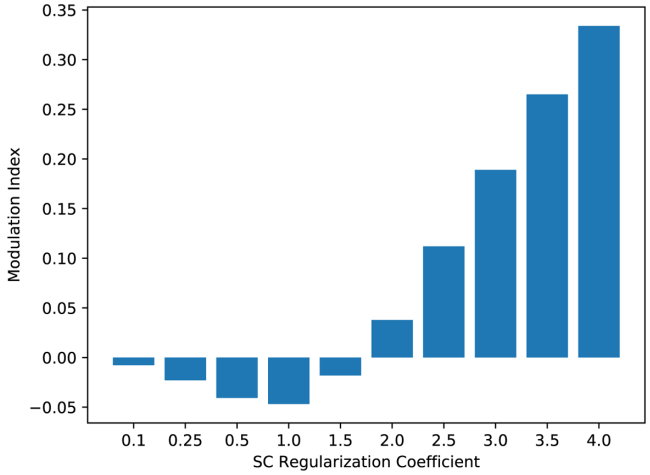

Figure 14. Texture modulation indices for non-negative sparse coding with various values of the regularization coefficient. As the regularization coefficient increased, the texture modulation increased. A regularization coefficient of 2.5 yielded a modulation index of approximately 0.112, which roughly approximates the texture modulation index of V2 as measured by fMRI (Freeman et al., 2013).

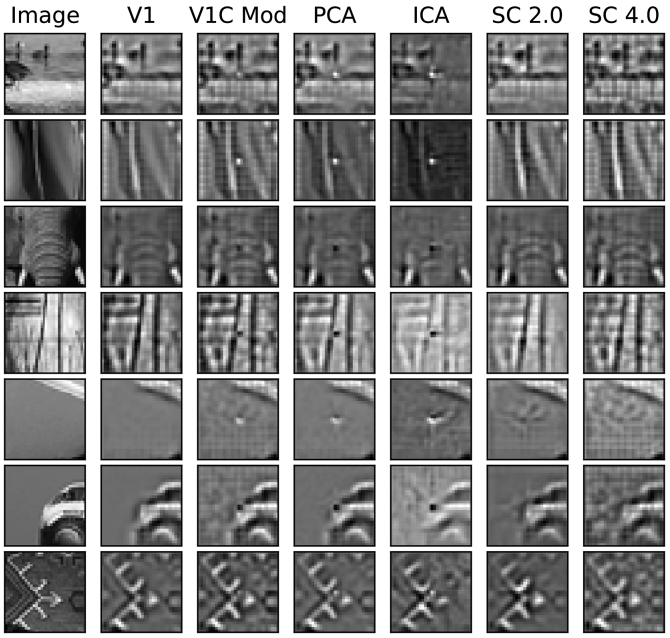

Figure 15. Image reconstructions of 32×32 image patches with a 1×1 missing V1 complex region. The first column shows the original image, the second the V1 reconstruction of the image, the third the V1 complex reconstruction with 1×1 missing region, the fourth the PCA reconstruction, the fifth the overcomplete ICA reconstruction, the sixth the non-negative sparse coding reconstruction with a regularization coefficient of 2.0, and the last the non-negative sparse coding with a regularization coefficient of 4.0.

does not have a sparsity control, so it could not yield representations with higher kurtosis. Interestingly, as the regularization coefficient of non-negative sparse coding increased, the modulation index increased. Over the range of values tested, a regularization coefficient of 2.5 most closely matched the modulation index of V2. The modulation indices for a few of the values tested are shown in Figure 14. The increase of the modulation index with sparsity is consistent with previous findings that considered deep neural networks (Zhuang et al., 2017).

Patch completion

A less-studied, but important, metric for vision models is their ability to infer missing structure in images. In this experiment, 1×1 and 2×2 regions were deleted at the level of the V1 complex cell responses. Selected patch reconstructions are shown in Figures 15 and 16. The columns in Figures 15 and 16 correspond with the different stages of visual processing in the model starting with the original image. The next two columns show the image reconstruction by the model after Gabor filtering (V1) and energy pooling along with the information removal (V1C Mod). Because the inverse transform of the V1 complex responses was undone exactly by saving the angles between quadrature pair responses during the forward transform, the V1 complex stage did not change the appearance of the reconstruction. For this reason, the V1 complex reconstruction without the information removal was omitted (see the V1 representation instead). The next column shows the reconstruction after redundancy reduction with PCA. The final three columns show the model's final stage of processing with either overcomplete ICA, non-negative sparse coding with a regularization coefficient of 2.0 (SC 2.0), and non-negative sparse coding with a regularization coefficient of 4.0 (SC 4.0). The MSE for the patches with a 1×1 V1 complex region deleted for non-negative sparse coding with a regularization coefficient of 2.0 was 0.0129, with a regularization coefficient of 4.0 was 0.0218, and for overcomplete ICA was 0.786. The MSE for the patches with a 2×2 V1 complex region deleted for non-negative sparse coding with a regularization coefficient of 2.0 was 0.0355, with a regularization coefficient of 4.0 was 0.0304, and for overcomplete ICA was 2.67. Both non-negative sparse coding manipulations performed much better than overcomplete ICA. The Student *t*-test for independent samples showed that both were significant ($p < 0.01$). The differences between non-negative sparse coding with both values of the regularization coefficient were also significant ($p < 0.01$; *t*-test for independent samples) with non-negative sparse coding and a regularization coefficient of 2.0 performing better

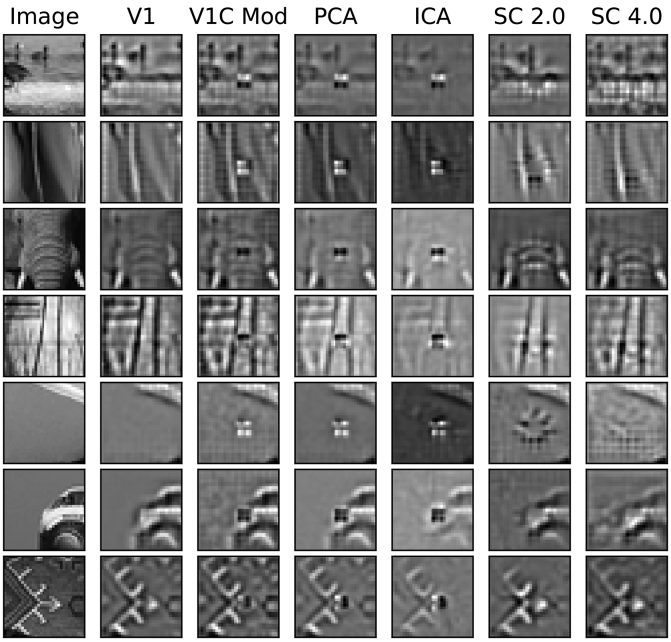

Figure 16. Image reconstructions of 32×32 image patches with 2×2 missing V1 complex region. The first column shows the original image, the second the V1 reconstruction of the image, the third the V1 complex reconstruction with a 2×2 missing region, the fourth the PCA reconstruction, the fifth the overcomplete ICA reconstruction, the sixth the non-negative sparse coding reconstruction with a regularization coefficient of 2.0, and the last the non-negative sparse coding with a regularization coefficient of 4.0.

when only a 1×1 V1 complex region was deleted, and with a regularization coefficient of 4.0 performing better when a 2×2 V1 complex region was deleted. Overcomplete ICA reconstructions did not seem to attempt to complete missing information. Interestingly, non-negative sparse coding with different regularization coefficients inferred missing information with different plausible image reconstructions, such as how it inferred the car's bumper and the arrow in the last two rows in Figure 15. Also, higher sparsity allowed for better inference when more information was missing in the line structure in the second row and the low spatial frequency region in the fifth row in Figure 16.

Discussion

This work built on the hierarchical unsupervised learning V2 model of Hosoya and Hyvärinen (2015). Their work investigated overcomplete independent component analysis as a sparse coding model, but they did not investigate the original sparse coding model of Olshausen and Field (1996). In this article, the different structures learned by incorporating ICA versus sparse coding in the V2 model were shown, a characterization of the marginal statistics of the filter responses was performed, and the implications for performing inference and classification tasks with these approaches were demonstrated.

This article examined the tradeoffs of ovecomplete ICA versus non-negative sparse coding, and non-negative sparse coding with different sparsity levels, for vision. Good performance on a single task like image classification does not imply good performance on other tasks like image inference. Furthermore, image classification accuracy may suffer with a certain degree of kurtosis, whereas texture sensitivity becomes more like V2. Although seemingly worse for the model, such a change may be tolerated in the light of V2 being an early visual processing area. Perhaps further transformations are needed before classification accuracy increases to the degree observed by overcomplete ICA. However, it is not obvious that the strategy of sparse coding by Olshausen and Field (1996) should be ascribed to V2, and other sparse coding algorithms may yield better classification accuracy with a representation that is more sparse. Although the non-negative sparse coding model here was linked to V2 via the modulation index, other coding strategies may also yield similar values of the modulation index.

In terms of our choice of sparse coding implementation, a theoretical link exists between the sparse coding method of Olshausen and Field (1996) and the neurally plausible locally competitive algorithm (Rozell et al., 2008) approach to sparse coding based on the principles of thresholding and local competition. The locally competitive algorithm is a dynamical systems approach to sparse coding that models a neural circuit via membrane potential-like quantities that govern the sparse coding response properties along with inhibitory signals from other sparse coding units (similar to neural inhibition). Interestingly, Rozell et al. (2008) showed that their approach to deriving the sparse coding responses minimizes, under some constraints, the same loss function used here (LASSO) to perform sparse coding. Thus, results from the algorithm explored in this work can be connected to neurally plausible implementations. However, the LCA strategy mainly accounts for the forward response properties; the basis functions are still derived in a similar fashion to Olshausen and Field (1996). The degree of sparseness from the LASSO objective influences the learned basis functions, but it is still possible that other methods of deriving the basis functions may yield better classification results with a higher degree of sparseness, which is a limitation to this work.

Future work can relate to biological data, by examining how well the model responses (with different sparsity levels) match measures of neural activity. Hosoya and Hyvärinen (2015) examined V2 properties

in their hierarchical overcomplete ICA model, and one can consider recent natural scenes data such as the large-scale natural scenes dataset of Allen et al. (2021). Their dataset provides V2 fMRI voxel data in response to viewing natural scenes. The non-negative sparse coding responses can be computed for the natural scenes to attempt fitting a linear classifier to the data with the model's responses.

There is also interest in examining connections to deep convolutional neural networks, which have been shown to capture various cortical neural response properties (Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Pospisil et al., 2018; Cadena et al., 2019; Kindel et al., 2019; Laskar et al., 2020). Such networks can perform a form of sparse coding by thresholding (setting to zero) responses with the ReLU activation function depending on the values of the bias weights (characterized by Bowren, 2021). A deep neural network can be trained with a constraint of several degrees of large negative bias weights to vary sparseness. If classification accuracy suffers and inference improves with larger kurtosis, then the result holds in another type of sparse coding model that has been popular for modeling cortical data. One could further test how the deep neural network models (modified for the different sparsity levels) capture neural data.

Non-negative sparse coding was a natural extension to make the original sparse coding model comparable with positive rectified overcomplete ICA, so it was explored here. Interestingly, the corresponding generative models of the two methods learned to represent images with different structures. Moreover, sparse coding found different structures depending on the degree of sparsity in the model, which is fixed in overcomplete ICA. Regular positive and negative sparse coding was also found to yield different structures, but non-negative sparse coding was explored here because it performed better on the classification tasks and eliminated the need for rectification of model V2 units. In comparison with other vision models, the V1 complex cell stage is fixed rather than learned as in other image statistics models of visual cortical complex cells (Hyvärinen & Hoyer, 2001; Karklin & Lewicki, 2009).

The differences between overcomplete ICA and non-negative sparse coding may be attributed to the different computational strategies of the two models: overcomplete ICA has an explicit linear forward transform, whereas non-negative sparse coding has an implicit nonlinear forward transform. The strategy of ICA is to find a set of filters whose responses to images are as independent as possible, assuming a linear transform (note that independence is not guaranteed, because of the existence of higher-order couplings that cannot be removed by linear transformations, and by overcompleteness). Non-negative sparse coding, by contrast, does not learn filters, but a dictionary of basis functions that optimally reconstructs images with a linear combination of a few (depending on the regularization coefficient) of its basis functions. However, the different objectives of overcomplete ICA and non-negative sparse coding were not designed to classify images or infer unseen image information respectively. Roughly independent filter responses are not obviously better than sparse coding responses for image classification, and the advantage was not large compared to the best non-negative sparse coding configuration. For the image inference task, image reconstruction error grew as the regularization coefficient increased, but the reconstruction error for the original unmodified image decreased despite having the opposite effect on the input reconstruction error. In other words, the reconstruction error for the image without the deletion in the V1 complex stage was better for a regularization coefficient of 0.5 than 4.0, but when the deletion was present the opposite was true. The better reconstruction of the original image can be seen qualitatively in Figures 15 and 16 where 1×1 or 2×2 input regions were deleted. For example, consider the back-bumper of the car in row 6 of Figure 15; if the model were simply performing reconstruction, the missing bumper region (blank space) of the car would have been reconstructed, but the model introduces new information into the image representation via its basis functions. The difference in performance on the image tasks was not an obvious result of the difference in loss functions, and previous patch completion (in-painting) results like that of Mairal et al. (2009) only investigated inference in a single-layered model with smaller receptive fields and one level of sparsity. Here, the result of inference could be seen in single patches rather than reconstructing an entire image from its constituent image patches, and, more important, the result across various levels of sparsity was also demonstrated. Also, another difference between sparse coding and ICA is that while ICA may be thought of as similar to sparse coding in the complete case, ICA tends to maximize coherence (redundancy) in its filter matrix when extended to the overcomplete case (Livezey et al., 2019). In maximum-likelihood inspired ICA models, this factor is usually addressed by adding a coherence control to the loss function. Score matching ICA was incorporated in this model, similar to Hosoya and Hyvärinen (2015), and although coherence control was not explicitly enforced, score matching provides an implicit, albeit data dependent, form of coherence control.

It was found that the resulting non-negative sparse coding units contained intuitively useful geometric primitives such as curves and corners (Figures 6a and 6b), unlike overcomplete ICA, which found the units defined by Hosoya and Hyvärinen (2015) shown in Figure 6c mentioned in the Methods. Hosoya and Hyvärinen (2015) obtained orientation-convergent

units, which they noted might be related to corner detection. Other hierarchical models have also resulted in structures such as curves and corners, including a two-layer sparse deep belief net model (Lee et al., 2007), a two-layer model that included a statistically optimal divisive normalization at the V1-like stage (Coen-Cagli & Schwartz, 2013), and the second layer of particular deep convolutional neural networks (Zeiler & Fergus, 2014). There is a need in future work to study differences in the resulting structure learned across different classes of models and computations, including intermediate layers of deep convolutional neural networks. In this work, the focus was on a fixed architecture and the influence of ICA versus sparse coding.

For non-negative sparse coding, as the regularization coefficient increases, the sizes of the geometric primitives increase because the model is constrained to represent entire images with only a few basis functions, so each basis function must contain more information to reduce reconstruction error. To see this more clearly, consider the extremes of the regularization coefficient: a coefficient of zero removes the L1 penalty term while increasing the coefficient excessively leads to fewer and fewer basis functions reconstructing the image until it is reconstructed as an image of all zeros with no basis functions. Reconstruction error increases with higher values of the regularization coefficient because the input representation is modified: more information has to be inferred. A similar effect was seen with sparse autoencoders trained on handwritten digits when the degree of sparsity was increased (Makhzani & Frey, 2015). This inductive inference mechanism has potential use when sending information down a noisy pipe. When parts of the signal become corrupted, these parts can be safely interpolated by the contributions of the basis functions. When the regularization coefficient is low, these parts of the signal are more likely to be interpreted by the model as genuine parts of the signal. When the regularization coefficient is high, these parts of the signal are inferred over because the model does not have coefficients to spare on perturbations which are not well represented by the dictionary if the model is trained on uncorrupted signals. This inductive inference mechanism may be useful in the brain as well because of the stochasticity in the firing of neurons.

In addition to geometric primitives, non-negative sparse coding (with both choices of regularization coefficient shown in Figures 6 and 7) also finds units that maximally respond to texture-like repeating patterns. Hosoya and Hyvärinen (2015) found some indication of more localized texture patterns with overcomplete ICA, but it was found here that the non-negative sparse coding units by contrast produced texture-like units covering the full extent of the receptive field. Texture patterns were also apparent in intermediate layers of deep convolutional neural networks (Zeiler & Fergus, 2014). In practice, some of the unit types of non-negative sparse coding and overcomplete ICA were both maximally excited by similar images (Figure 8). In non-negative sparse coding, lines that stop abruptly were maximally excited by images similar to those that maximally excite iso-oriented excitation with end inhibition units in overcomplete ICA: images where a line stops before reaching the end of the image. However, most of the non-negative sparse coding unit structures were different from those of overcomplete ICA. Non-negative sparse coding was also much more sparse than overcomplete ICA (Figures 9 and 10). The distribution of non-negative sparse coding units covers a different range of sparseness as measured by kurtosis (Figure 9) compared with overcomplete ICA. If non-negative sparse coding (with a certain regularization coefficient) is advantageous for vision tasks, this sparseness could be motivated by an efficient coding paradigm.

Non-negative sparse coding with the appropriate regularization coefficient better matched the level of texture sensitivity in V2 as measured with the texture modulation index. The modulation indices were computed for the Brodatz texture dataset (Brodatz, 1966) in order to determine if the texture sensitivity level would increase with an increasing regularization coefficient, but it is important to note that other texture datasets, like the one in Freeman et al. (2013), may yield different optimal coefficients, so the exact values should not be viewed as constants for the optimal V2 texture sensitivity match. Instead, the biological link is the increase in texture sensitivity with the increase in sparseness in the non-negative sparse coding model up to a point. The ability of sparse coding to derive bases with varying levels of sparsity allowed it to derive bases with varying levels of texture sensitivity as noted by Zhuang et al. (2021) with other sparse models. It would also be interesting to see if the change in kurtosis of the spectrally matched noise images owing to the elimination of higher-order statistics would elicit a model response pattern similar to that found in an fMRI study by Puckett et al. (2020) where humans viewed natural scenes degraded of higher order statistics.

Interestingly, although non-negative sparse coding had a high level of texture sensitivity (modulation index of 0.334), its performance on the classification tasks were poorer than that of non-negative sparse coding with a regularization coefficient of 0.5. The implication is that a model that spans the range of kurtosis of a low and high kurtosis sparse coding may better match V2. More on this approach is discussed elsewhere in this article; however, the main takeaway is that, within a sparse coding framework, texture sensitivity may be increased at the expense of classification accuracy (especially angle classification). This was likely due to

the larger receptive fields of sparse coding with a higher regularization coefficient.

Non-negative sparse coding performed worse overall on image classification than overcomplete ICA followed by point-wise rectification (see Figure 11). However, vision is a rich process that pertains to far more than distinguishing between classes of images. Vision systems must learn to perform inference when information is absent or lost due to error. Non-negative sparse coding seems to address this, but not rectified overcomplete ICA (see Figures 15 and 16). Perturbations to the V1 complex representation were simply maintained by rectified overcomplete ICA, whereas non-negative sparse coding derives a representation much closer to the original V1 representation before perturbations were introduced. It is likely that non-negative sparse coding with a regularization coefficient of 4.0 suffered the most on the line classification task because the receptive fields of most units were too large to pick up on the angle between the two lines. However, the rectification step in overcomplete ICA seemed to remove structural information in the ICA representation, but this step was found to be necessary to maintain its high image classification accuracy; image classification accuracy decreased significantly and performed worse than non-negative sparse coding when rectification was not incorporated. Non-negative sparse coding has an implicit built-in rectification mechanism, so it did not experience a similar effect; rather, the learned representations were derived such that no negative responses were needed. Interestingly, non-negative sparse coding reconstructed images and completed missing regions, so although it suffered at image classification, it excelled at image inference and overcomplete ICA experienced the opposite. An interesting question, but beyond the scope of this work, is exploring the existence of any asymmetries in the activations of non-negative sparse coding, especially given the fact that changing the sign of an activation at the V2 level does not merely correspond with a contrast inversion.

Also, different reconstructions were formed with different values of the regularization coefficient. Larger values of the regularization coefficient led to representations that were more sparse and had more latent information introduced by the model. Smaller values of the regularization coefficient led to representations that were more faithful to the original V1 representation. Indeed, when only a 1×1 V1 complex region was deleted, a regularization coefficient of 2.0 yielded a smaller MSE; however, when a 2×2 V1 complex region was deleted a regularization coefficient of 4.0 yielded a smaller MSE. This finding is consistent with the idea that representations that are more sparse introduce more prior knowledge into the representation through the model's basis functions. When the model was constrained to make due with fewer basis functions, more information had to be inferred. Depending on the amount of missing information, different degrees of sparsity were more useful in building the model representation that best explained the data.

One question that arises when attempting to add inference and content generation mechanisms to vision models is where in the brain do such mechanisms exist? Inference within the receptive field may occur throughout the visual cortex and is harder to localize, but complete image generation (imagining images) can be studied with fMRI. D'Esposito et al. (1997) asked subjects to imagine images while in a fMRI machine and found that the visual association cortex was activated, but not the primary visual cortex. It seems that content generation arises in higher visual areas and should not be expected from low-level vision models like the one described in this work. Instead, only low-level local region inference might be expected in the early visual system. The degree of sparsity expected might be related to the vast overcompleteness in the primary visual cortex (Olshausen & Lewicki, 2014). See Olshausen et al. (2009) for an application of very overcomplete and sparse coding.

One future direction may be to attempt to incorporate different degrees of sparsity into one overall model that is both able to reconstruct images with low error and perform inferences that best explain the data. Such a model would learn a representation that spans the range of kurtosis distributions in Figure 9. Another direction is applying the perturbations of Figures 15 and 16 to the original image. This way, missing information in lower level receptive fields may be inferred via higher level sparse coding, and future higher level models may be shown to complete larger regions of images. The original image was not modified in this work because of limitations in the underlying V1 complex energy model. The energy model pooled V1 responses by taking the magnitude of each quadrature pair of Gabor filters in polar coordinates, but discarding the phase. Because the model contains a large spatial stride, when reconstructing images with a randomized phase a large part of the image structure was lost and inference was unfeasible without attempting a method of phase recovery (Gerchberg & Saxton, 1972).

In the future, more Gabor filters could be applied at more spatial locations, orientations, and phases to better recover the original image structure. An appealing approach would be to learn a representation with sparse coding for the V1 representation coupled with some form of pooling. Besides pooling after V1, the sparse coding derived V2 responses can be enhanced. A natural improvement is to make the model convolutional (Szlam et al., 2010) to decrease the redundancy in the learned basis functions. One could also include a more sophisticated approach to sparse coding, such as approximating variance structure with a nonlinear

model as in Karklin and Lewicki (2005). This practice gives the model an understanding of the underlying distribution of the sparse coding responses. Yet another appealing sparse coding approach is to learn both of the bases at the same time as done by Boutin et al. (2021), Zeiler and Fergus (2010). The role of nonlinear computations such as divisive normalization motivated by image statistics (e.g., Coen-Cagli & Schwartz, 2013) can also be explored within such models. All these methods allow for perturbations to be introduced in the original image.

Conclusion

Non-negative sparse coding discovers a unique set of intuitively useful basis functions that with different degrees of sparsity may be advantageous for particular vision tasks. Overcomplete ICA performs better than non-negative sparse coding on image classification, but performs poorly on image inference. With a high degree of sparsity in a high-level visual model, non-negative sparse coding is able to infer small regions of missing information. The inference mechanism postulated here is feasible.

Keywords: hierarchy, sparse coding, mid-level vision

Acknowledgments

The authors thank Hauro Hosoya for publicly providing his hierarchical overcomplete ICA model code (Hosoya & Hyvärinen, 2015). We thank Bruno Olshausen for his helpful discussion regarding sparse coding in hierarchical models. We thank Nasir Laskar for the synthetic texture images in the classification experiment generated with the publicly available texture generation code associated with Portilla and Simoncelli (2000). We thank Ruben Coen-Cagli for providing us with his code for generating the figure–ground patches from the Berkeley segmentation dataset.

Supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1451511. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Commercial relationships: none.
Corresponding author: Odelia Schwartz.
Email: odelia@cs.miami.edu.
Address: Department of Computer Science, University of Miami, Coral Gables, FL, USA.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., Pestilli, F., ... Naselaris, T. et al. (2021). A massive 7t FMRI dataset to bridge cognitive and computational neuroscience. *bioRxiv*.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Berkes, P., White, B., & Fiser, J. (2009). No evidence for active sparsification in the visual cortex. *Advances in Neural Information Processing Systems, 22*, 108–116.

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision, 5*(6), 9–9.

Boutin, V., Franciosini, A., Chavane, F., Ruffier, F., & Perrinet, L. (2021). Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLoS Computational Biology, 17*(1), e1008629.

Bowren, J. (2021). A sparse coding interpretation of neural networks and theoretical implications. arXiv preprint arXiv:2108.06622.

Brodatz, P. (1966). *Textures: A photographic album for artists and designers*. New York: Dover Pub.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., ... Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology, 15*(4), e1006897.

Coen-Cagli, R., & Schwartz, O. (2013). The impact on midlevel vision of statistically optimal divisive normalization in V1. *Journal of Vision, 13*(8), 13–13.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. BioRxiv.

D'Esposito, M., Detre, J. A., Aguirre, G. K., Stallcup, M., Alsop, D. C., Tippet, L. J., ... Farah, M. J. (1997). A functional mri study of mental image generation. *Neuropsychologia, 35*(5), 725–730.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6*(4), 559–601.

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure–ground cues are valid for natural images. *Journal of Vision, 7*(8), 2–2.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Galerne, B., Gousseau, Y., & Morel, J.-M. (2010). Random phase textures: Theory and synthesis. *IEEE Transactions on Image Processing, 20*(1), 257–267.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167–192.

Gerchberg, R. W., & Saxton, W. O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik, 35*, 237–246.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Hosoya, H., & Hyvärinen, A. (2015). A hierarchical statistical model of natural images explains tuning properties in V2. *Journal of Neuroscience, 35*(29), 10412–10428.

Hoyer, P. O. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (pp. 557–565). IEEE, https://doi.org/10.1109/NNSP.2002.1030067.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research, 6*(24), 695–709.

Hyvärinen, A., & Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research, 41*(18), 2413–2423.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation, 9*(7), 1483–1492.

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience, 24*(13), 3313–3324.

Kanerva, P. (1988). *Sparse distributed memory.* MIT Press.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation, 17*(2), 397–423.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*(7225), 83–86.

Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision, 19*(4), 29–29.

Kohler, P. J., Clarke, A., Yakovleva, A., Liu, Y., & Norcia, A. M. (2016). Representation of maximally regular textures in human visual cortex. *Journal of Neuroscience, 36*(3), 714–729.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Laskar, M. N. U., Giraldo, L. G. S., & Schwartz, O. (2020). Deep neural networks capture texture sensitivity in V2. *Journal of Vision, 20*(7), 21–1.

Lee, H., Ekanadham, C., & Ng, A. (2007). Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems, 20*, 873–880.

Livezey, J. A., Bujan, A. F., & Sommer, F. T. (2019). Learning overcomplete, low coherence dictionaries with linear inference. *Journal of Machine Learning Research, 20*, 174–1.

Luo, Y., Xu, Y., & Ji, H. (2015). Removing rain from a single image via discriminative sparse coding. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3397–3405). IEEE.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Non-local sparse models for image restoration. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2272–2279). IEEE.

Makhzani, A., & Frey, B. J. (2015). Winner-take-all autoencoders. *Advances in Neural Information Processing Systems, 28*, 2791–2799.

Olshausen, B. A., Cadieu, C. F., & Warland, D. K. (2009). Learning real and complex overcomplete representations from the statistics of natural images. *Proceedings of the SPIE 7446 (Wavelets XIII)*. (pp. 236–246). International Society for Optics and Photonics.

Olshausen, B. A. et al. (2003). Principles of image representation in visual cortex. *Visual Neurosciences, 2*, 1603–1615.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14*(4), 481–487.

Olshausen, B. A., & Field, D. J. (2004). What is the other 85% of V1 doing. In L. van Hemmen, & T. Sejnowski (Eds.), *Problems in System Neuroscience* (pp. 182–211), https://doi.org/10.1093/acprof:oso/9780195148220.003.0010.

Olshausen, B. A., & Lewicki, M. S. (2014). What natural scenes statistics can tell us about cortical representation. In Chalupa, & Werner (Eds.), *The New Visual Neurosciences* (pp. 1247–1262). MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pei, S.-C., & Zeng, Y.-C. (2006). A novel image recovery algorithm for visible watermarked images. *IEEE Transactions on Information Forensics and Security, 1*(4), 543–550.

Peterhans, E., & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *Journal of Neuroscience, 9*(5), 1749–1763.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70.

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'artiphysiology'reveals v4-like shape tuning in a deep network trained for image classification. *Elife, 7*, e38242.

Puckett, A. M., Schira, M. M., Isherwood, Z. J., Victor, J. D., Roberts, J. A., & Breakspear, M. (2020). Manipulating the structure of natural scenes using wavelets to study the functional architecture of perceptual hierarchies in the brain. *NeuroImage, 221*, 117173.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint *arXiv*:1511.06434.

Rao, R. P., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain: Perception and neural function.* MIT Press, https://doi.org/10.7551/mitpress/5583.001.0001.

Rowekamp, R. J., & Sharpee, T. O. (2017). Cross-orientation suppression in visual area V2. *Nature Communications, 8*(1), 1–9.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation, 20*(10), 2526–2563.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Shan, H., & Cottrell, G. (2013). Efficient visual coding: From retina to V2. arXiv preprint *arXiv*:1312.6077.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193–1216.

Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex FMRI activity patterns for occluded scenes. *Journal of Vision, 21*(7), 5–5.

Szlam, A., Kavukcuoglu, K., & LeCun, Y. (2010). Convolutional matching pursuit and dictionary training. arXiv preprint *arXiv*:1010.0422.

Turner, M. H., Giraldo, L. G. S., Schwartz, O., & Rieke, F. (2019). Stimulus-and goaloriented frameworks for understanding natural vision. *Nature Neuroscience, 22*(1), 15–24.

von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience, 9*(5), 1731–1748.

Willmore, B. D., Mazer, J. A., & Gallant, J. L. (2011). Sparse coding in striate and extrastriate visual cortex. *Journal of Neurophysiology, 105*(6), 2907–2919.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature, 222*(5197), 960–962.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Yoshida, T., & Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications, 11*(1), 1–19.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zeiler, M., & Fergus, R. (2010). *Learning image decompositions with hierarchical sparse coding.* (TR2010-935), Courant Institute of Mathematical Science, New York University.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision* (pp. 818–833). Springer.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron, 47*(1), 143–153.

Zhaoping, L., & Jingling, L. (2008). Filling-in and suppression of visual perception from context: A bayesian account of perceptual biases by contextual inuences. *PLoS Computational Biology, 4*(2), e14.

Zhuang, C., Wang, Y., Yamins, D., & Hu, X. (2017). Deep learning predicts correlation between a functional signature of higher visual areas and sparse firing of neurons. *Frontiers in Computational Neuroscience, 11*, 100.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United State of America, 118*(3), https://doi.org/10.1073/pnas.2014196118.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences of the United State of America, 113*(22), E3140–E3149.