NeuroImage 198 (2019) 125–136

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex

Kuan Han[b,c], Haiguang Wen[b,c], Junxing Shi[b,c], Kun-Han Lu[b,c], Yizhen Zhang[b,c], Di Fu[b,c], Zhongming Liu[a,b,c,*]

[a] Weldon School of Biomedical Engineering, USA
[b] School of Electrical and Computer Engineering, USA
[c] Purdue Institute for Integrative Neuroscience, Purdue University, West Lafayette, IN, 47906, USA

ARTICLE INFO

Keywords:
Neural coding
Variational autoencoder
Bayesian brain
Visual reconstruction

ABSTRACT

Goal-driven and feedforward-only convolutional neural networks (CNN) have been shown to be able to predict and decode cortical responses to natural images or videos. Here, we explored an alternative deep neural network, variational auto-encoder (VAE), as a computational model of the visual cortex. We trained a VAE with a five-layer encoder and a five-layer decoder to learn visual representations from a diverse set of unlabeled images. Using the trained VAE, we predicted and decoded cortical activity observed with functional magnetic resonance imaging (fMRI) from three human subjects passively watching natural videos. Compared to CNN, VAE could predict the video-evoked cortical responses with comparable accuracy in early visual areas, but relatively lower accuracy in higher-order visual areas. The distinction between CNN and VAE in terms of encoding performance was primarily attributed to their different learning objectives, rather than their different model architecture or number of parameters. Despite lower encoding accuracies, VAE offered a more convenient strategy for decoding the fMRI activity to reconstruct the video input, by first converting the fMRI activity to the VAE's latent variables, and then converting the latent variables to the reconstructed video frames through the VAE's decoder. This strategy was more advantageous than alternative decoding methods, e.g. partial least squares regression, for being able to reconstruct both the spatial structure and color of the visual input. Such findings highlight VAE as an unsupervised model for learning visual representation, as well as its potential and limitations for explaining cortical responses and reconstructing naturalistic and diverse visual experiences.

1. Introduction

Humans readily make sense of the visual surroundings through complex neuronal circuits. Understanding the human visual system requires not only measurements of brain activity but also computational models built upon hypotheses about neural computation and learning (Kietzmann et al., 2019). A model that truly reflects the brain's algorithmic mechanism of vision should be image-computable and capable of predicting brain responses to any visual input (namely encoding) and retrieving visual and conceptual information from brain responses (namely decoding). In this sense, evaluating the models' encoding and decoding performance can, at least in part, serve to test and compare hypotheses about how the brain learns and organizes visual representations (Wu et al., 2006).

In one class of hypotheses, the visual system consists of feature detectors that progressively extract and integrate features for pattern recognition. For example, Gabor and wavelet filters have been shown to model low-level features (Hubel and Wiesel, 1962; van Hateren and van der Schaaf, 1998), and explain brain responses in early visual areas (Kay et al., 2008; Nishimoto et al., 2011). As another example, supervised convolutional neural networks (CNNs) encode hierarchical visual features in a fully-computable feedforward model (LeCun et al., 2015) and enable intelligent tasks in computer vision (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Recent studies have shown that CNNs use similar representations as does the visual cortex (Cichy et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014), and yield the state-of-the-art accuracy in encoding and decoding brain responses to natural image or video stimuli (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Horikawa and Kamitani, 2017; Seeliger et al., 2018a; Wen et al., 2018a; Yamins et al., 2014). Although they are becoming popular

[*] Corresponding author. College of Engineering, Purdue University, 206 S. Martin Jischke Dr., West Lafayette, IN, 47907, USA.
E-mail address: zmliu@purdue.edu (Z. Liu).

https://doi.org/10.1016/j.neuroimage.2019.05.039
Received 13 January 2018; Received in revised form 13 April 2019; Accepted 15 May 2019
Available online 16 May 2019

models in computational visual neuroscience (Kriegeskorte, 2015; Yamins and DiCarlo, 2016), CNNs are unlike the brain in many aspects. As perhaps the most notable distinctions, the brain does not always learn by supervision but often learns from experiences without supervision (Barlow, 1989), and the brain uses bidirectional (both feedforward and feedback) pathways (Bastos et al., 2012; Salin and Bullier, 1995), whereas CNNs are only feedforward.

In another class of hypotheses, the brain is mostly unsupervised (Barlow, 1989; Hinton et al., 1999; Seung and Lee, 2000). For example, unsupervised visual learning may utilize "analysis by synthesis" (Hinton et al., 1995; Yuille and Kersten, 2006). In this notion, the bottom-up process infers a representation of the input to support perception. The top-down process tries to reconstruct (or predict) the input from the inferred representation to ensure its consistency with the input (Yuille and Kersten, 2006) (Fig. 1A). Both bottom-up and top-down processes are optimized such that visual experiences can be explainable by the brain (Dayan et al., 1995; Hinton and Zemel, 1994; Rao and Ballard, 1999). This hypothesis about unsupervised visual learning takes into account both bottom-up and top-down pathways in the brain and reconciles the humans' ability to readily construct mental images (e.g. during an imagery or dream). It is thus compelling for computational neuroscience (Bastos et al., 2012; Friston, 2010; Rao and Ballard, 1999; Yuille and Kersten, 2006) as well as artificial intelligence (Hinton et al., 1995; Lotter et al., 2016; Mirza et al., 2016).

A similar notion of unsupervised learning has been explored for artificial intelligence. Variational auto-encoder (VAE) uses independent "latent" variables to represent input images (Kingma and Welling, 2013). VAE learns the latent variables from images via an encoder and samples the latent variables to generate new images via a decoder. Both the encoder and the decoder are neural networks trainable from unlabeled images (Doersch, 2016). As a model of unsupervised learning, VAE is a potentially plausible model of the brain's visual system in the computational level, and may enable an effective way to decode brain activity during either visual perception or imagery (Du et al., 2018; Güçlütürk et al., 2017; Naselaris et al., 2009; Nishimoto et al., 2011; Seeliger et al., 2018b; Shen et al., 2019; van Gerven et al., 2010; Wen et al., 2018a). To test VAE as a brain model, we built and trained a VAE to learn latent representations of natural images without requiring any image label assigned for training, and evaluated the trained VAE in terms of its usability for encoding and decoding human functional magnetic resonance imaging (fMRI) responses to naturalistic movie stimuli (Fig. 1B).

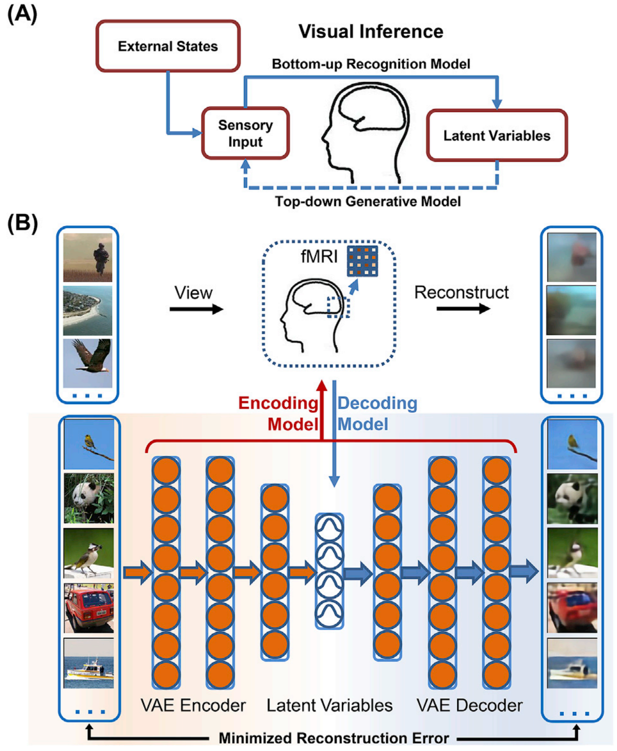

Fig. 1. An "analysis by synthesis" model of human vision. (A) The brain learns the visual world through synthesizing sensations. The brain analyzes the sensory input to infer the hidden representations of the input through its bottom-up processes and synthesizes the sensory input through its top-down processes. **(B) Encoding and decoding visually-evoked cortical fMRI responses by using VAE as a model of the visual cortex.** For encoding, cortical responses to visual stimuli were predicted as a linear projection of all hidden units in VAE to the same stimuli. For decoding, visual stimuli were reconstructed by first estimating the VAE's latent variables as a linear function of the fMRI responses, and then generating pixel patterns from the estimated latent variables through the VAE's decoder.

2. Methods and materials

2.1. Theory: variational auto-encoder

In general, VAE uses a deep neural network to learn representations from complex data without supervision (Kingma and Welling, 2013). A VAE includes an encoder and a decoder, both of which are neural networks. The encoder learns latent variables from the input and the decoder generates an output based on samples of the latent variables. Given sufficient training data, the encoder and the decoder are trainable altogether by minimizing the reconstruction loss and the Kullback-Leibler (KL) divergence between the distributions of latent variables and independent normal distributions (Doersch, 2016). When the input data are natural images, the decoder models the forward process of image formation (namely the generative model), the encoder models the inverse process of inference (namely the inference model), and the learned latent variables should represent the hidden causes (or factors) that have generated the images.

Let z be the latent variables and $\boldsymbol{x}$ be an image. The encoder parameterized with $\boldsymbol{\phi}$ infers z from $\boldsymbol{x}$, and the decoder parameterized with $\boldsymbol{\theta}$ generates $\boldsymbol{x}$ from z. In VAE, both z and $\boldsymbol{x}$ are random variables. The likelihood of $\boldsymbol{x}$ given z under the generative model with $\boldsymbol{\theta}$ is denoted as $p_{\boldsymbol{\theta}}(\boldsymbol{x}|z)$. The probability of z given $\boldsymbol{x}$ under the inference model with $\boldsymbol{\phi}$ is denoted as $q_{\boldsymbol{\phi}}(z|\boldsymbol{x})$. The marginal likelihood of data can be written as the following form.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = D_{KL}\left[q_{\boldsymbol{\phi}}(z|\boldsymbol{x})\middle|\middle|p_{\boldsymbol{\theta}}(z|\boldsymbol{x})\right] + L(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) \tag{1}$$

Since the KL divergence in Eq. (1) is non-negative, $L(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x})$ can be regarded as the lower-bound of data likelihood and also can be rewritten as Eq. (2). For VAE, the learning rule is to optimize $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by maximizing $L(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x})$ given the training samples of $\boldsymbol{x}$.

$$L(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) = -D_{KL}\left[q_{\boldsymbol{\phi}}(z|\boldsymbol{x})\middle|\middle|p_{\boldsymbol{\theta}}(z)\right] + E_{z\sim q_{\boldsymbol{\phi}}(z|\boldsymbol{x})}\left[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|z))\right] \tag{2}$$

In this objective function, the first term is the KL divergence between the distribution of z inferred from $\boldsymbol{x}$ and the prior distribution of z, both of which are assumed to follow a multivariate normal distribution. The second term is the expectation of the log-likelihood that the input image can be generated based on the sampled values of z from the inferred distribution $q_{\boldsymbol{\phi}}(z|\boldsymbol{x})$. When $q_{\boldsymbol{\phi}}(z|\boldsymbol{x})$ is a multivariate normal distribution with unknown expectations $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$, the objective function is differentiable with respect to ($\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) (Kingma and Welling, 2013). The parameters in VAE could be optimized iteratively using stochastic gradient-descent algorithms, e.g. the Adam optimizer (Kingma and Ba, 2014).

2.2. Training VAE with diverse natural images

We designed a VAE with 1,024 latent variables, and the encoder and the decoder were both convolutional neural networks with five hidden layers (Fig. 2A). Each convolutional layer included nonlinear units with a

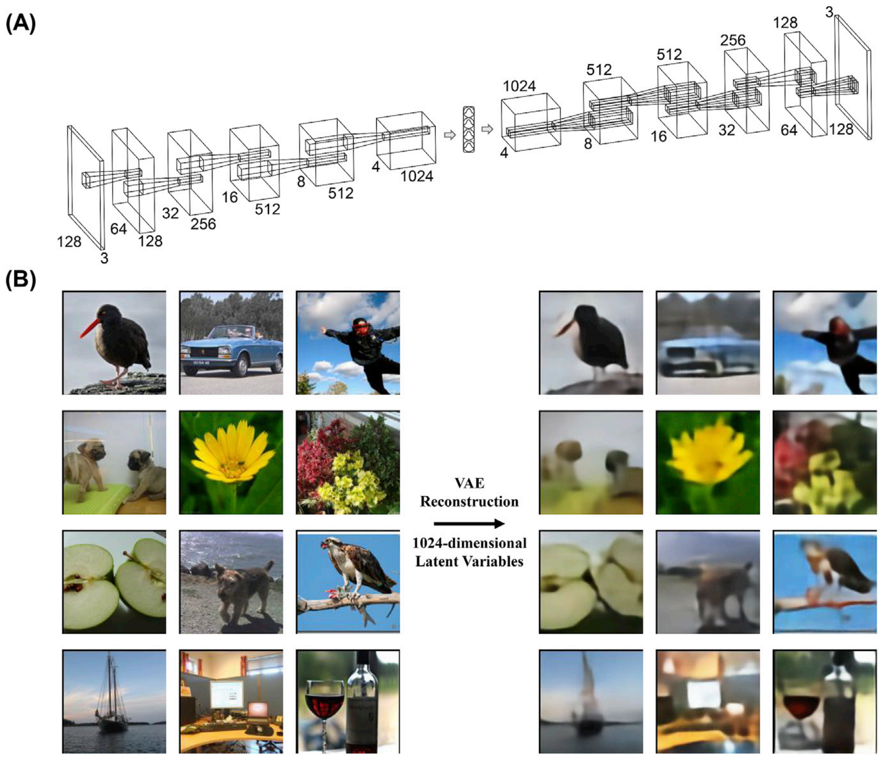

Fig. 2. Variational auto-encoder. (A) The architecture of VAE for natural images. The encoder and the decoder both contained 5 layers. Each layer of the encoder contained convolution units (kernel size = 4, stride = 2, padding = 1) and rectified linear units (ReLU). Each layer of the decoder contained transposed convolution units (kernel size = 4, stride = 2, padding = 1) and ReLU, except the last layer which replaced ReLU with sigmoid transformation. Fully-connected layers were used to transform the encoder's last layer to latent variables or to transform the re-parameterized latent variables to the decoder's first layer. The dimension of the latent variables was 1024. **(B) Reconstruction of natural images by VAE.** For any image (left), its information was encoded by 1024 latent variables by passing it through the VAE encoder. From the latent variables, the VAE decoder generated the corresponding reconstruction (right) of the input image, despite blurred details.

Rectified Linear Unit (ReLU) function (Nair and Hinton, 2010), except the last layer in the decoder where a sigmoid function was used to generate normalized pixel values between 0 and 1. The model was trained on the ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015) with every training image resized to 128 × 128 × 3. To enlarge the amount of training data, the original training images were randomly flipped in the horizontal direction, resulting in >2 million training samples in total. The training data were divided into batches with a size of 200. For each training image, the pixel intensities were normalized to [0, 1]; the normalized intensity was viewed as the probability of color emission (Gregor et al., 2015). To train the VAE, the Adam optimizer was used with a learning rate of 10^{-4}. The model was implemented in *PyTorch* (http://pytorch.org/).

2.3. Experimental data

Three healthy volunteers (all female, age: 23–26) participated in this study with informed written consent according to a research protocol approved by the Institutional Review Board at Purdue University. As described in detail elsewhere (Wen et al., 2018a), the experimental design and data were summarized as below. Each subject watched a diverse set of natural videos for a total length up to 13.7 h. The videos were downloaded from Videoblocks and YouTube, and were then separated into two independent sets. One data set was for training the models to predict the fMRI responses based on the input video (i.e. the encoding models) as well as the models to reconstruct the input video based on the measured fMRI responses (i.e. the decoding models). The other data set was for testing the trained encoding or decoding models. The videos in the training and testing datasets were independent for unbiased model evaluation. Both the training and testing movies were further split into 8-min segments, each of which was used as the visual stimulation ($20.3^\circ \times 20.3^\circ$) along with a central fixation cross ($0.8^\circ \times 0.8^\circ$) presented via an MRI-compatible binocular goggle during a single fMRI session. The training movie included 98 segments (13.1 h) for Subject 1, and 18 segments (1.6 h) for Subject 2 & 3. The testing movie included 5 segments (40 min in total). Each subject watched the testing movie 10 times. All five segments of the testing movie were used to test the encoding model. One of the five segments of the testing movie was used to test the decoding models for visual reconstruction, since that segment contained video clips that were continuous over relatively long periods (mean ± std: 13.3 ± 4.8 s).

MRI/fMRI data were collected from a 3-T MRI system, including anatomical MRI (T_1 and T_2 weighted) of 1 mm isotropic spatial resolution, and blood oxygenation level dependent (BOLD) fMRI with 2-s temporal resolution and 3.5 mm isotropic spatial resolution. The fMRI data were registered onto anatomical MRI data, and were further co-registered on a cortical surface template (Glasser et al., 2013). The fMRI data were preprocessed with the minimal preprocessing pipeline released for the human connectome project (Glasser et al., 2013).

2.4. VAE-based encoding models

After training, VAE extracted the latent representation of any video by a feed-forward pass of every video frame into the encoder and reconstructed every video frame by a feedback pass of the latent representation into the decoder. To predict cortical fMRI responses to the video stimuli, an encoding model was defined separately for each voxel as a linear regression model (Güçlü and van Gerven, 2014; Naselaris et al., 2011). The voxel-wise fMRI signal was estimated as a linear combination of all unit activities in both the encoder and the decoder given the input video. Every unit activity in VAE was convolved with a canonical hemodynamic response function (HRF). For dimension reduction, PCA was first applied to the HRF-convolved unit activities from each layer, keeping 99% of the variance of the layer-wise activity given the training movies. Then, the layer-wise activity was concatenated across layers; PCA was applied again to the concatenated activity to keep 99% of the variance of the activity from all layers given the training movies. This method was explained in greater detail in our earlier paper (Wen et al., 2018b). Following the dimension reduction, the principal components of unit-activity were used as the regressors to predict the fMRI signal at each voxel through a linear regression model specifically estimated for the voxel. In addition, we also defined two variations of the encoding model:

one based only on the VAE's encoder, the other based only on the VAE's decoder. For both variations, dimension reduction was done with the same procedure as described above.

In any above variation of the encoding model, the voxel-wise regression model was trained separately for each subject based on the subject's fMRI data observed during the training movie. Mathematically, let z stand for the regressors derived from visual input x based on both/either the encoder and/or the decoder in VAE. For each voxel i, the fMRI signal y_i was modeled as a linear function of the regressors, while the regression coefficients were assumed to be variable across voxels, as expressed in Eq. (3).

$$y_i = \mathbf{w}_i^T \mathbf{z} + b_i + \varepsilon_i \tag{3}$$

where $\mathbf{w}_i$ is a column vector of the regression coefficients, b_i is the bias term, and ε_i is the error or noise. The linear regression coefficients were estimated by least-squares estimation with L$_2$-norm regularization, as expressed by Eq. (4).

$$\widehat{\mathbf{w}}_\mathbf{i}, \widehat{b}_i = \underset{\mathbf{w}_\mathbf{i},\, b_i}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{j=1}^{N} \left(y_i^j - \mathbf{w}_i^T \mathbf{z}^j - b_i \right)^2 + \lambda_i ||\mathbf{w}_\mathbf{i}||_2^2 \right) \tag{4}$$

where N is the number of training samples, the superscript j refers to the individual training samples (or time points). The regularization parameter λ_i was determined, separately for each voxel, as the optimal value that minimized the loss in Eq. (4) in three-fold cross-validation. Once the optimal parameter λ_i was determined, the model was estimated by using the entire training dataset.

2.5. Evaluation of encoding performance

After the voxel-wise encoding models were trained, they were evaluated with the fMRI data observed during the testing movies. Note that the testing movies were different from the training movies to ensure unbiased model evaluation. For each voxel, the encoding performance was measured as the temporal correlation between the measured and predicted fMRI signals in response to the testing movie. In addition, we used the encoding model trained from one subject to predict other subjects' fMRI responses to the testing movies, to evaluate the ability of transferring encoding models across subjects. The performance was compared across the three encoding models based on the VAE as a whole, only its encoder part, or its decoder part.

Moreover, the voxel-wise correlation between the measured and predicted responses was evaluated for statistical significance based on a block permutation test (Adolf et al., 2014) with a block size of 24-sec and 100,000 permutations and corrected at false discovery rate (FDR) $q < 0.01$, as in our earlier papers (Shi et al., 2018; Wen et al., 2018a). The correlation was further compared against the so-called "noise ceiling", which indicated the upper limit of predictability given the presence of "noise" unrelated to the visual stimuli (David and Gallant, 2005; Nili et al., 2014). The noise ceiling was estimated using the method described elsewhere (Kay et al., 2013). Briefly, the noise was assumed to follow a Gaussian distribution with a zero mean and an unknown variance that varied from voxel to voxel. The response and noise were assumed to be independent and additive. The variance of the noise was estimated as the squared standard error of the mean of the fMRI response, which was obtained by averaging the fMRI signal across the 10 repeated sessions of each testing movie. The variance of the response was obtained by subtracting the variance of the noise from the variance of the data. From the signal and noise distributions, we drew samples of the response and the noise, respectively, based on Monte Carlo simulation for 1,000 random trials. For each trial, the signal was simulated as the sum of the simulated response and noise; its correlation coefficient with the simulated response was calculated. This resulted in an empirical distribution of the correlation coefficient for each voxel. The mean of the distribution was identified and interpreted as the noise ceiling, denoted as r_{NC}.

2.6. Comparison with CNN

In terms of the encoding performance, we compared the encoding models based on VAE against those based on CNN, which have been explored in recent studies, e.g. (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Wen et al., 2018a). Specifically, we used a 18-layer residual network (ResNet-18) (He et al., 2016) as an example of CNN. Relative to AlexNet (Krizhevsky et al., 2012) or ResNet-50 (He et al., 2016), ResNet-18 had an intermediate level of architectural complexity in terms of the number of layers and the total number of units. It was a suitable benchmark for comparison with VAE, which had a comparable level of complexity. Briefly, ResNet-18 consisted of 18 hidden layers organized into 6 blocks. The 1st block was a convolutional layer followed by max-pooling; the 2nd through 5th blocks were residual blocks, each being a stack of 3 convolutional layers with a shortcut connection; the 6th block performed the multinomial logistic regression for image classification. As a typical CNN, ResNet-18 encoded increasingly complex visual features from lower to higher layers.

We built and trained voxel-wise regression models to project the representations in ResNet-18 to voxel responses in the brain, using the same procedures and data as used for VAE-based encoding models (see the subsection ***VAE-based encoding models***). Then, we compared VAE against CNN (ResNet-18) in terms of their encoding performance in the level of voxels or regions of interest (ROI). In the voxel level, the voxel-wise accuracy of the VAE or CNN-based encoding model was converted from the correlation coefficient to the z value based on the Fisher's z-transform. Their difference in the z value was calculated by subtraction. For the ROI-level comparison, multiple ROIs were selected from existing cortical parcellation (Glasser et al., 2016), including V1, V2, V3, V4, lateral occipital (LO), middle temporal (MT), fusiform face area (FFA), para-hippocampal place area (PPA) and temporo-parietal junction (TPJ). The correlation coefficient was averaged over all voxels in each ROI and was compared between VAE and ResNet-18. Note that the areas herein referred to as FFA and PPA were originally named as the posterior parahippocampal cortex (PHC) (Arcaro et al., 2009) and fusiform face complex (FFC) in the original parcellation by Glasser et al. (2016). Since FFC and PHC overlapped largely with FFA and PPA, respectively, we prefer FFA/PPA to FFC/PHC since the former terms are more widely used in the neuroscience literature.

ResNet and VAE were different in their learning objectives, architectures, and numbers of free parameters. Their difference in the voxel-wise encoding performance was likely attributable to one or multiple of these factors. To pinpoint the primary factor, we constructed two additional CNN models, which were consistent with the VAE in terms of the architecture and the number of free parameters. For one CNN model (referred to as CNN-A, where "A" implies architecture-matched), its architecture was constrained to be identical to the architecture of the encoder in the VAE. For every convolutional layer, the size of feature maps and the number of kernels were the same for CNN-A and VAE. Global averaged pooling (Lin et al., 2013) was applied to the last convolutional layer, followed by logistic regression to output a probabilistic distribution over pre-defined image categories. CNN-A was trained for image classification based on labeled images in ImageNet (Russakovsky et al., 2015). Therefore, CNN-A and VAE shared the same architecture but used different learning objectives. Before using CNN-A for encoding voxel-wise fMRI responses to movie stimuli, the same algorithm for dimension reduction was applied to CNN-A as was applied to VAE. Despite the use of the same algorithm, the number of components (or regressors) that remained after dimension reduction was not necessarily identical for CNN-A and VAE. To further resolve this distinction, we also added another constraint such that the same number of principal components were retained for CNN-A as was retained for VAE. For the ease of distinction, we referred to the dimension-matched variation as CNN-AP (where "P" implies parameter-matched), since the same number of encoding parameters was used for the encoding model based on CNN-AP and VAE. The encoding models based on CNN-A and CNN-AP were trained and tested using the same procedures and the same data as were used for VAE.

We also compared the ROI-level encoding performance of VAE against those of the Gabor filters (Fogel and Sagi, 1989; Marcelja, 1980), specifically based on the implementation documented in Kay et al. (2008); Naselaris et al. (2011); Nishimoto et al. (2011) and publicized online.[1] Briefly, the video frames were converted to gray-scale. Wavelets were defined with 5 spatial frequencies (2, 4, 8, 16 or 32 cycles per FOV), 8 orientations (0°, 22.5°, …, 157.5°) and two phases (0° or 90°). Each pair of wavelets (with two phases) were squared and summed, giving rise to analytically predefined spatial features in the pixel space. In terms of these Gabor filters, the representation of every video frame was reduced to a lower dimension that kept 99% variance, based on a similar PCA as used for the dimension reduction in VAE and CNN. Note that Gabor filters could be defined with many variations. In this study, we only explored gray-scale Gabor filters as in Kay et al. (2008); Nishimoto et al. (2011). We also confined our analysis to Gabor filters in the spatial domain, excluding the motion-energy filters applicable to the spatiotemporal domain (Nishimoto et al., 2011). As such, this study only addressed several models of spatial processing, regardless of whether the models use visual features that are hand-engineered (for Gabor filters) or learned from data (for VAE or CNN). Among such models, the learnable network models were the primary focus.

2.7. Decoding fMRI for visual reconstruction

We trained and tested the decoding model for reconstructing visual input from distributed fMRI responses. The model contained two steps: 1) transforming the fMRI response pattern to the latent variables in VAE through a linear regression model, and 2) transforming the latent variables to pixel patterns through the VAE's decoder. Here we used a cortical mask that covered the visual cortex, and used the voxels within the mask as the input to the decoding model as in our previous study (Wen et al., 2018b).

Let $\boldsymbol{y}$ be a column vector representing the cortical pattern observed with fMRI, and $\boldsymbol{z}$ be a column vector of latent variables in VAE. As Eq. (5), a multivariate linear regression model was defined to predict $\boldsymbol{z}$ given $\boldsymbol{y}$.

$$\boldsymbol{z} = \mathbf{U}\boldsymbol{y} + \boldsymbol{c} + \boldsymbol{\varepsilon} \tag{5}$$

where $\mathbf{U}$ is a matrix that consists of the regression coefficients to transform a cortical pattern to a vector of latent variables, $\boldsymbol{c}$ is the bias term, and $\boldsymbol{\varepsilon}$ is the error term unexplained by the model. This model was estimated based on the data during the training movie.

To estimate the parameters of the decoding model, we minimized the loss function as Eq. (6) for least-squares estimation with L_1-regularization to prevent over-fitting.

$$\widehat{\mathbf{U}}, \widehat{\boldsymbol{c}} = \underset{\mathbf{U},\boldsymbol{c}}{\operatorname{argmin}}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{z}^j - \mathbf{U}\boldsymbol{y}^j - \boldsymbol{c}\right)^2 + \lambda\|\mathbf{U}\|_1^1\right) \tag{6}$$

where N is the number of data samples used for training the model. The regularization parameter, λ, was optimized to minimize the loss in three-fold cross-validation. To solve Eq. (6), we used the stochastic gradient-descent algorithm with a batch size of 100 and a learning rate of 10^{-7}. The testing movie was reconstructed frame by frame by passing the estimated latent variables through the decoder in VAE, as expressed by Eq. (7)

$$\widehat{\boldsymbol{x}}^j = \Theta\left(\widehat{\boldsymbol{z}}^j\right) = \Theta\left(\widehat{\mathbf{U}}\boldsymbol{y}^j + \widehat{\boldsymbol{c}}\right) \tag{7}$$

where Θ is the VAE's decoder for nonlinear mapping from the decoded latent variables to the reconstructed visual input.

2.8. Accounting for the hemodynamic delay

Since the fMRI response is delayed from neural response due to neurovascular coupling, the hemodynamic delay has to be taken into account for decoding the fMRI response into latent representations of visual input. For training the decoding model, we first convolved the latent variables with a canonical HRF. Next, we optimized the decoding parameters to estimate the HRF-convolved latent variables given the fMRI responses at every time point. For testing the decoding model, we used the trained decoding model to estimate the latent variables given the unknown visual input and the known fMRI response pattern at a given time t. The decoded latent variables were used to reconstruct the visual input, which was assumed to occur 4 s before the time of response (i.e. at t – 4s) (Nishimoto et al., 2011). Alternatively, we trained a linear regression model for deconvolution as elaborated in Huth et al. (2016). Briefly, the model used the voxel responses at time t+2*TR, t+3*TR and t+4*TR (TR = 2s) to predict the latent variables at time t when training and testing the decoding model. Both strategies were explored and compared in this study.

2.9. Evaluation of decoding performance

To evaluate the decoding performance in visual reconstruction, we calculated the Structural Similarity index (SSIM) (Wang et al., 2004) between every reconstructed video frame and the true video frame, yielding a measure of the similarity in the pattern of pixel intensity. The SSIM was further averaged across all video frames in the testing movie.

In addition, we evaluated the degree to which the reconstructed movie preserved the color information in the original movie. For this purpose, the color at each pixel was converted from the RGB values to a single hue value. The hue maps of the reconstructed movie frames were compared with those of the original movie frames. Their similarity was evaluated in terms of the circular correlation (Berens, 2009; Jammalamadaka and Sengupta, 2001) for all movie frames. Specifically, the hue values were represented as a vector for every frame and then were further concatenated across all frames. The circular correlation in the concatenated hue vector between the original and reconstructed frames was calculated for each subject. It was further tested for statistical significance based on the block-permutation test with a 24-sec block size and 100,000 times of permutation (Adolf et al., 2014).

As color may covary with other visual properties in natural images, it is thus likely that the decoded latent variables in VAE do not represent color per se, but other features that are more associated with certain types of color than others (e.g. sea tends to be blue, grass tends to be green). In an attempt to disentangle color vs. non-color features, we converted every testing movie frame from a color image to a gray-scale image. In the absence of any color feature, the gray-scale images were used as the input to the encoder of VAE. The encoder converted the input image into latent representations, and the decoder further reconstructed an image based on the latent representations. Then we evaluated the correlation in the hue value between the original color movie and the reconstructed movie.

We also compared the performance of the VAE-based decoding method with a previously published decoding method (Cowen et al., 2014). In that alternative method, every frame in the training movie was vectorized and then assembled altogether into a matrix. PCA was applied to this matrix to obtain its principal components (or eigen-images) that explained 99% of the variance. The partial least squares regression (PLSR) (Tenenhaus et al., 2005) was used to estimate the linear transformation from fMRI maps to eigen-images based on the fMRI data during the training movie. Through the estimated PLSR model, the fMRI data during the testing movie was converted to the corresponding representations of eigen-images, which in turn were recombined to reconstruct the visual stimuli (Cowen et al., 2014). As a variation of this

[1] https://github.com/gallantlab/motion_energy_matlab/tree/master/demo/ComputeStaticGabors.m.

PLSR-based model, we also explored the use of L_1-norm regularized optimization to estimate the linear PLSR function, in a similar way as for the VAE-based decoding model (see Eq. (6)). Using the same training procedure allowed us to test the effect of different feature models (latent variables vs. eigen-images) for fully-computable decoding of fMRI responses.

3. Results

3.1. *VAE provided compressed representations of natural images*

By design, VAE aimed to form a compressed and generalized vector representation of any natural image. In VAE, the encoder converted any natural image into 1,024 independent latent variables; the decoder reconstructed the image from the latent variables (Fig. 2A). After training it with >2 million natural images in a wide range of categories, the VAE could regenerate natural images without a significant loss in image content, structure, and color, albeit blurred details (Fig. 2B). The VAE-generated images showed comparable quality for different types of input images (Fig. 2B). As such, the latent representations in VAE were generalizable across various types of visual objects, or their combinations.

3.2. *VAE predicted movie-induced cortical responses*

Given a natural movie as visual input, we further asked to what extent the model dynamics in VAE could be used to model and predict the movie-induced cortical responses. After dimension reduction, the unit responses of VAE was represented in terms of 5,816 principal components. A linear combination of these components, as defined by a voxel-wise encoding model, was used to predict how each voxel in the brain responded to a visual stimulus. Specifically, the encoding model was trained separately for each voxel by fitting the voxel response to a training movie as a linear combination of the VAE's responses to the same movie. Then, the trained voxel-wise encoding model was tested with an independent testing movie (not used for training) to evaluate the model's prediction accuracy (i.e. the correlation between the predicted and measured fMRI responses). For a large area in the visual cortex (Fig. 3), the VAE-based encoding models could predict the movie-evoked responses with statistically significant accuracy (FDR, $q < 0.01$). In particular, early visual areas (V1/V2/V3) showed the highest prediction accuracy, whereas the prediction accuracy was relatively lower for higher visual areas along the ventral or dorsal stream (Fig. 3). The VAE-predictable areas were relatively larger when more data (~12-h movie) were used for training the encoding models in Subject 1 than in Subject 2 & 3 for whom fewer training data (2.5-h movie) were available (Fig. 3). In addition, the encoding models trained with data from Subject 1 were used to predict the fMRI responses from Subject 2 or 3. Relative to the encoding models trained and tested for the same subject (Subject 1), the accuracy of cross-subject encoding was relatively lower but still significant over a large area in the visual cortex (Fig. 3).

The encoding performance was further evaluated for individual ROIs selected from different levels of the visual hierarchy. The ROI-level analysis confirmed the results from the voxel-level analysis. As shown in Fig. 4, the encoding accuracy was highest in early visual areas (especially V1) and was progressively lower at increasingly higher-level areas. Overall, the encoding accuracy obtained with VAE had a large margin from the noise ceiling for any ROI. The encoder and the decoder part of the VAE did not differ significantly in terms of their contributions to encoding the movie-induced fMRI responses. For every ROI, the encoding accuracy obtained with either the encoder or the decoder was very similar to the encoding accuracy obtained with both the encoder and the decoder (or the VAE as a whole).

3.3. *Comparing the encoding performance across different models*

We further compared the encoding performance between VAE and two alternative models: gray-scale Gabor filters (Kay et al., 2008) and

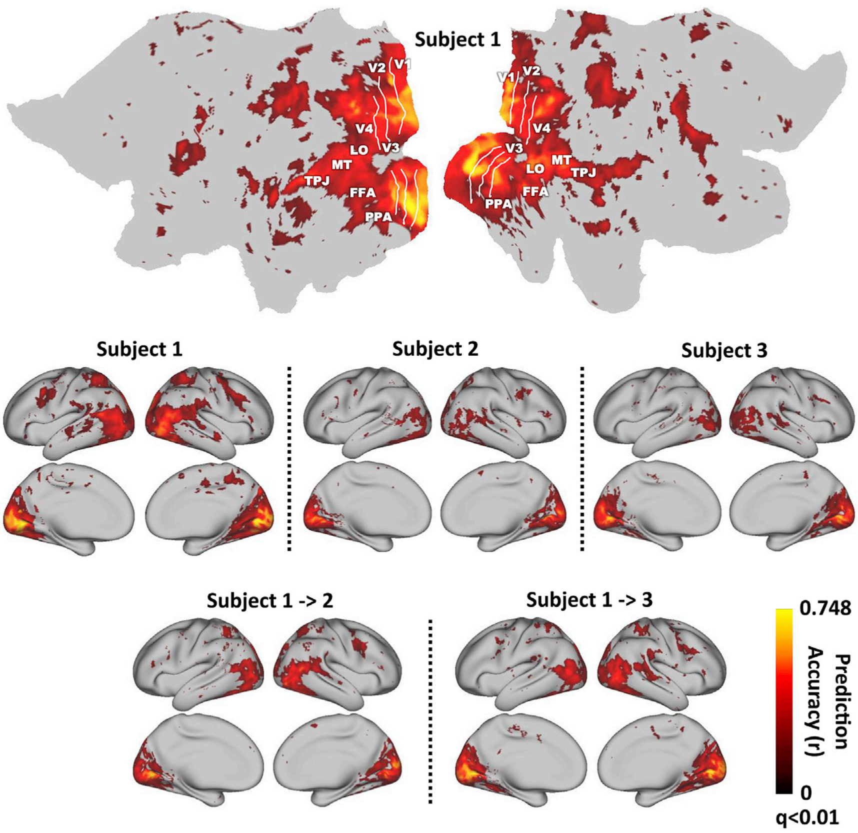

Fig. 3. Prediction accuracy with VAE-based encoding models. The accuracy was measured by the Pearson's correlation coefficient (r) between the model-predicted response and the actual fMRI response. The map shows the r value averaged across the five testing movies. The map was thresholded by statistical significance (FDR $q < 0.01$). For intra-subject encoding, the results are shown on the flattened (only for Subject 1) and inflated cortical surfaces (for every subject) as in the first and second rows. For inter-subject encoding, the results are shown in the third row.

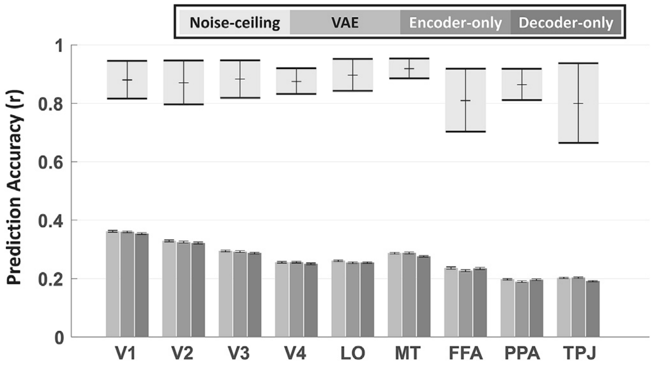

Fig. 4. The ROI-level encoding performance of VAE encoder or VAE decoder. For each feature model, the accuracy of the encoding model in predicting the fMRI response to the testing movie was summarized for each of the nine pre-defined ROIs. The models under comparison are variational autoencoder (VAE), VAE encoder and VAE decoder. Arranged from the left to the right, individual ROIs are located in increasingly higher levels of the visual hierarchy. The bar chart is based on the mean ± SEM (standard error of the mean) of the voxel-wise prediction accuracy averaged across all the voxels in each ROI, and across different testing movies and subjects. The bars in the lightest color indicate the mean and the standard derivation of the noise-ceilings in each ROI.

feedforward-only ResNet-18 (or ResNet in short). After dimension reduction, Gabor filters gave rise to 1,122 features, ResNet gave rise to 7, 564 features, whereas VAE gave rise to 5,816 features. From features to voxels, a linear regression model, defined separately for each voxel, was trained and tested in the same way for different feature models (i.e. VAE, Gabor, ResNet).

As shown in Fig. 5, the encoding accuracy obtained with VAE was higher than the accuracy obtained with the Gabor (spatial) filters for every ROI. Their difference was most significant for early visual areas (V1/V2/V3/V4) and LO/MT/FFA/TPJ in higher visual areas ($p < 0.001$, paired sample *t*-test), but relatively marginal for PPA ($p < 0.05$, paired sample *t*-test). It should be noted that the (spatial) Gabor filters used here are likely incomplete or sub-optimal. The relatively superior encoding performance of VAE to a specific set of Gabor filters should not be directly extrapolated to all variations of Gabor filters, especially for their extension to the (spatiotemporal) motion-energy filters (Nishimoto et al., 2011).

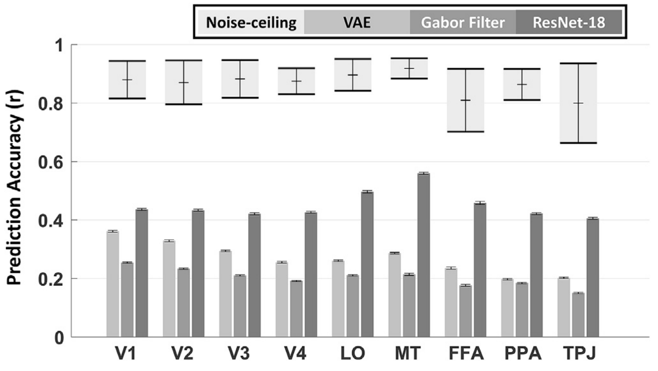

Fig. 5. | **The ROI-level encoding performance of different models.** For each feature model, the accuracy of the encoding model in predicting the fMRI response to the testing movie was summarized for each of the nine pre-defined ROIs. The compared models are: variational autoencoder (VAE), Gabor filters and ResNet-18. Arranged from the left to the right, individual ROIs are located in increasingly higher levels of the visual hierarchy. Their bar chart is based on the mean ± SEM (standard error of the mean) of the voxel-wise prediction accuracy averaged across all the voxels in each ROI, and across different testing movies and subjects. The bars in the lightest color indicate the mean and the standard derivation of the noise-ceilings at each ROI.

However, ResNet outperformed VAE by a large margin for every ROI ($p < 0.001$, paired sample *t*-test). Their difference was most pronounced at higher-level ventral areas (e.g. FFA/PPA/TPJ) but relatively less sizable at early visual areas (Fig. 5). Fig. 6 shows the comparison between VAE and ResNet in terms of their voxel-wise encoding accuracy. In line with the results from the ROI-level analysis, ResNet outperformed VAE for most of the visual cortex. Their difference (by subtraction) was much more notable in the ventral-stream areas than early visual areas or those in the dorsal-stream areas. Therefore, VAE was in general less predictive of visual cortical activity than was ResNet, which was trained with supervised learning.

We further asked whether the difference in encoding performance between VAE and CNN was due to their differences in learning objective, architecture, or dimensionality. To address this question, we defined two encoding models based on a CNN that used the same architecture as that of the encoder in VAE. This CNN, herein referred to as CNN-A, achieved 60.23% top-1 accuracy for image classification with data from ImageNet (more specifically the validation set of ILSVRC2012). CNN-A gave rise to 9,081 features after dimension reduction applied in the same way as for VAE. The encoding models based on CNN-A (without residual connections) were less predictive of fMRI responses than was ResNet. However, CNN-A still outperformed VAE for all ROIs, and the difference progressively increased from lower to higher areas (Fig. 7).

We further constrained the number of encoding parameters for CNN-A to be identical to that for VAE. As shown in Fig. 7, the encoding models constrained for the number of free parameters (referred to as CNN-AP) yielded similar encoding accuracies as those of CNN-A. Fig. 8 shows the voxel-wise difference in encoding accuracy between VAE and CNN-A/CNN-AP, which was more notable in higher-level ventral-stream areas than other visual areas. It is worth noting again that the encoding performance of the VAE as a whole or by its encoder part, was comparable, without any significant difference (Fig. 4). Since the only difference between CNN-A and the encoder of VAE was their different learning objectives (classification versus compression), the results shown in Figs. 7, 8 and Fig. 4, taken together, suggest that the different encoding performance of CNN vs. VAE was mostly attributable to their different learning objectives (supervised learning for image classification vs. unsupervised learning for image reconstruction), with less contribution from their differences in architecture or dimensionality.

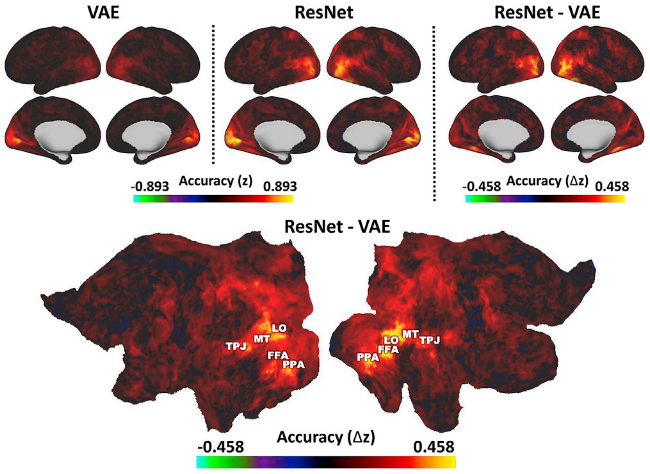

Fig. 6. Encoding performance of VAE vs. ResNet-18. The prediction accuracy (the z-transformed correlation between the predicted and measured fMRI responses) is displayed on inflated cortical surfaces for the encoding models based on VAE (top-left) and ResNet-18 (top-middle). Their difference (ResNet – VAE) in the prediction accuracy is displayed on both inflated (top-right) and flattened (bottom) cortex.

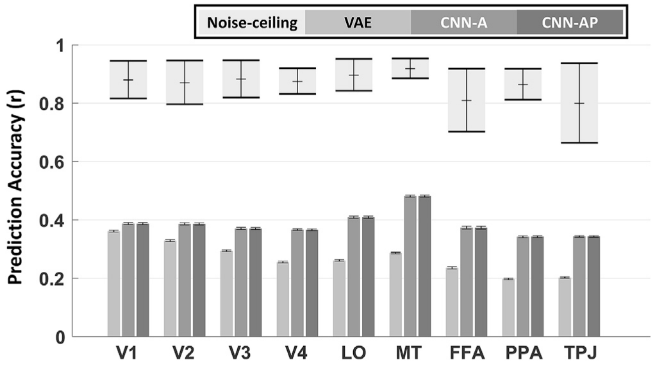

Fig. 7. The ROI-level encoding performance of VAE and constrained CNNs. For each feature model, the accuracy of the encoding model in predicting the fMRI response to the testing movie was summarized for each of the nine predefined ROIs. The models under comparison are variational autoencoder (VAE), CNN with constraint on its architecture (CNN-A) and CNN with constraint on its architecture and parameters (CNN-AP). Arranged from the left to the right, individual ROIs are located in increasingly higher levels of the visual hierarchy. The bar chart is based on the mean ± SEM (standard error of the mean) of the voxel-wise prediction accuracy averaged across all the voxels in each ROI, and across different testing movies and subjects. The bars in the lightest color indicate the mean and the standard derivation of the noise-ceilings in each ROI.

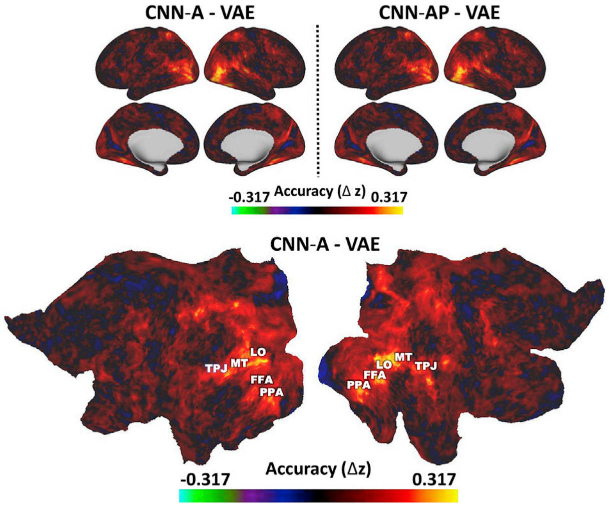

Fig. 8. Encoding performance with VAE vs. constrained CNNs. The difference in prediction accuracy (the z-transformed correlation between the predicted and measured fMRI responses) is displayed on inflated/flattened cortical surfaces. For VAE and CNN with constraint on its architecture (CNN-A), their difference (CNN-A − VAE) in the prediction accuracy is displayed on both inflated (top-left) and flattened (bottom) cortex. For VAE and CNN with constraint on its architecture and parameters (CNN-AP), their difference (CNN-AP − VAE) in the prediction accuracy is displayed on the inflated (top-right) cortex.

3.4. Direct visual reconstruction by decoding fMRI responses

We further explored the use of VAE for decoding the fMRI activity to reconstruct the visual input. For this purpose, a decoding model was trained and used to convert the fMRI activity to the VAE's latent representation, which was in turn converted to a pixel pattern through the VAE's decoder. In comparison with the original videos, Fig. 9 shows the visual input reconstructed from fMRI activity based on VAE and the decoding models, which were trained and tested with data from either the same or different subjects. Although too blurry to fully resolve details or discern visual objects, the reconstructed videos captured some important information about the original videos, including the coarse position and shape of objects, and the rough color and contrast. The quality of visual reconstruction was better when the decoding models were trained and tested for the same subject than for different subjects.

We assessed the quality of visual reconstruction by quantifying the structural similarity (as SSIM) (Wang et al., 2004) between the reconstructed and original movies. The VAE-based decoding method yielded a much higher SSIM (about 0.5) than the eigen-image-based benchmark models with either partial least squares regression (Cowen et al., 2014) or L_1-regularized linear regression (Fig. 10A, paired *t*-test, $p < 0.001$). As neurovascular coupling (modeled as HRF) caused the fMRI response to occur with a delay from the input stimulus, we shifted the decoded visual input by 4 s, as in previous studies by others (Nishimoto et al., 2011) and us (Wen et al., 2018a). In addition, we explored an alternative strategy, as in Huth et al. (2016), to estimate a deconvolutional kernel, through which the visual input at a given time was reconstructed based on the fMRI responses delayed by multiple time points. It was found that a simple time shift gave rise to a higher SSIM than did an estimated deconvolutional kernel (paired *t*-test, $p < 0.001$). Moreover, the VAE-based reconstruction preserved the color information in the movie, showing statistically significant (permutation test, $p < 0.001$) correlations in color index (hue-value) around 0.24 (Fig. 10B). The reconstructed color resulted from the information decoded from fMRI data, instead of a spurious result generated by the VAE's decoder. When we converted the original movie from color to gray, running the gray-scale movie through VAE could not reconstruct the original color movie, showing a low and non-significant correlation in hue-value ($r = -0.0664$).

4. Discussion

In this study, we trained and tested a variational autoencoder (or VAE in short) as an unsupervised model of visual perception. As established in machine learning (Kingma and Welling, 2013), VAE uses an encoder-decoder architecture to learn representations of input data without supervision. The encoder infers the “causes” of the input. The decoder uses the inferred “causes” to attempt to reconstruct the input. As such, perceptual inference could be tested to update the encoder and the decoder by learning to minimize the error of reconstruction. This notion behind VAE is conceptually similar to that of the Bayesian brain (Knill and Pouget, 2004; Yuille and Kersten, 2006), despite distinctions to be discussed later. The encoder runs bottom-up inference for recognition as does the brain's feedforward pathway, and the decoder runs top-down synthesis for prediction as does the brain's feedback pathway. Motivated by this conceptual linkage, we focus this study on evaluating VAE for predicting and decoding fMRI responses to natural video stimuli. Our results demonstrate a modest level of success, providing some arguably useful clues to guide future efforts towards finding a better model of human visual processing and learning.

In the following, we share our interpretations of the results from this study and discuss the merits and limitations of VAE as a brain model. The purpose of the discussions is to cast our perspectives with both arguments and counter-arguments from the literature and the comments raised during the peer-review of this paper.

4.1. VAE vs. free-energy principle of the brain

Mathematically, the learning objective for VAE is equivalent to minimization of the variational “free energy”, which is advocated (by some) as a principle of the brain (Friston, 2010). In this principle, the brain (or an autonomous agent alike) learns to predict and explain away whatever comes from the environment by inferring its hidden causes and suppressing the so-called “free energy” (Friston, 2009) or “prediction error” (Rao and Ballard, 1999). In the context of perception under the free-energy principle, VAE and the brain share some common characteristics: both running stochastic processing, using latent (or internal)

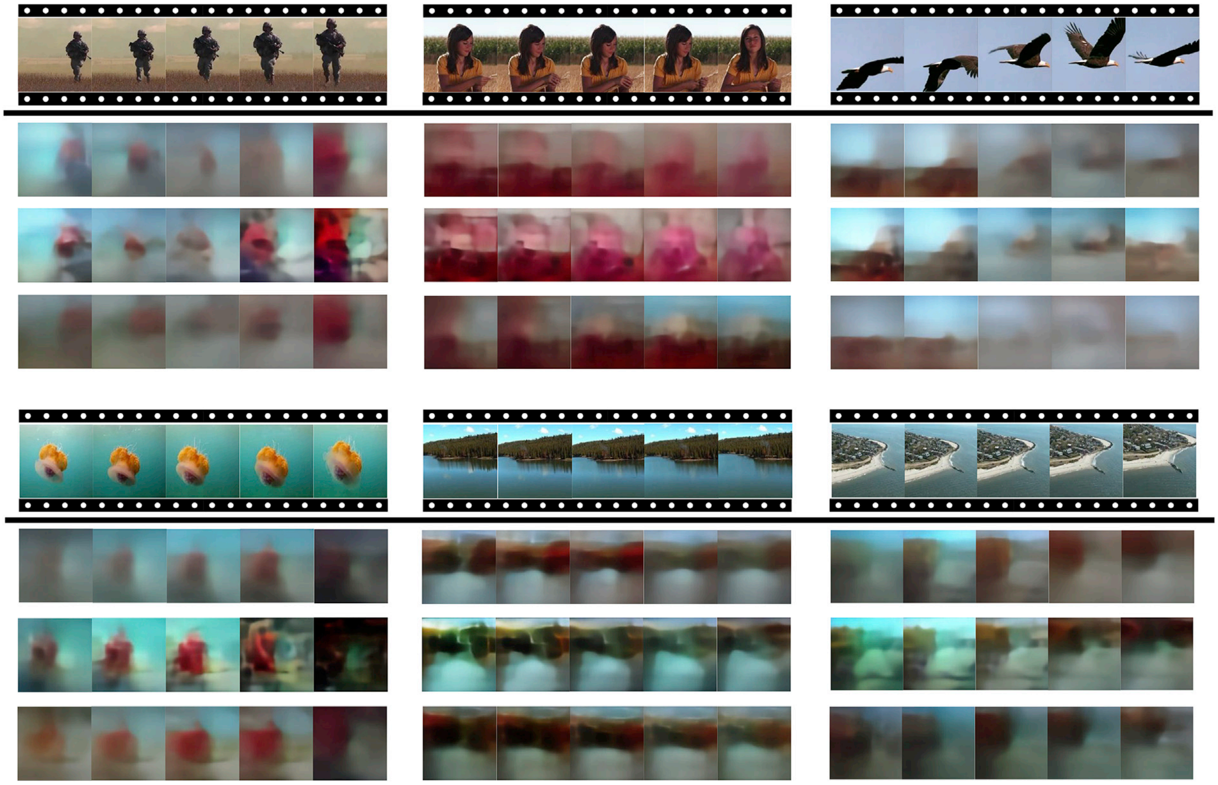

Fig. 9. Visual reconstruction based on VAE and fMRI. The original and reconstructed video frames are shown in comparison for 6 example video clips. The top row shows the original video frames. The second and third rows show the visual reconstruction based on VAE and the decoding model trained and tested within the same subject (Subject 1). For the second row, a canonical HRF is used during training and a temporal delay is used during testing to compensate for HRF. The third row shows the reconstruction based on estimating a deconvolution kernel to compensate for HRF. For the Fourth row, the VAE-based cross-subject decoding model was trained with data from Subject 1 but tested on data from Subject 2 (the top 3 clips) or Subject 3 (the bottom 3 clips).

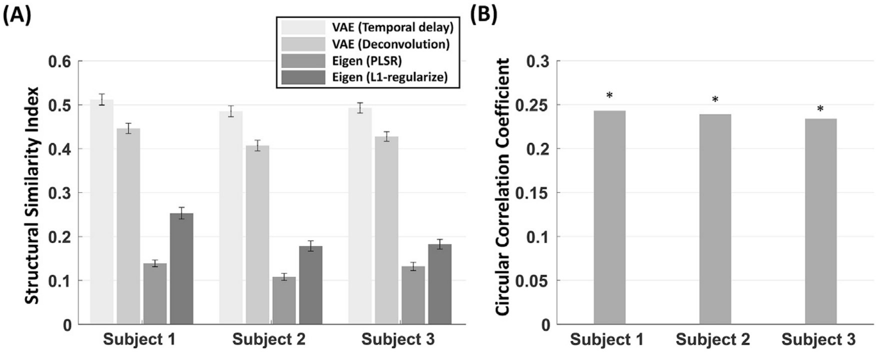

Fig. 10. Quantitative evaluation of visual reconstruction. (A) Comparison of structural similarity index (SSIM). SSIM scores of VAE-based decoding models (compensating HRF with either a temporal delay or an estimated deconvolution kernel) and eigen-image-based models were compared for all 3 subjects. Each bar shows the mean ± SE SSIM score over all frames in the testing movie. **(B) Correlation in color (hue-value)**. The (circular) correlation between the original and reconstructed hue components was calculated and evaluated for statistical significance with permutation test (*, $p < 0.001$).

states to encode the causes of sensations, mapping from sensations to latent states, mapping from latent states to sensations, and learning to avoid surprises or the difference between the external input and the internal interpretation manifested as reconstruction or prediction. For these reasons, VAE is arguably a plausible model of the brain, at least in the computational level.

However, the VAE implemented in this study and the free energy principle of the brain differ in important ways. In the brain, higher-level representations are the causes that explain or synthesize lower-level representations (Friston, 2009). The free energy should be defined for every level of the visual hierarchy (Friston, 2010), whereas VAE only minimizes the free energy with respect to the input level and a single-level latent space. Although the encoder and the decoder in VAE both contain multiple layers, they only map to/from a single level of latent variables. To be more consistent with the free energy principle, an alternative (and potentially better) model should perhaps stack multiple VAEs into hierarchically-organized latent spaces (Zhao et al., 2017).

We relate the encoder and the decoder in VAE to the forward and backward processes in the brain. However, the encoder and the decoder do not interact during computation, whereas the brain's forward and backward processes interact both within and between cortical areas (Bastos et al., 2012). The dynamic interaction has been thought (by some) to subserve the so-called "predictive coding" (Friston and Kiebel, 2009; Rao and Ballard, 1999; Spratling, 2010). The backward connections from a higher level carry the top-down prediction of lower-level representation, and the forward connections carry the error of prediction to the higher level (Rao and Ballard, 1999). As such, the forward and backward pathways convey different types of message (prediction error vs. prediction itself), which likely contribute differently to the response observable with fMRI. In contrast, the encoder and the decoder in VAE do not seem to make any differentiable contribution to the encoding of fMRI response (Fig. 4), likely due to the lack of interaction between the encoder and the decoder. Moreover, VAE does not have any mechanism for dynamic and recurrent processing, whereas video information is conveyed in both space and time. Speculatively, what might potentially address these limitation are hierarchical predictive coding networks, which are bi-directional (like VAE), dynamical (with recurrent mechanisms), and predictive forward in time (capable of temporal processing), as demonstrated in studies for computational neuroscience (Friston and Kiebel, 2009; Rao and Ballard, 1999; Spratling, 2010) and machine learning (Han et al., 2018; Lotter et al., 2016; Spratling, 2017; Wen et al., 2018c).

4.2. VAE for neural encoding

Trainable with unsupervised learning, VAE compresses images into a lower-dimensional latent space while minimizing informational loss and redundancy (Kingma and Welling, 2013). The compression through VAE is based on a cascade of nonlinear transform, unlike PCA (Dai et al., 2017; Wetzel, 2017). Since nonlinearity is central to neural information processing, it is non-trivial yet not necessarily surprising that VAE yields better decoding performance than PCA (Cowen et al., 2014), as shown in Fig. 10.

In VAE, the latent variables encode different aspects of the input data (Bouchacourt et al., 2017), and support applications of VAE for graphics transformation (Kulkarni et al., 2015), image generation (Yan et al., 2016), movement forecast (Walker et al., 2016), and image style interpolation (Deshpande et al., 2017; Yeh et al., 2016). As such, VAE is capable of learning representations beyond the scope of analytically defined Gabor filtering.

However, being more useful for computer vision does not necessarily suggest that VAE is a better model for neural encoding. One may speculate and argue that the only basis for the VAE-based encoding model is that the VAE learns Gabor-like filtering. It is known that simple cells in V1 encode features selective to orientations and (spatial and temporal) frequencies (Hubel and Wiesel, 1962). Such features, to a varying extent, can be mathematically expressed as Gabor wavelets or learned from data with independent component analysis (van Hateren and van der Schaaf, 1998), sparse coding (Olshausen and Field, 1996), or convolutional neural networks (Krizhevsky et al., 2012). It is thus possible that VAE also learns Gabor-like filters. However, it was hard to verify or reject this possibility within the scope of this study, since our VAE model used convolutional kernels of a 4-by-4 size, which was too small to manifest itself as any Gabor-like filter. Although VAE seemed to outperform a specific (likely suboptimal) set of Gabor filters in terms of neural encoding at early visual areas, their encoding performance was both relatively low and their difference was too marginal or modest to merit further analysis.

In contrast, the difference in encoding performance between VAE and CNNs was much more sizable and potentially more informative. The encoding performance of VAE was lower than ResNet (Fig. 5), or alternative CNN models with the same architecture and even the same number of encoding parameters as VAE (Fig. 7). Thus, the difference in encoding performance between VAE and CNN was primarily attributable to their learning objectives, as opposed to the network architecture. The difference was more pronounced for the ventral stream than early visual areas or the dorsal stream (Figs. 6 and 8). From these findings, we speculate that higher-level ventral areas may require supervised learning in order to extract representations for object recognition. A previous study has drawn a similar conclusion based on representational similarity analysis (Khaligh-Razavi and Kriegeskorte, 2014). However, caution should be exercised when attempting to generalize this limitation to other objectives for unsupervised learning beyond the scope of this study. Whether the brain can learn abstract representations entirely without supervision remains at least likely and awaits future studies to fully address.

4.3. VAE for neural decoding

Intuitively, a model that enabled better neural encoding should also be better for neural decoding. This is reasonable for encoding-based decoding strategies (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011), in which the goal is to find one or multiple exemplars, e.g. from a large image set, that best explain the observed response pattern through encoding models. A number of recent studies have reported superior encoding performance obtained with CNNs (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Wen et al., 2018a; Yamins et al., 2014). It is thus intuitive to anticipate that CNNs would also be a good model for brain decoding, as explored in Shen et al. (2019); Wen et al. (2018a). However, CNN might not be an optimal model for decoding, for at least two reasons. First, in the encoding model, the voxel-wise linear regression model may or may not be directly invertible, especially when the number of encoding parameters is significantly greater than the number of training samples. Second, the internal transformation of CNN may not be directly mappable onto the image space, although mapping from images to representations is always well determined.

Unlike CNN, VAE is bi-directional, allowing direct mapping from images to latent representations (via the encoder) and mapping from latent representations to images (via the decoder). From an entirely technical perspective, VAE offers a convenient and more straight-forward decoding strategy by first converting fMRI to the latent variables in VAE and then converting the latent variables to reconstructed images through the decoder in VAE. This two-step decoding is fully computable and does not require any iterative optimization or exemplar matching, which is often required for encoding-based decoding (Kay et al., 2008; Nishimoto et al., 2011; Shen et al., 2019). In addition, the VAE-based decoding does not require any computation through the encoder of the VAE. It also implies that this decoding strategy is applicable not only to images or videos being seen, but also those being imagined. However, decoding visual imagery with VAE remains speculative and awaits future studies.

In previous studies, brain responses are often related to the same set of features for both encoding and decoding (Horikawa and Kamitani, 2017; Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Shen et al., 2019). However, different features were used for encoding vs. decoding in this study. For encoding, we used the encoder, the decoder, or both together, but excluding the latent variables. Adding the latent variables into the features for neural encoding did not improve the encoding accuracy (Supplementary Fig. 1). Speculatively, the latent space as a whole encodes similar information as the top layer of the encoder, while disentangling the information into independent dimensions each represented by a latent variable. Arguably, one does not have to use the same set of features for both encoding and decoding. For encoding, it is perhaps desirable to use features that result from the bi-directional processes for both bottom-up recognition and top-down generation. For decoding or more specifically reconstructing seen or imagined images, it is perhaps desirable to only use features associated with top-down processes.

The finding that VAE was helpful for decoding color from fMRI data is intriguing. However, this finding could not be readily attributable to any "color" center in the model or in the brain. In the VAE, color was not confined to any single latent variable or a small set of latent variables. When trained with diverse natural images, VAE could not disentangle color vs. non-color (structural) features or separate the pathways for color or non-color information processing. The brain voxels that were most decodable (onto the latent variables) were distributed within early visual areas (V1/V2/V3). As such, the color and non-color features are entangled. One might argue that VAE does not decode color per se but decode some non-color features associated with color. Although this possibility could not be excluded and is indeed likely, it does not necessarily invalidate our results. The decoded color is not entirely spurious. Given the absence of any color in the input images (i.e. equal RBG values), the VAE, when it was trained with color images, did not necessarily generate spurious color in the reconstructed images. As such, VAE does extract and encode color. When it is trained with gray-scale images, VAE yielded slightly lower encoding performance (except V1), compared to the VAE trained with color images (Supplementary Fig. 2). The decrease in encoding performance was relatively more noticeable for V4, FFA, PPA, LO, than other areas. However, it should be admitted that the results obtained with the VAE, as it is in the present study, do not allow us to reveal where and how color is encoded in the brain.

Acknowledgements

The authors would like to recognize inputs from Dr. Eugenio Culurciello, Dr. Alfredo Canziani and Abhishek Chaurasia for their constructive discussions on deep neural networks. The research was in part supported by NIH R01MH104402 and the College of Engineering at Purdue University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.05.039.

References

Adolf, D., Weston, S., Baecke, S., Luchtmann, M., Bernarding, J., Kropf, S., 2014. Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. Front. Neuroinf. 8, 72.

Arcaro, M.J., McMains, S.A., Singer, B.D., Kastner, S., 2009. Retinotopic organization of human ventral visual cortex. J. Neurosci. 29, 10638–10652.

Barlow, H.B., 1989. Unsupervised learning. Neural Comput. 1, 295–311.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711.

Berens, P., 2009. CircStat: a MATLAB toolbox for circular statistics. J. Stat. Softw. 31, 1–21.

Bouchacourt, D., Tomioka, R., Nowozin, S., 2017. Multi-Level Variational Autoencoder: Learning Disentangled Representations from Grouped Observations (arxiv).

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. 6, 27755.

Cowen, A.S., Chun, M.M., Kuhl, B.A., 2014. Neural portraits of perception: reconstructing face images from evoked brain activity. Neuroimage 94, 12–22.

Dai, B., Wang, Y., Aston, J., Hua, G., Wipf, D., 2017. Hidden Talents of the Variational Autoencoder arXiv preprint arXiv:1706.05148.

David, S.V., Gallant, J.L., 2005. Predicting neuronal responses during natural vision. Netw. Comput. Neural Syst. 16, 239–260.

Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S., 1995. The helmholtz machine. Neural Comput. 7, 889–904.

Deshpande, A., Lu, J., Yeh, M.C., Chong, M.J., Forsyth, D., 2017. Learning Diverse Image Colorization. CVPR, pp. 2877–2885.

Doersch, C., 2016. Tutorial on Variational Autoencoders arXiv preprint arXiv: 1606.05908.

Du, C., Du, C., Huang, L., He, H., 2018. Reconstructing Perceived Images from Human Brain Activities with Bayesian Deep Multiview Learning. IEEE transactions on neural networks and learning systems.

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: convolutional network layers map the function of the human visual system. Neuroimage 152, 184–194.

Fogel, I., Sagi, D., 1989. Gabor filters as texture discriminator. Biol. Cybern. 61, 103–113.

Friston, K., 2009. The free-energy principle: a rough guide to the brain? Trends Cognit. Sci. 13, 293–301.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364, 1211–1221.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124.

Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D., 2015. DRAW: A Recurrent Neural Network for Image Generation arXiv preprint arXiv:1502.04623.

Guclu, U., van Gerven, M.A., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

Güçlü, U., van Gerven, M.A., 2014. Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Comput. Biol. 10, e1003724.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., van Gerven, M.A., 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. Adv. Neural Inf. Process. Syst. 4246–4257.

Han, K., Wen, H., Zhang, Y., Fu, D., Culurciello, E., Liu, Z., 2018. Deep predictive coding network with local recurrent processing for object recognition. Adv. Neural Inf. Process. Syst. 9201–9213.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

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., Sejnowski, T.J., Poggio, T.A., 1999. Unsupervised Learning: Foundations of Neural Computation. MIT press.

Hinton, G.E., Zemel, R.S., 1994. Autoencoders, minimum description length and Helmholtz free energy. Adv. Neural Inf. Process. Syst. 3–10.

Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. 8, 15037.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154.

Huth, A.G., Lee, T., Nishimoto, S., Bilenko, N.Y., Vu, A.T., Gallant, J.L., 2016. Decoding the semantic content of natural movies from human brain activity. Front. Syst. Neurosci. 10, 81.

Jammalamadaka, S.R., Sengupta, A., 2001. Topics in Circular Statistics. World Scientific.

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

Kay, K.N., Winawer, J., Mezer, A., Wandell, B.A., 2013. Compressive spatial summation in human visual cortex. J. Neurophysiol. 110, 481–494.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.

Kietzmann, T.C., McClure, P., Kriegeskorte, N., 2019. Deep Neural Networks in Computational Neuroscience. Oxford Research Encyclopedia of Neuroscience. https://doi.org/10.1093/acrefore/9780190264086.013.46.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980.

Kingma, D.P., Welling, M., 2013. Auto-encoding Variational Bayes arXiv preprint arXiv: 1312.6114.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27, 712–719.

Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Ann. Rev. Vis. Sci. 1, 417–446.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, Chicago, pp. 1097–1105.

Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.B., 2015. Deep convolutional inverse graphics network. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2, pp. 2539–2547.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lin, M., Chen, Q., Yan, S., 2013. Network in Network arXiv preprint arXiv:1312.4400.

Lotter, W., Kreiman, G., Cox, D., 2016. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning arXiv preprint arXiv:1605.08104.

Marcelja, S., 1980. Mathematical description of the responses of simple cortical cells. J. Opt. Soc. Am. 70, 1297.

Mirza, M., Courville, A., Bengio, Y., 2016. Generalizable Features from Unsupervised Learning arXiv preprint arXiv:1612.03809.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–410.

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, 902–915.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646.

Olshausen, Bruno A., Field, David J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607.

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Salin, P.A., Bullier, J., 1995. Corticocortical connections in the visual system: structure and function. Physiol. Rev. 75, 107–154.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., van Gerven, M., 2018a. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. Neuroimage 180, 253–266.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., Van Gerven, M., 2018b. Generative adversarial networks for reconstructing natural images from brain activity. Neuroimage 181, 775–785.

Seung, H.S., Lee, D.D., 2000. The manifold ways of perception. Science 290, 2268–2269.

Shen, G., Horikawa, T., Majima, K., Kamitani, Y., 2019. Deep image reconstruction from human brain activity. PLoS Comput. Biol. 15, e1006633.

Shi, J., Wen, H., Zhang, Y., Han, K., Liu, Z., 2017. Deep Recurrent Neural Network Reveals a Hierarchy of Process Memory during Dynamic Natural Vision. Hum. Brain Mapp. 39, 2269–2282.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.

Spratling, M.W., 2010. Predictive coding as a model of response properties in cortical area V1. J. Neurosci. 30, 3531–3543.

Spratling, M.W., 2017. A hierarchical predictive coding model of object recognition in natural images. Cogn. Comput. 9, 151–167.

Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., Lauro, C., 2005. PLS path modeling. Comput. Stat. Data Anal. 48, 159–205.

van Gerven, M.A., de Lange, F.P., Heskes, T., 2010. Neural decoding with hierarchical generative models. Neural Comput. 22, 3127–3142.

van Hateren, J.H., van der Schaaf, A., 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. Proc. Biol. Sci. 265, 359–366.

Walker, J., Doersch, C., Gupta, A., Hebert, M., 2016. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. European Conference on Computer Vision, pp. 835–851.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., Liu, Z., 2018a. Neural encoding and decoding with deep learning for dynamic natural vision. Cerebr. Cortex 28, 4136–4160.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., Liu, Z., 2018b. Deep Predictive Coding Network for Object Recognition. International Conference on Machine Learning, pp. 5263–5272.

Wen, H., Shi, J., Chen, W., Liu, Z., 2018c. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Sci. Rep. 8, 3752.

Wetzel, S.J., 2017. Unsupervised learning of phase transitions: from principal component analysis to variational autoencoders. Phys. Rev. 96, 022140.

Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. 29, 477–505.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. 111, 8619–8624.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.

Yan, X., Yang, J., Sohn, K., Lee, H., 2016. Attribute2Image: conditional image generation from visual attributes. Euro. Conf. Comp. Vision 776–791.

Yeh, R., Liu, Z., Dan, B.G., Agarwala, A., 2016. Semantic Facial Expression Editing Using Autoencoded Flow (arxiv).

Yuille, A., Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? Trends Cognit. Sci. 10, 301–308.

Zhao, S., Song, J., Ermon, S., 2017. Learning Hierarchical Features from Generative Models arXiv preprint arXiv:1702.08396.