Neural Decoding with Hierarchical Generative Models

Marcel A. J. van Gerven
marcelge@cs.ru.nl
Radboud University Nijmegen, Institute for Computing and Information Sciences, 6525 AJ Nijmegen, the Netherlands, and Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Floris P. de Lange
florisdelange@gmail.com
Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Tom Heskes
t.heskes@science.ru.nl
Radboud University Nijmegen, Institute for Computing and Information Sciences, 6525 AJ Nijmegen, the Netherlands, and Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Recent research has shown that reconstruction of perceived images based on hemodynamic response as measured with functional magnetic resonance imaging (fMRI) is starting to become feasible. In this letter, we explore reconstruction based on a learned hierarchy of features by employing a hierarchical generative model that consists of conditional restricted Boltzmann machines. In an unsupervised phase, we learn a hierarchy of features from data, and in a supervised phase, we learn how brain activity predicts the states of those features. Reconstruction is achieved by sampling from the model, conditioned on brain activity. We show that by using the hierarchical generative model, we can obtain good-quality reconstructions of visual images of handwritten digits presented during an fMRI scanning session.

1 Introduction

Recent developments in cognitive neuroscience have shown that it is possible to infer mental state from neuroimaging data (Haxby et al., 2001; Thirion et al., 2006; Kay, Naselaris, Prenger, & Gallant, 2008; Mitchell et al., 2008; Miyawaki et al., 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009). These breakthroughs in neural decoding, popularized as brain reading, hold much promise as a new approach for studying human brain function (see Hassabis et al., 2009; Stokes, Thompson, Cusack, & Duncan, 2009, for a number of examples). Decoding can be distinguished into classification, identification, and reconstruction (Kay & Gallant, 2009). The aim of classification is to infer to which of a small number of stimulus classes a particular brain state belongs. The goal of identification is to determine which stimulus from a candidate set of possible stimuli explains the observed brain state. This is typically achieved by template matching, where the observed brain state is compared with the predicted brain state of stimuli in the candidate set. Reconstruction, finally, uses the observed brain state in order to reconstruct the actual stimulus rather than choosing from a class or set of potential stimuli.

An early decoding example is the study by Haxby et al. (2001), which showed that different stimulus categories can be classified reliably using fMRI. More recently, Kay et al. (2008) and Mitchell et al. (2008) showed that the identification of previously unseen stimuli by comparing predicted hemodynamic response with measured hemodynamic response is feasible. Thirion et al. (2006) showed that perceived or imagined contrast patterns can be reconstructed from retinotopic information. Finally, Miyawaki et al. (2008) and Naselaris et al. (2009) demonstrated that the reconstruction of perceived images from measured hemodynamic response is possible by decoding multivariate activation patterns. Note that reconstruction is much harder than either classification or identification since it requires inference about which stimulus (from a sheer infinite set of possible stimuli) caused the observed brain activity. In contrast, classification and identification boils down to the selection of one stimulus from a restricted set of possible stimuli.

In this study, our goal is to build a reconstruction model that aims to mimic neural encoding and to invert this model in order to reconstruct perceived stimuli from measured brain activity—in our case, hemodynamic responses in visual cortex. An influential hypothesis about neural encoding is the predictive coding hypothesis, which states that the brain tries to infer the causes of its sensations (Helmholtz, 1867; Barlow, 1961; Rao & Ballard, 1998). This principle, together with hierarchical organization as a key organizational principle of the brain (Zeki & Shipp, 1988; Felleman & Van Essen, 1991), suggests that the human brain may embody a hierarchical generative model where top-down drive along the hierarchy encodes our prior beliefs about the presence or absence of abstract causes and where bottom-up drive along the hierarchy encodes sensory information (Lee & Mumford, 2003; Friston, 2005).

Our reconstruction model is known as a deep belief network—a hierarchical generative model whose building blocks are known as restricted Boltzmann machines (Smolensky, 1986; Hinton, Osindero, & Teh, 2006) and whose latent causes (i.e., stimulus features) are learned from data. This approach is different from most existing reconstruction studies where reconstruction is based on predefined features (e.g., manually constructed image patches or fixed Gabor filters). This has as an advantage that the model can be adapted to data sets with different stimulus characteristics.

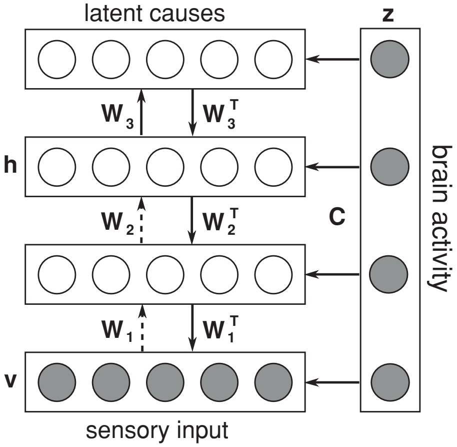

Figure 1: The hierarchical generative model implements how latent causes **h** explain sensory input **v** conditional on observed brain activity **z**. Reconstruction proceeds by sampling from the model while clamping **z**. Dashed arcs represent interactions that are used during training but are not part of the generative model.

The work presented in Fujiwara, Miyawaki, and Kamitani (2009) is another recent example where features are learned from data using canonical correlation analysis, albeit without making use of a hierarchical architecture as employed in this letter. The hierarchical nature of our model allows low-level reconstructions to be influenced by complex image features. Such deep models do more justice to the hierarchical organization of the neocortex and should lead to improved reconstruction performance since we may benefit from the relation between complex image features and response properties of neurons in higher cortical areas (Hedgé & Van Essen, 2000).

We show that our hierarchical generative model is able to reconstruct individual handwritten digits with low reconstruction error where reconstruction quality, as determined by a behavioral experiment, improves when the hierarchy consists of multiple layers.

2 Hierarchical Generative Model

The aim of the hierarchical generative model, shown in Figure 1, is to provide a model of the interactions between the stimulus **v**, latent causes **h**, and brain activity **z**. We start from the assumption that the latent causes can be modeled in terms of a hierarchy of processing units that detect increasingly complex features (statistical invariances), analogous to the hierarchical organization of visual cortex (Felleman & Van Essen, 1991). The key idea of our study is to use an unsupervised learning phase to learn the latent causes that best explain observed data and to use a supervised learning phase in order to learn how observed hemodynamic response is linked to these latent causes. The resulting hierarchical generative model captures how brain activity arises from the invariances in our environment. Reconstruction is achieved by sampling from the model conditional on observed brain activity.

2.1 Unsupervised Learning Phase. In the unsupervised learning phase, we disregard the conditional part of the model and focus only on learning a hierarchy of features. One way to represent such a feature hierarchy is in terms of a deep belief network (Hinton et al., 2006), which consists of smaller modules, known as restricted Boltzmann machines. We first describe the theory behind (restricted) Boltzmann machines and subsequently describe how they can be used to compose a deep belief network.

A Boltzmann machine (Hinton & Sejnowski, 1983) is a network of symmetrically coupled units that associates a scalar energy to each state $\mathbf{x}$ of the variables of interest:

$$p(\mathbf{x}) = \frac{\exp(-E(\mathbf{x}))}{Z}, \tag{2.1}$$

where $Z = \sum_{\mathbf{x}} \exp(-E(\mathbf{x}))$ is the partition function and $E(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ is the energy of a state with weight matrix $\mathbf{W}$ and bias terms $\mathbf{b}$. A Boltzmann machine normally consists of stochastic binary (conditional Bernoulli) units where the probability of a unit x_i being active given the states of the remaining units is given by the following Gibbs sampling update rule:

$$p(x_i = 1 \mid \mathbf{x}_{-i}) = \sigma\left(b_i + \sum_{j \neq i} w_{ij} x_j\right), \tag{2.2}$$

with sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$.

Parameter learning becomes interesting when some of the units are visible and the remaining units are hidden: $\mathbf{x} = (\mathbf{v}, \mathbf{h})$. The hidden units $\mathbf{h}$ then act as latent variables that model distributions over the visible state vectors $\mathbf{v}$ that cannot be modeled by direct pairwise interactions between visible units. The average gradient of the log likelihood over a training set $D = \{\mathbf{v}^n\}_{n=1}^N$ with respect to one of the model parameters θ is then given by

$$\mathrm{E}_{\hat{p}}\left(\frac{\partial \log p(\mathbf{v})}{\partial\theta}\right) = \mathrm{E}_p\left(\frac{\partial F(\mathbf{v})}{\partial\theta}\right) - \mathrm{E}_{\hat{p}}\left(\frac{\partial F(\mathbf{v})}{\partial\theta}\right), \tag{2.3}$$

where p is the model distribution, $\hat{p}$ is the empirical distribution, and $F(\mathbf{v}) = -\log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$ is the free energy. Learning in Boltzmann machines is hard since it takes a long time to reach the equilibrium distribution and the learning signal is noisy, as it is the difference of two sampled expectations. Fortunately, learning becomes easier when one makes use of restricted Boltzmann machines (RBMs) (Smolensky, 1986), where interactions are restricted to taking place only between visible and hidden units such that they can be described in terms of a bipartite graph.

Learning in restricted Boltzmann machines can be carried out more efficiently using the notion of contrastive divergence (Hinton, 2002; Hinton et al., 2006). The energy function of a restricted Boltzmann machine is bilinear,

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{h}^T\mathbf{W}\mathbf{v} - \mathbf{c}^T\mathbf{v} - \mathbf{b}^T\mathbf{h}, \tag{2.4}$$

such that the free energy of the input can be computed efficiently using the distributive law of probability theory:

$$\begin{aligned} F(\mathbf{v}) &= -\log \sum_{\mathbf{h}} \exp(\mathbf{h}^T\mathbf{W}\mathbf{v} + \mathbf{c}^T\mathbf{v} + \mathbf{b}^T\mathbf{h}) \\ &= -\mathbf{c}^T\mathbf{v} - \log \sum_{h_1,\ldots,h_n} \prod_i \exp\left(h_i\left(b_i + \sum_j w_{ij}v_j\right)\right) \\ &= -\mathbf{c}^T\mathbf{v} - \sum_i \log \sum_{h_i} \exp\left(h_i\left(b_i + \sum_j w_{ij}v_j\right)\right). \end{aligned} \tag{2.5}$$

Furthermore, for a restricted Boltzmann machine, we readily obtain the following conditionals:

$$\begin{aligned} p(\mathbf{h} \mid \mathbf{v}) &= \prod_i p(h_i \mid \mathbf{v}) = \prod_i \sigma\left((2h_i - 1)\left(b_i + \sum_j w_{ij}v_j\right)\right) \\ p(\mathbf{v} \mid \mathbf{h}) &= \prod_j p(v_j \mid \mathbf{h}) = \prod_j \sigma\left((2v_j - 1)\left(c_j + \sum_i w_{ij}h_i\right)\right). \end{aligned}$$

The factorization enjoyed by RBMs brings about two benefits. First, $\mathrm{E}_{\hat{p}}(\frac{\partial F(\mathbf{v})}{\partial \theta})$ can be computed analytically. Second, the set of variables in $(\mathbf{v}, \mathbf{h})$ can be sampled in two substeps in each step of a Gibbs sampling chain. We sample $\mathbf{h}$ given $\mathbf{v}$ and then a new $\mathbf{v}$ given $\mathbf{h}$, starting from a training example $\mathbf{v}_1$ (by sampling from the empirical distribution $\hat{p}$), such that after k steps, we obtain $\mathbf{v}_{k+1} \sim p(\mathbf{v} \mid \mathbf{h}_k)$. The idea of k-step contrastive divergence (CD-k) involves an approximation that introduces some bias in the gradient. We run the chain for only k steps to obtain the following parameter update after seeing an example $\mathbf{v}_1$,

$$\Delta\theta \propto \frac{\partial F(\mathbf{v}_{k+1})}{\partial\theta} - \frac{\partial F(\mathbf{v}_1)}{\partial\theta}, \tag{2.6}$$

where we replaced the averages over all inputs in equation 2.3 by single samples $\mathbf{v}_{k+1}$ and $\mathbf{v}_1$. Instead of sampling, it is also possible to use a mean field approach where expected activations instead of sampled states are propagated through the model (Welling & Hinton, 2002). We use this strategy when sampling hidden layer activations. The average gradient is obtained by averaging $\Delta\theta$ over many examples. Despite the bias introduced by these approximations, contrastive divergence works well in practice.

The feature hierarchy can be constructed by stacking restricted Boltzmann machines on top of each other, where the output of the previous RBM acts as input to the next RBM. The RBMs are trained using contrastive divergence and the stacked model, known as a deep belief network (DBN), is fine-tuned using the wake-sleep algorithm (Hinton, Dayan, Frey, & Neal, 1995; Hinton et al., 2006). The obtained model can be used to detect features by forward propagation of activations, but it can also be interpreted as a generative model that consists of a top-level associative memory that can be used to generate samples $\mathbf{y}$ (Hinton, 2007). In other words, given a state of the associative memory defined by the upper two layers, we can reconstruct the input. In our experiments, we will use gray-scale images as input to our model. One way to achieve this using stochastic binary units is by mapping real values to binary activation probabilities (Hinton et al., 2006). Furthermore, we will penalize large parameter values by adding a small weight decay term to the parameter updates to avoid overfitting.

2.2 Supervised Learning Phase. In order to use a DBN for neural decoding, we need to condition the model on observed brain activity $\mathbf{z}$ during reconstruction. One way to achieve this is to make use of conditional restricted Boltzmann machines (Salakhutdinov, Mnih, & Hinton, 2007; Taylor, Hinton, & Roweis, 2006) such that model parameters become a function of $\mathbf{z}$ (Bengio, 2009). Here, we assume that the bias terms $\mathbf{b}$ and $\mathbf{c}$ become linearly dependent on $\mathbf{z}$ simply by writing the energy function as

$$E(\mathbf{v}, \mathbf{h} \mid \mathbf{z}) = -\mathbf{h}^T\mathbf{W}\mathbf{v} - \mathbf{z}^T\mathbf{C}\mathbf{v} - \mathbf{z}^T\mathbf{B}\mathbf{h}\,. \tag{2.7}$$

It is assumed that $\mathbf{z}$ also includes a constant to model arbitrary offsets as in a standard RBM. It follows that the conditional free energy becomes

$$F(\mathbf{v} \mid \mathbf{z}) = -\mathbf{z}^T\mathbf{C}\mathbf{v} - \sum_i \log \sum_{h_i} \exp\left(h_i\left(\sum_k b_{ki}z_k + \sum_j w_{ij}v_j\right)\right), \tag{2.8}$$

whose derivatives can readily be computed (Taylor et al., 2006).

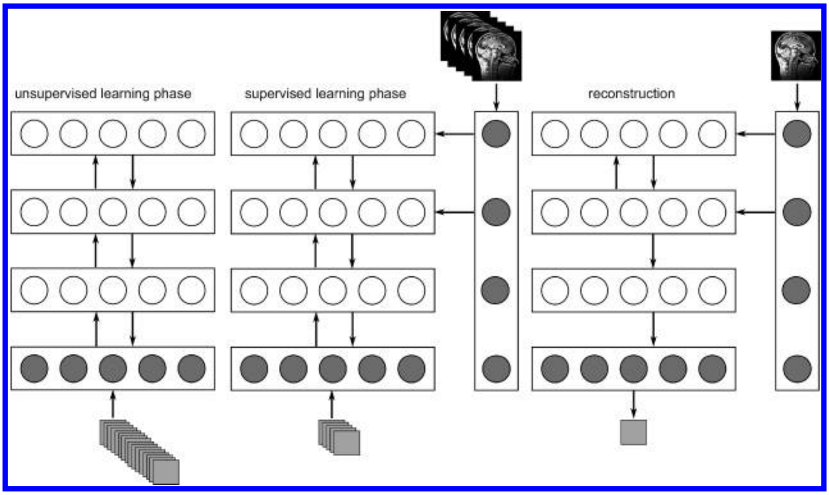

Figure 2: The experimental procedure consists of three stages. During the unsupervised learning phase, many stimuli are presented in order to learn their latent causes. During the supervised learning phase, a small number of stimuli are presented, together with brain activity, which was measured when subjects observed the stimuli. Finally, a stimulus that has not been previously seen by the model is reconstructed by sampling from the hierarchical generative model conditional on measured brain activity.

In the supervised learning phase, we replace the restricted Boltzmann machines that have previously been trained in the unsupervised phase by conditional restricted Boltzmann machines. Learning proceeds as before, but now we keep the interactions fixed and update only the bias terms based on the observed brain activity. Essentially this allows our observations to modulate the probability that certain feature detectors are active or inactive. In the following, we will update the bias terms only for the top-level associative memory—the top two layers in the hierarchy.

2.3 Reconstruction. In order to reconstruct perceived stimuli, we use the following procedure (see Figure 2). First, we learn the feature hierarchy in an unsupervised manner using thousands of stimuli. Second, we learn in a supervised manner how the biases are influenced by observed brain activity that is acquired while presenting a small number of stimuli. Finally, we generate a reconstruction using the hierarchical generative model by performing conditional sampling. That is, we perform one Gibbs sampling step in the top-level associative memory conditional on brain activity and a second unconditional Gibbs sampling step in the top-level associative memory, and then we propagate expectations back to the input layer. Reconstruction error between a stimulus $\mathbf{v}$ and its reconstruction $\mathbf{r}$ consisting of N pixels is quantified in terms of city-block distance: $\epsilon(\mathbf{v}, \mathbf{r}) = \frac{1}{N}\sum_{i=1}^{N} |v_i - r_i|$. In order to obtain a more subjective assessment of reconstruction performance, we asked five subjects to rate how well reconstructions match the stimuli.

3 Experiment

Stimuli consisted of 2106 handwritten gray-scale digits at a 28 × 28 pixel resolution taken from the training set of the MNIST database (http://yann.lecun.com/exdb/mnist). We selected 1000 handwritten 6s and 1000 handwritten 9s in order to train a hierarchical generative model (model stimuli). This subset of the available data was found to be sufficiently large for model estimation. Additionally, for the imaging experiment (see below), we selected 53 handwritten 6s and 53 handwritten 9s (presentation stimuli). The choice for these two digits was based on the consideration that the variation within and between these two digit classes was quite large. The former ensures that reconstruction will be nontrivial, while the latter allows us to assess more easily if digit specific features are learned. The model stimuli were downsampled from 28 × 28 pixel to 16 × 16 pixel images. The presentation stimuli were scaled to fit the full visual field.

We collected 106 trials in one participant. In each trial, a handwritten 6 or 9 was presented to the subject. The character remained visible for 12.5 seconds and flickered at a rate of 6 Hz on a black background. In order to ensure sustained attention during the entire scanning session, the subject's task was to maintain fixation to a fixation dot and to detect a brief (33 ms) change in color from red to green and back occurring once and randomly within a trial. Detection was indicated by pressing a button with the right-hand thumb as fast as possible. Trials were separated by a 12.5 second intertrial interval. The 106 trials were partitioned randomly into four runs interspersed with 30 second rest periods.

Blood-oxygenation-level dependent (BOLD) sensitive functional images were obtained by means of a Siemens 3T MRI system using a 32-channel coil for signal reception. We used a single-shot gradient EPI sequence with a repetition time (TR) of 2500 ms, echo time (TE) of 30 ms, and isotropic voxel size of 2 × 2 × 2 mm. Functional images were acquired in 42 axial slices in ascending order. A high-resolution anatomical image was acquired using an MP-RAGE sequence (TE/TR = 3.39/2250 ms; 176 sagittal slices, with isotropic voxel size of 1 × 1 × 1 mm).

Functional data were preprocessed and analyzed within the framework of SPM5 (Statistical Parametric Mapping, www.fil.ion.ucl.ac.uk/spm). Functional brain volumes were motion-corrected and coregistered with the anatomical scan. Functional data were detrended and high-pass-filtered. The volumes acquired 10 to 15 seconds after trial onset were averaged in order to obtain an estimate of the steady-state response in individual voxels. As input to our reconstruction model, we used the 1000 voxels that showed

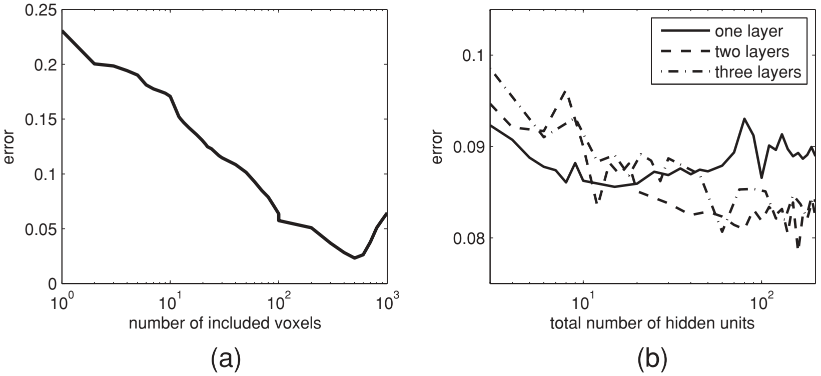

Figure 3: Reconstruction error as a function of the number of included voxels (a) and as a function of the number of hidden units for models consisting of one, two, or three hidden layers (b).

the largest difference between task and rest conditions according to a standard general linear model (GLM) analysis. As expected, the selected voxels were almost exclusively located in the occipital lobe, which contains the main cortical visual areas.

When training the models in the unsupervised phase, we ran the contrastive divergence algorithm (CD-1) followed by the wake-sleep algorithm for both 100 iterations using a learning rate of 0.1 and a weight penalty of 0.001 on all 2000 model stimuli. One iteration consisted of a single pass over the training data, which was partitioned into minibatches of 100 trials. All reconstructions were computed using a leave-one-out cross-validation scheme. For each of the 106 presentation stimuli, all 105 remaining stimuli were used to train a model in the supervised phase by running CD-1 for 100 iterations using a small learning rate of 0.0001 and a weight penalty of 0.001. Subsequently, a reconstruction was produced for the remaining stimulus using Gibbs sampling as described above.

4 Results

We start by computing reconstruction error when predicting the gray value of individual pixels directly from BOLD response. This can be interpreted as using just the visible layer of an RBM, where gray values are taken to be the posterior probabilities of the hidden unit activations and which is equivalent to solving a set of independent logistic regression problems. Figure 3a shows the decrease in reconstruction error as the number of included voxels increases. Note the slight increase in error when all voxels are included, possibly due to overfitting. Reconstruction error averaged over all images based on 1000 included voxels was 0.063.

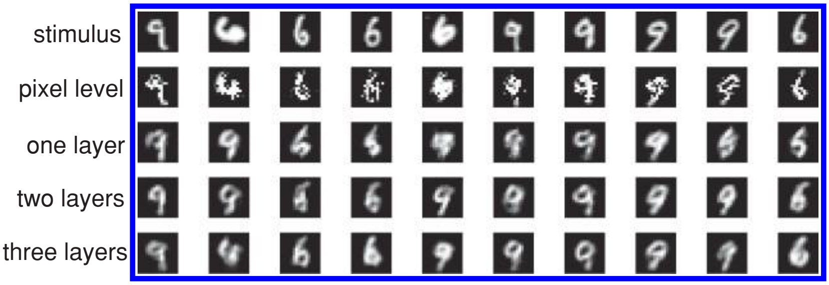

Figure 4: Image reconstructions for the first 10 images for all different models.

We now move on to the hierarchical generative model. Our primary goal is to determine whether adding a layer of hidden units improves the reconstruction. Figure 3b depicts the average reconstruction error as a function of the number of hidden units for one, two, or three hidden layers. Although there is quite some variability in the reconstruction error, it is clear that the reconstruction error decreases as a function of the number of hidden units. Furthermore, models with two or three hidden layers seem to outperform models consisting of one hidden layer. Minimum reconstruction error for one-, two-, and three-layer models was 0.085, 0.081, and 0.082, respectively.

Note that the reconstruction error for the hierarchical generative models is larger than that of the pixel-level reconstructions. However, it is also well known that error measures such as city-block distance do not accurately capture the quality of the reconstructions as perceived by humans and alternative metrics have been defined that try to incorporate properties of the human visual system (Wang, Bovik, Sheikh, & Simoncelli, 2004). This is also apparent in the reconstructions shown in Figure 4. We quantified the subjective experience of reconstruction performance by asking five naive subjects to rank each of the reconstructions according to how well they match the stimuli. The histogram in Figure 5 shows that on average, the pixel-level model and the one-layer model give less preferred reconstructions compared to the two-layer and three-layer reconstructions. Differences in preference among all models were significant as computed by a Wilcoxon rank sum test ($p = 0.01$).

In order to gain an understanding of what invariances are represented by our model, we can visualize the features that have been learned by the hidden units simply by representing the elements of the weight matrix belonging to a hidden unit as a 16 × 16 image. The features learned in the first layer of the hierarchical model resemble Gabor filters; that is, the features are optimally responsive for some location, orientation, and spatial frequency (one of these features is shown in the top left of Figure 6). Features learned by layers higher up the hierarchy are more of a distributed nature. The hierarchical generative model also allows us to probe which voxels code for

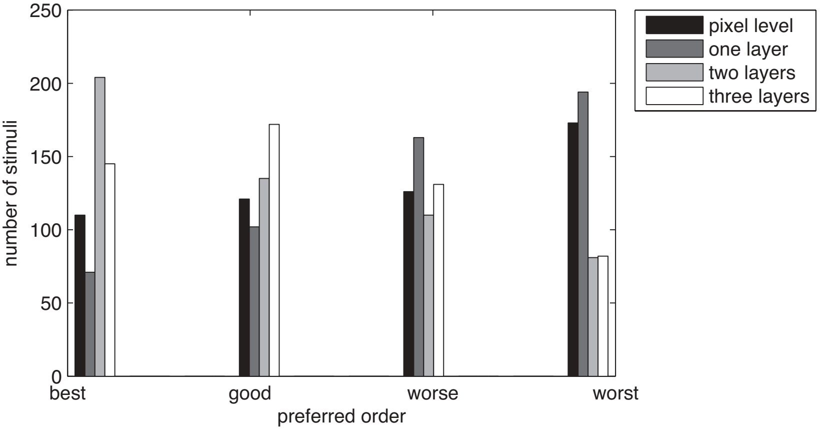

Figure 5: Preferred order of the models determined from their reconstructions by five naive subjects.

which feature. For example, Figure 6 shows positive and negative voxel contributions to one of the feature nodes in a one-layer model. Both early visual and lateral occipital cortex contribute to this feature. As would be expected, activity in early visual cortex decreases the likelihood of the feature being active, in line with the feature coding (among others) for the absence of visual stimulation in the visual field. Involvement of the lateral occipital complex is well in line with previous findings that this region is one of the sites involved in the encoding of letters and letter strings (Vinckier et al., 2007).

5 Discussion

We have demonstrated that hierarchical generative models can be used to obtain good-quality reconstructions of images based on measured hemodynamic response. We have also shown that reconstructions obtained by deep (multilayered) models lead to generally preferred solutions. Furthermore, hierarchical generative models can be used to probe how individual voxels influence the activation and deactivation of features and, as a consequence, to learn about the neural encoding of certain stimulus characteristics. Inspection of anatomical localization of feature voxels shows that both early visual and lateral occipital cortex contribute to the reconstruction of digits.

Our motivation for using hierarchical generative models was twofold. First, we wanted to learn stimulus features from data instead of using a fixed basis set. This allows the model to adapt to the statistics of the data at hand. This said, it may be possible to use a large data set of natural image patches in order to learn a generic basis set that can be applied to any data set (Lee, Grosse, Ranganath, & Ng, 2009). Second, we wanted

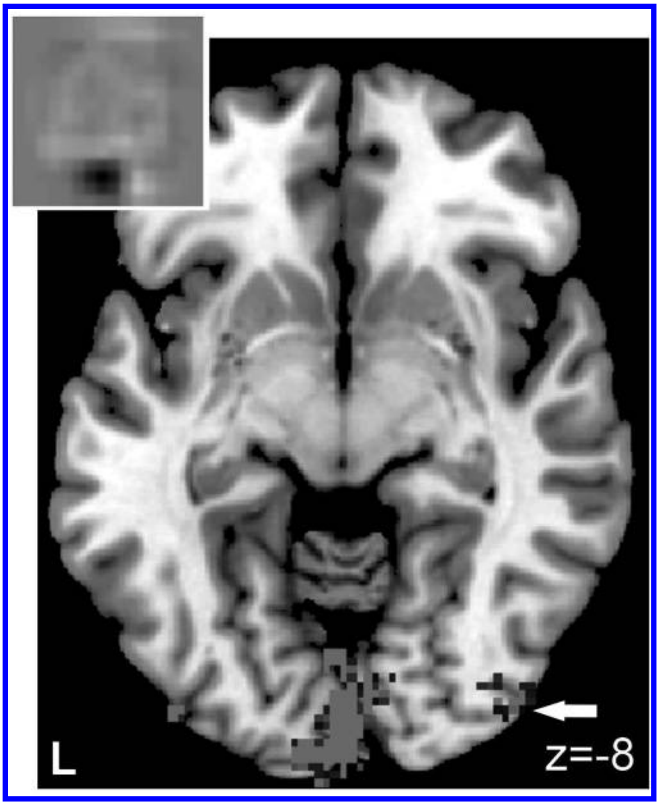

Figure 6: Anatomical localization of voxels that contribute to one of the features in a one-layer model. The corresponding feature is shown in the top left panel. The gray blob near the calcarine sulcus represents negative parameter values, and the gray blob in lateral occipital cortex, indicated by the arrow, represents positive parameter values.

to use a hierarchical architecture such that complex features could also influence reconstruction performance. Although deep models gave better reconstructions than shallow models, there was no clear relation between layers in the visual hierarchy and layers in the model. Voxels in early visual cortex seemed to be most predictive for each layer in the model.

In order to assess reconstruction performance, we have used both city-block distance, and a subjective measure of reconstruction performance. In terms of city-block distance, pixel-level reconstructions gave the best performance, whereas in terms of the subjective measure, a two-layer model gave the best performance. The reason for this apparent contradiction is due to the fact that city-block distance, although useful for testing convergence, is not an optimal measure of reconstruction performance. Reconstructions using deep models are obtained as compositions of the individual feature activations. Since these features have some fixed spatial layout, they may activate pixels that were not active in the original model. Furthermore, although deep models generally lead to high-quality smooth reconstructions, they can in some cases generate a reconstruction that belongs to the wrong stimulus class. For example, stimuli 2 and 5 in Figure 4 are erroneously reconstructed as a 9 instead of a 6, whereas the pixel-level reconstructions still show some correspondence with the original image. This occasional misclassification will induce a large contribution to the city-block distance. The subjective measure, in contrast, will be tolerant of small differences (e.g., rotations or translations) between an original image and its reconstruction while penalizing the nonsmooth appearance of the pixel-level reconstructions.

In our experiment, a simple block design was used where stimuli were presented as flashing stimuli for prolonged periods. This yielded high-quality reconstructions because we could make use of the high signal-to-noise ratio that is afforded by the steady-state response. At the same time, this prohibited the presentation of a large number of stimuli and therefore necessitated the use of a restricted stimulus set consisting of handwritten 6s and 9s only. This also simplified the reconstruction problem since the stimulus set could be modeled by a relatively small set of features. Other studies have used more rapid designs where flashing stimuli were presented for only brief periods (Kay et al., 2008), thus allowing the presentation of larger and more variable stimulus sets. It remains an open question whether accurate reconstruction can be obtained using our model in such settings.

We have also made use of a GLM analysis in order to identify voxels that show significant differences between task and rest conditions. This allowed us to reduce the number of voxels that were used as input to the model and reduce the negative effect of overfitting. Our focus on a small subset of voxels made the interpretation of the relation between those voxels and the learned features difficult. Interpretation may be facilitated by including more voxels and using regularizers that induce smoothness and sparseness (van Gerven, Cseke, de Lange, & Heskes, 2010). Together with the use of retinotopic mapping, this allows one to probe more accurately which cortical areas code for which image features.

In conclusion, we have demonstrated that hierarchical generative models can be used for neural decoding and offer a new window into the brain. If performance could be generalized to imagined stimuli, we will come closer to a system that is able to "read thoughts" (Kay & Gallant, 2009). There is reason to believe that such a generalization is feasible due to the fact that perceived and imagined stimuli can lead to similar neural activation patterns (Roland & Gulyas, 1995; Mellet, Petit, Mazoyer, Denis, & Tzourio, 1998; Koch, 2004; Thirion, Duchesnay, Hubbard, Dubois, Poline, Lebihan, et al., 2006; Stokes et al., 2009).

Acknowledgments

We gratefully acknowledge the support of the Dutch technology foundation STW (project number 07050), the Netherlands Organization for Scientific Research NWO (Veni grant 451.09.001 and Vici grant 639.023.604), and the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

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

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex, 1*, 1–47.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci., 360*(1456), 815–836.

Fujiwara, Y., Miyawaki, Y., & Kamitani, Y. (2009). Estimating image bases for visual image reconstruction from human brain activity. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22* (pp. 576–584). Cambridge, MA: MIT Press.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., & Maguire, E. A. (2009). Decoding neuronal ensembles in the human hippocampus. *Current Biology, 19*, 546–554.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430.

Hedgé, J. & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *J. Neurosci., 20*(RC61), 1–6.

Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. New York: Dover.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation, 14*(8), 1711–1800.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. In P. Cisek, T. Drew, & J. Kalaska (Eds.), *Computational neuroscience: Theoretical insights into brain function*. Burlington, MA: Elsevier.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science, 268*, 1158–1161.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.

Hinton, G. E. & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE.

Kay, K. N. & Gallant, J. L. (2009). I can see what you see. *Nature Neuroscience, 12*(3), 245–246.

Kay, K., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*, 352–355.

Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Greenwood Village, CO: Roberts & Company.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 609–616). New York: ACM.

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A, 20*(7), 1434–1448.

Mellet, E., Petit, L., Mazoyer, B., Denis, M., & Tzourio, N. (1998). Reopening the mental imagery debate: Lessons from functional anatomy. *NeuroImage, 8*(2), 129–139.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron, 60*(5), 915–929.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*(6), 902–915.

Rao, R. P., & Ballard, D. H. (1998). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci., 2*, 79–87.

Roland, P. E., & Gulyas, B. (1995). Visual memory, visual imagery, and visual recognition of large field patterns by the human brain: Functional anatomy by positron emission tomography. *Cerebral Cortex, 5*, 79–93.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Vol. 1: Foundations* (pp. 194–281). Cambridge, MA: MIT Press.

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *J. Neurosci., 29*(5), 1565–1572.

Taylor, G. W., Hinton, G. E., & Roweis, S. (2006). Modeling human motion using binary latent variables. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, 20*. Cambridge, MA: MIT Press.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J., Lebihan, D., et al. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage, 33*(4), 1104–1116.

van Gerven, M. A. J., Cseke, B., de Lange, F. P., & Heskes, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage, 50*(1), 150–161.

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron, 55*, 143–156.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Welling, M., & Hinton, G. E. (2002). A new learning algorithm for mean field Boltzmann machines. In *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks* (pp. 351–372). Berlin: Springer-Verlag.

Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature, 335*, 311–331.

Received February 19, 2010; accepted May 13, 2010.

This article has been cited by: