PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

End-to-end neural system identification with neural information flow

K. Seeliger[1,2☯]*, L. Ambrogioni[1☯], Y. Güçlütürk[1], L. M. van den Bulk[1], U. Güçlü[1], M. A. J. van Gerven[1]*

1 Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands, **2** Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

☯ These authors contributed equally to this work.
* kseeliger@posteo.jp (KS); m.vangerven@donders.ru.nl (MvG)

OPEN ACCESS

Citation: Seeliger K, Ambrogioni L, Güçlütürk Y, van den Bulk LM, Güçlü U, van Gerven M A.J (2021) End-to-end neural system identification with neural information flow. PLoS Comput Biol 17(2): e1008558. https://doi.org/10.1371/journal.pcbi.1008558

Editor: Samuel J. Gershman, Harvard University, UNITED STATES

Received: October 11, 2019

Accepted: November 24, 2020

Published: February 4, 2021

Data Availability Statement: The functional MRI data is available on Donders Repository (data.donders.ru.nl, di.dcc.DSC_2018.00082_134, https://doi.org/10.34973/j05g-fr58) and has been described in detail in a separate data description manuscript (Seeliger & Sommers 2019; doi: https://doi.org/10.1101/687681). For data access researchers have to agree to our local data use agreement (DUA) for human participants.

Funding: This research was supported by VIDI grant number 639.072.513 of The Netherlands

Abstract

Neural information flow (NIF) provides a novel approach for system identification in neuroscience. It models the neural computations in multiple brain regions and can be trained end-to-end via stochastic gradient descent from noninvasive data. NIF models represent neural information processing via a network of coupled tensors, each encoding the representation of the sensory input contained in a brain region. The elements of these tensors can be interpreted as cortical columns whose activity encodes the presence of a specific feature in a spatiotemporal location. Each tensor is coupled to the measured data specific to a brain region via low-rank observation models that can be decomposed into the spatial, temporal and feature receptive fields of a localized neuronal population. Both these observation models and the convolutional weights defining the information processing within regions are learned end-to-end by predicting the neural signal during sensory stimulation. We trained a NIF model on the activity of early visual areas using a large-scale fMRI dataset recorded in a single participant. We show that we can recover plausible visual representations and population receptive fields that are consistent with empirical findings.

Author summary

We propose a method for data-driven estimation of computational models, representing neural information processing between different cortical areas. We demonstrate this method on the largest single-participant naturalistic fMRI dataset recorded to date. By training a simplified model of the visual system we show that biologically plausible computations emerge in the training process, yielding a new approach to understanding information processing in neural systems. The approach is applicable to other sensory or imaging modalities, thus providing a general way to computational modeling in cognitive neuroscience.

This is a *PLOS Computational Biology* Methods paper.

Organization for Scientific Research (NWO, https://www.nwo.nl; (M. A. J. v. G.)). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

Uncovering the nature of neural computations is a major goal in neuroscience [1]. It may be argued that true understanding of the brain requires the development of *in silico* models that explain the activity of biological neurons in terms of information processing. We refer to this idea as *neural system identification* [2, 3]. In cognitive terms, information processing can be understood as using internal representations of environments with the goal of generating behaviour.

The predominant approach for uncovering these representations is to use predefined non-linear features derived from the stimulus as a hypothesis for predicting measured neural responses [4–6]. Using this approach, in visual and auditory domains the best results so far have been obtained by using convolutional (or deep) neural networks (DNNs) [6–15]. DNNs process input through a sequence of layers with linear and nonlinear transformations, and learn local features and maps of these features through the convolution operation. Each layer of a DNN encodes increasingly more complex abstractions of the original input. However, using this approach DNNs have to be trained for solving manually defined tasks such as object classification on specific data bases. Consequently, the resulting DNN feature representations are biased towards their specific objective function.

An alternative approach is to directly estimate hierarchical representations from neural data. This idea has been used to reveal mechanisms of neural information processing in biological systems [13, 16–24]. However, most of these ideas have been applied within individual brain regions (most frequently within V1) and using invasive data. In the area of human visual perception across multiple areas, the most related approach is *Representational Distance Learning* [25, 26], which uses representational dissimilarity matrices estimated within visual areas as an element of the training objective of a convolutional neural network modeling these areas. Recent approaches use the prediction of neural measurements directly for learning to separate the location and features that voxels respond to [13, 17, 18, 24, 27, 28]. This manuscript expands on this work, proposing a novel approach for neural system identification, referred to as *neural information flow* (NIF). NIF generalizes existing approaches, allowing estimating neural information processing systems from individual cortical areas up to the whole-brain level.

Similar to DNN encoding models, the information processing hierarchy is expressed as a multi-layer neural network. However, the layers of NIF models have a one-to-one correspondence to biological neural populations (such as V1), and all neural network parameters are solely trained with the objective function of predicting brain activity measured in response to input stimuli. Using this method, training is expected to learn spatiotemporal neural representations of the sensory input inside the corresponding population, and learn to derive the underlying flow of information processing. In neurobiological terms, DNN nodes can be interpreted as the activation of a cortical column responsive to a specific local feature, such as a Gabor wavelet in V1. The cascade of convolutional layers can be interpreted as the topologically organized connectivity between brain regions.

Convolutional layer activity is linked to neural measurement units through unit-wise observation models that are trained jointly with the other network parameters. The choice of measurement unit (e.g. cellular, voxels, behavioural) in the NIF framework is arbitrary, and measurements can be combined. In case of functional magnetic resonance imaging (fMRI) from a visual experiment, each voxel learns its spatial receptive field and local peak of the hemodynamic response; and the preferred convolutional features (channels) of its underlying information processing units.

In this manuscript we outline the principles and methodology of NIF with a simplified model of the visual system. Using a large fMRI dataset acquired under stimulation with

naturalistic video we demonstrate that the model is capable of generating realistic brain measurements, and that the computations learned inside the model are biologically meaningful. We expect that these ideas will guide the development of a new family of computational models that allow uncovering the principles of neural computations in biological systems.

Methods

Ethics statement

Data collection was approved by the local ethical review board (CMO regio Arnhem-Nijmegen, The Netherlands, CMO code 2014-288 with amendment NL45659.091.14) and was carried out in accordance with the approved guidelines. For each session written formal consent was obtained from the participant. All specifics of the data set are described in a separate manuscript accompanying the data publication [29].

Neural information flow

The purpose of a NIF model is to capture the neural computations that take place within and between neuronal populations in response to sensory input. The general philosophy of NIF is outlined in Fig 1. The core of a NIF model is a deep modular neural network architecture where individual neuronal populations are modeled using neural network modules that transform afferent input into efferent output. The connectivity between populations is captured by convolutional layers which model the topographically organized information exchange between neuronal populations. Finally, population activity is used to predict observed measurements through factorized observation models. Model parameters are estimated by fitting the neural signals measured during sensory stimulation. Specifically, the NIF model receives the same sensory input that is presented to the participant and predicts the measurements of all brain regions of interest. Model components are trained end-to-end using stochastic

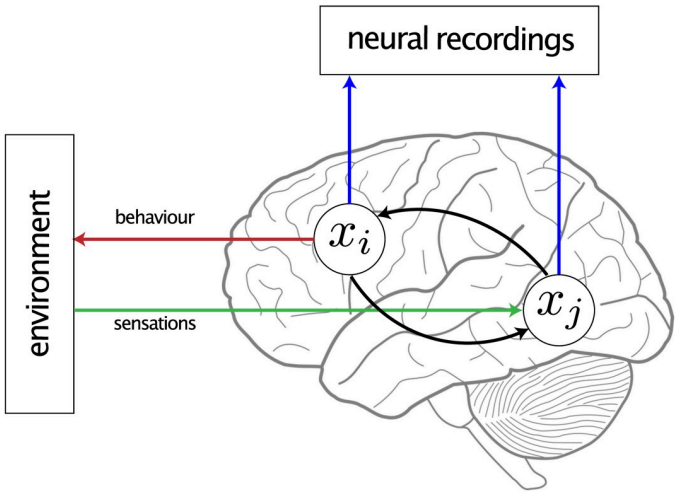

Fig 1. The philosophy underlying neural information flow. NIF models define synthetic brains that model information processing in real brains. They are specified in terms of mutually interacting neuronal populations (white discs) that receive sensory input (green) and give rise to measurements of neural activity (blue) and/or behavior (red). In practice, NIF models may consist of up to hundreds such interacting populations. They can be estimated by fitting them to neurobehavioral data acquired under these tasks. By analyzing NIF models, we can gain a mechanistic understanding of neural information processing in real brains and how neural information processing relates to phenomenology.

https://doi.org/10.1371/journal.pcbi.1008558.g001

gradient descent to minimize the error in voxel-specific measurement predictions. In the following we describe the NIF components in more detail.

Modeling sensory input and neural representations. Sensory input is modeled using a four-dimensional tensor $\mathbf{N} \in \mathbb{R}^{N_c \times N_t \times N_x \times N_y}$ whose array dimensions represent input channels c, time t and spatial coordinates (x, y) respectively. For example, the input channels can be the RGB components of a visual stimulus or the photoreceptor responses of a retinal model. In our experiments, we model grayscale images using a single luminance channel ($N_c = 1$). We used temporal windows of 2.1s, resulting in 48 frames ($N_t = 48$). Analogously, the representations of the sensory input encoded in each brain region are modeled using four-dimensional tensors. The feature maps $\mathbf{N}[c, :, :, :]$ of these neural tensors encode neural processing of specific sensory stimulus features such as oriented edges or coherent motion. Consequently, a tensor element can be interpreted as the response of one cortical column. Under the same interpretation, cortical hyper-columns are represented by a sub-tensor $\mathbf{N}[:, :, x, y]$ storing the activations of all the columns that respond to the same spatial location.

Modeling directed connectivity and information flow. We model the directed connectivity between brain regions using spatiotemporal convolutions. The spatial weights model the topographically organized synaptic connections while the temporal component models synaptic delays. Using this setup, we can model how neural populations respond to sensory input as well as to each other. Note that to enforce causality of the neural responses, the temporal filters should be causal, meaning that the only non-zero weights correspond to past time points. However, this assumption can be dropped when the time scale of our observations is much slower than that of the underlying temporal dynamics (as in BOLD data).

Let $\tilde{\mathbf{N}}$ denote the concatenation of afferent inputs $\mathbf{N}_1, \ldots, \mathbf{N}_N$ along the feature dimension and let $\star$ denote the convolution operation. We define the activation of the j-th brain area as a function of its afferent input as follows:

$$\mathbf{N}_j = \mathbf{f}_j(\mathbf{N}_1, \ldots, \mathbf{N}_N) = \mathbf{f}(\tilde{\mathbf{N}} \star \mathbf{W}_j + \mathbf{B}_j)\,, \tag{1}$$

where $\mathbf{f}$ is the element-wise application of a sigmoid activation function followed by downsampling using an average pooling operation, $\mathbf{W}_j$ is a synaptic weight kernel and $\mathbf{B}_j$ is a bias term. Initial testing indicated more stable convergence using sigmoid activation functions compared to ReLU activation functions.

Modeling observable signals

NIF models are estimated by linking neural tensors to observation models that capture indirect measurements of brain activity. Observations are represented using tensors $\mathbf{Y}$ that store measurable responses. The observation model expresses the predicted measurements as a function of the activity of the latent tensors:

$$\mathbf{Y} = \mathbf{g}(\mathbf{N}_1, \ldots, \mathbf{N}_N) + \epsilon\,, \tag{2}$$

where ϵ is measurement noise. The exact form of $\mathbf{g}$ depends on the kinds of measurements that are being made. Neuroimaging methods such as fMRI, single- and multi-unit recordings, local field potentials, calcium imaging, EEG, MEG but also motor responses and eye movements are observable responses to afferent input and can thus be used as a training signal. Note that the same brain regions can be observed using multiple observation models, conditioning them on multiple heterogeneous datasets at the same time. This provides a solution for multimodal data fusion in neuroscience [30]. In this paper, we focus on modeling blood-oxygenation-level dependent (BOLD) responses obtained for individual voxels using fMRI. In this case, we can consider the voxel responses separately for each region, such that we have $\mathbf{Y}_i = \mathbf{g}_i(\mathbf{N}_i) + \epsilon$ for

each region i. Let $\mathbf{Y}_i \in \mathbb{R}^{K\times T}$ denote BOLD responses of K voxels acquired over T time points for the i-th region. Our observation model for the kth voxel in that region is defined as

$$\mathbf{Y}_i[k, t+\Delta_t] = b_k + \sum_{c,\tau,x,y} \mathbf{N}_i[c,\tau,x,y]\mathbf{U}_i[c,\tau,x,y,k] + \epsilon[k]\,, \tag{3}$$

where $\mathbf{N}_i$ contains neural network activations to the stimulus frames presented in preceding video chunks, relative to time t, b_k is a voxel-specific bias, $\epsilon[k]$ is normally distributed measurement noise and Δ_t is a temporal shift of the BOLD response that is used to take into account a default offset in the hemodynamic delay (4.9 s in our experiments). Every brain region can be observed using a function of the form shown in Eq (3).

Factorized observation models. To simplify parameter estimation and facilitate model interpretability we use a factorized representation of $\mathbf{U}$ (also see [17]. That is,

$$\mathbf{U}[c,t,x,y,k] = \mathbf{U}_c[c,k]\mathbf{U}_t[t,k]\mathbf{U}_s[x,y,k]\,, \tag{4}$$

where k is denotes the voxel index. Here, $\mathbf{U}_c[\cdot,k]$ are the feature loadings that capture the sensitivity of a voxel to specific input features, $\mathbf{U}_t[\cdot,k]$ is the temporal profile of the observed BOLD response of a voxel and $\mathbf{U}_s[\cdot,\cdot,k]$ is the spatial receptive field of a voxel. Hence, the estimated voxel-specific observation models have a direct biophysical interpretation.

We further facilitate parameter estimation by using a spatial weighted low-rank decomposition of the spatial receptive field:

$$\mathbf{U}_s[x,y,k] \approx \sum_{r=1}^{R} a_{k,r}\mathbf{U}_{x,r}[x,k]\mathbf{U}_{y,r}[y,k]\ . \tag{5}$$

Here, $a_{k,r}$ are rank amplitudes that are constrained to be positive using a softplus transformation. We used $R = 4$ in our experiments. The rank limits the complexity of the spatial observation model. Rank one models can estimate unimodal receptive fields. However, a small number of voxels have nonclassical receptive fields that respond to multiple parts of the input space, for which more degrees of freedom are needed. To further stabilize the model and obtain localized and positive spatiotemporal receptive fields, we apply a softmax nonlinearity to the columns of $\mathbf{U}_t$, $\mathbf{U}_x$ and $\mathbf{U}_y$. That is, the elements u_i of each column vector $\mathbf{u}$ of these matrices are given by

$$u_i = \sigma_i(\mathbf{v}) = \exp(v_i)/\sum_j \exp(v_j)\ , \tag{6}$$

where the v_i are learnable parameters.

Model estimation. Once the architecture of the NIF model is defined, synaptic weights and observation model parameters can be estimated by minimizing a loss using gradient descent via backpropagation. Let $\mathbf{Y}_i^t$ and $\hat{\mathbf{Y}}_i^t = \mathbf{g}_i(\mathbf{N}_i)$ denote the observed and predicted measurements for the ith region relative to the t measurement (BOLD volume). The loss is given by the squared error per region summed over regions and across measurements:

$$\mathcal{L} = \sum_{t,i} (\hat{\mathbf{Y}}_i^t - \mathbf{Y}_i^t)^2\,. \tag{7}$$

Note that, since the model couples neuronal populations, region-specific estimates are constrained by one another and consequently make use of all observed data. Our approach was implemented in the `Chainer` framework for automatic differentiation [31].

Experimental validation

To demonstrate the capabilities of the NIF framework, we estimated and tested a simple visual system model using a unique large-scale functional MRI dataset collected while one participant was exposed to almost 23 hours of complex naturalistic spatiotemporal stimuli. Specifically, we presented episodes from the BBC series *Doctor Who* [32].

Stimulus material. A single human participant (male, age 27.5) watched 30 episodes from seasons 2 to 4 of the 2005 relaunch of *Doctor Who*. This comprised the training set which was used for model estimation. Episodes were split into 12 min chunks (with each last one having varying length) and presented with a short break after every two runs. The participant additionally watched repeated presentations of the short movies *Pond Life* (five movies of 1 min, 26 repetitions) and *Space / Time* (two movies of 3 min, 22 repetitions), in random permutations and after most episodes. They were taken from the series' next iteration to avoid overlap with the training data. This comprised the test set which was used for model validation.

Data acquisition. We collected 3T whole-brain fMRI data. It was made sure that the training stimulus material was novel to the participant. Data were collected inside a Siemens 3T MAGNETOM Prisma system using a 32-channel head coil (Siemens, Erlangen, Germany). A T2*-weighted echo planar imaging pulse sequence was used for rapid data acquisition of whole-brain volumes (64 transversal slices with a voxel size of $2.4 \times 2.4 \times 2.4$ mm^3 collected using a TR of 700 ms). We used a multiband-multi-echo protocol with multiband acceleration factor of 8, TE of 39 ms and a flip angle of 75 degrees. The video episodes were presented on a rear-projection screen with the `Presentation` software package, cropped to 696×732 pixels squares so that they covered approximately 20 degrees of the vertical and horizontal visual field. The participant's head position was stabilized within and across sessions by using a custom-made MRI-compatible headcast, along with further measures such as extensive scanner training. The participant had to fixate on a fixation cross in the center of the video. At the beginning of every break and after every test set video a black screen was shown for 14 s to record the fadeout of the BOLD signal after video presentation stopped. The black screen stimuli of these periods were omitted in the present analysis. In total this leaves us with approximately 118.000 whole-brain volumes of single-presentation data, forming our training set (used for model estimation) and 1.032 volumes of resampled data, forming our test set (used for model evaluation). We decided to use the whole test set, including the second half with the slight vertical elongation.

Data preprocessing. Minimal BOLD data preprocessing was performed using `FSL v5.0`. Volumes were first aligned within each 12 min run to their center volume (run-specific reference volume). Next, all run-specific reference volumes were aligned to the center volume of the first run (global reference volume). The run-specific transformations were applied to all volumes to align them with the global reference volume. The signal of every voxel used in the model was linearly detrended, then standardized (demeaning, unit variance) per run. Test set BOLD data was averaged over repetitions to increase signal to noise ratio, and as a final step the result was standardized again. A fixed delay of 7 TRs (4.9 s) was used to associate stimulus video segments with responses and allow the model to learn voxel-specific HRF delays within $\mathbf{U}_t$. With the video segments covering 3 TRs starting from the fixed delay, the BOLD signal corresponding to a stimulus is thus expected to occur within a time window of 4.9 s to 7.0 s after the onset of the segment. As there were small differences between frame rates in the train and test sets we transcoded the stimulus videos to a uniform frame rate of 22.86 Hz (16 frames per TR) for training the example model. To reduce model complexity we downsampled the videos to 112×112. As the model operates on three consecutive TRs, the training input size

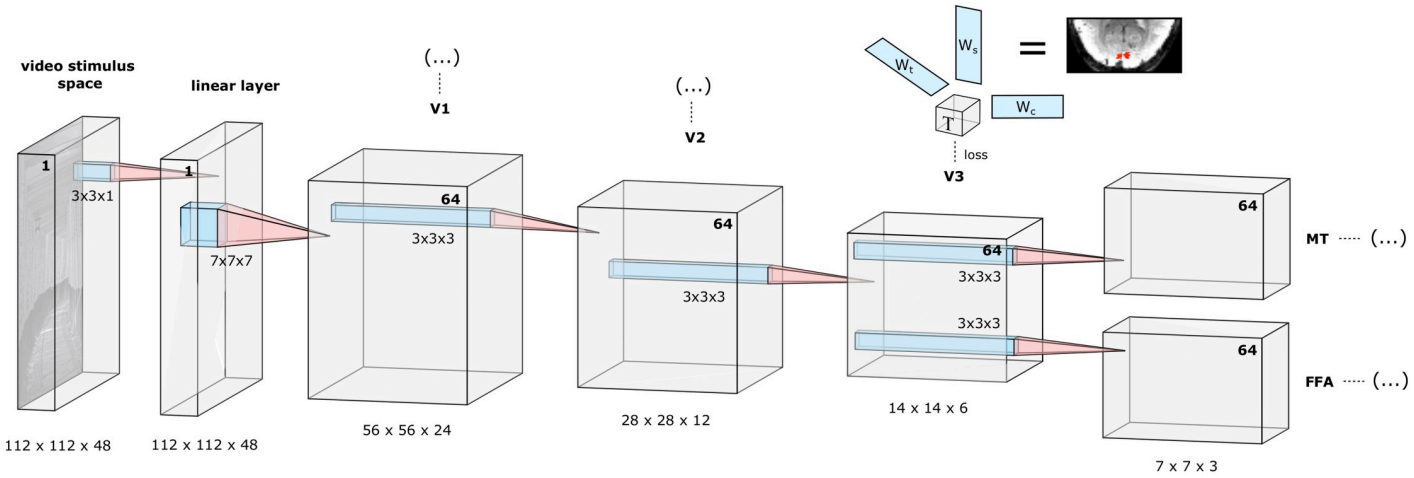

Fig 2. The described NIF architecture, a simplified feed-forward model of early visual areas. Underneath the tensors resulting from the 3D convolution operations we state the size of each input space ($x \times y \times t$) to the next layer. The number of feature maps in each input space is printed in boldface, with the stimulus (input) space consisting of a single channel. The input to the network are 3D stimulus video segments consisting of 3 × 16 frames (covering three TRs of 700 ms each), aligned with the hemodynamic response by applying a fixed delay of 7 TRs. The first convolutional layer is not attached to a region observation model, but is a single-channel linear spatial convolution layer. It serves as a learnable linear preprocessing step that accounts for retinal and LGN transformations. Convolutional kernel sizes are 7 × 7 × 7 in the second convolutional layer (leading to the V1 tensor), and 3 × 3 × 3 for all other layers. After every convolution operation (except for the linear layer) we apply a sigmoid nonlinearity and spatio-temporal average pooling with 2 × 2 × 2 kernels. Before entering the $\mathbf{U}_t$ observation models the temporal dimension is average pooled so that each point t covers one TR. All weights in this model (colored blue) are learned by backpropagating the mean squared error losses from predicting the BOLD activity of the observed voxels. The voxel-specific observation models consisting of the spatiotemporal weight vectors $\mathbf{U}_s$ and $\mathbf{U}_t$ and the feature observation model $\mathbf{U}_c$ enable the end-to-end training of the model from observational data.

https://doi.org/10.1371/journal.pcbi.1008558.g002

was 112 × 112 × 48. The stimuli were converted to grayscale [33] prior to presenting them to the model. Otherwise stimuli were left just as they were presented in the experiment.

Model architecture. We implemented a purely feed-forward architecture for modeling parts of the visual system (V1, V2, V3, FFA and MT). The used architecture is illustrated in detail in Fig 2. FFA and MT have their own tensors originating from V3 to allow for a simplified model of the interactions between upstream and downstream areas. We intentionally used a simplified model to focus on demonstrating the capabilities of the NIF framework. To model LGN output, we used a linear layer consisting of a single 3 × 3 × 1 spatial convolutional kernel. The NIF model was trained for 11 epochs with a batch size of 3, using the Adam optimizer [34] with learning rate $\alpha = 5 \times 10^{-4}$. Weights were initialized with Gaussian distributions scaled by the number of feature maps in every layer [35].

Results

In this paper we focus on the processing of visual information. In the following, we show that a NIF model uncovers meaningful characteristics of the visual system.

Accuracy of response predictions

After training the NIF model, we tested its accuracy on the test set. We observed that BOLD responses in a majority of voxels in each brain region could be predicted by the model (tested for significance with $p < 0.01$, Bonferroni-corrected over the total number of gray matter voxels). This is illustrated in Fig 3, showing voxel-wise correlations between predicted and observed test data per region. The results show that the NIF model generates realistic brain activity in response to unseen input stimuli. The larger correlations in area MT could be

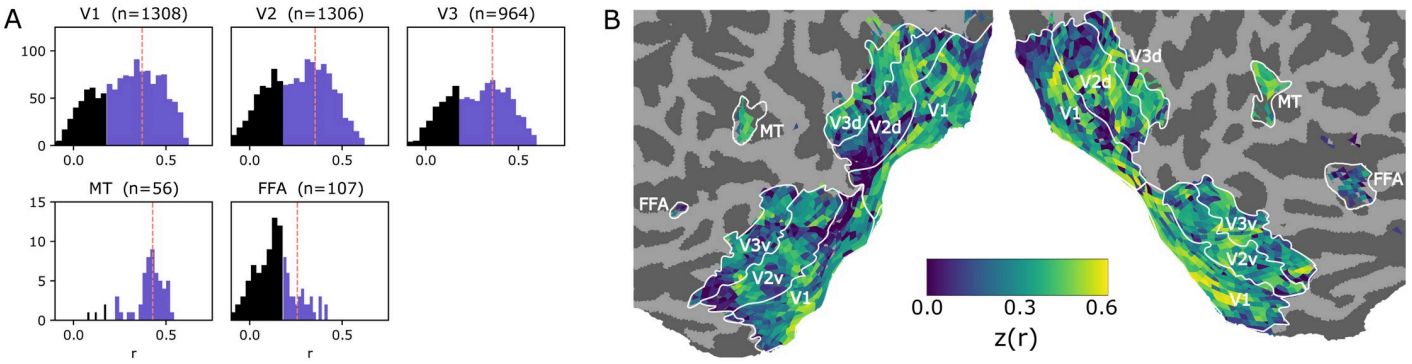

Fig 3. Voxel-wise correlations. A. Histograms of voxel-wise correlations between predicted and observed BOLD responses on the test set in different observed brain regions. The vertical line marks the median. The blue area shows the significantly predicted voxels. B. Cortical flatmap of the distribution of all correlations across the visual system. For the map we applied a Fisher z-transform to facilitate linear visual comparison of correlation magnitudes.

https://doi.org/10.1371/journal.pcbi.1008558.g003

explained by its motion-sensitivity, which can be strongly driven by the employed video stimulus and can be modeled well using a relatively straightforward motion energy model [36].

Visualization of learned representations

In this subsection we examine the features of the external stimulus that are encoded in our trained model of the visual system. We will begin with an analysis of the first layers, LGN and V1, whose features can be visualized by plotting the weights of the convolutional kernels. We will then show visualization of higher order regions using a more sophisticated preferred input analysis.

Linear feature analysis. For the first layers of the model, before the application of nonlinear transformations, neural network features can be inspected by visualizing the learned weights. A linear single-channel spatial layer was used to represent the transformation of the visual input at the retinal/LGN stage, before it enters the visual cortex [37, 38]. Fig 4A shows the estimated kernel as well as the resulting image transformation when applying this kernel to the input. As we can see, the linear kernel learns to extract edges at different orientations, as well as (albeit weaker) luminance. The result is strikingly similar to that of analytical ZCA whitening, however emphasizes edges further. When learning two linear kernels instead of one (as in our model), one kernel learns to extract luminance while the other extracts edges. This is likely to be a reflection of the independence of luminance and contrast information in natural images and in LGN responses [39]. We can also visualize the feature detectors that determine the responses of V1. Fig 4B shows the 64 channels learned by the neural tensor connected to V1 voxels. Several well-known feature detection mechanisms of V1 arise, such as Gabor-like response profiles [40]. As shown in Fig 4C, several of these feature detectors also show distinct dynamic temporal profiles, reflecting the processing of visual motion [16].

Preferred input analysis. Feature sensitivities in DNNs can only be investigated by directly plotting the learned weights before non-linearities are applied. For higher order regions, neural network interpretability methods need to be used. For instance, we can gain insight into the nature of the representations of higher order regions by visualizing which stimulus properties best drive simulated neural responses in a particular brain region. To this end, we estimated the gradient that leads to an increase in activity in individual target voxels, and used this gradient to modify the input such as to optimally drive the voxel response, starting from a three-dimensional white noise input. The technique is similar to [41], and similar in spirit to [42–44]. The basic approach was originally proposed in [45].

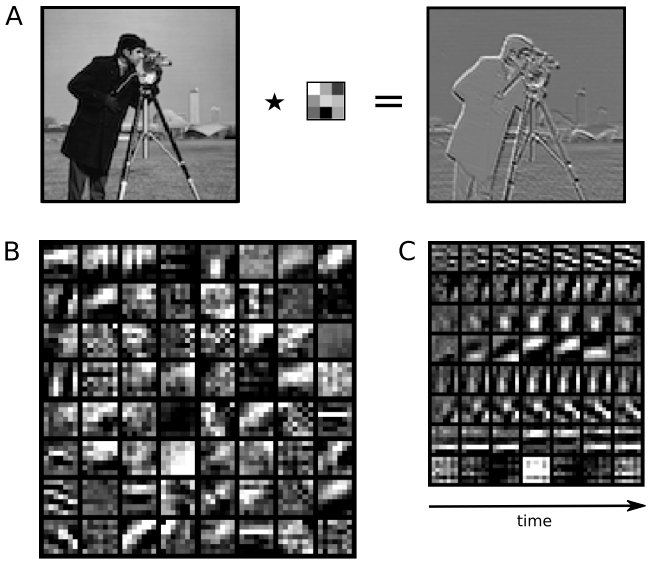

Fig 4. Stimulus features derived by the NIF model. A. Learned linear preprocessing showing that the estimated kernel extracts edges from the original input image. B. The 64 spatial features estimated from neural data for area V1 (frame three out of seven). C. Visualization of seven of these features across the temporal dimension. For visualization, feature weights were clipped at the extremes and all weights were globally rescaled between zero and one. See S1 Video for the animated version.

https://doi.org/10.1371/journal.pcbi.1008558.g004

Let $I_{t,x,y}$ denote the pixel intensity for the tth frame at spatial location (x, y). The size of I matches the input dimension of $48 \times 112 \times 112$ and is initialized with random values in the same range as the original input.

The analysis was performed only for those voxels for which the correlation between predicted and observed responses exceeded 0.4 on the test set. Let $\mathbf{y} = (y_1, \ldots, y_K)$ such that $\mathbf{y}$ denotes the activity of all voxels in a specific ROI and y_k denote the response of the kth target voxel (the voxel that's activity should be maximized). The objective is to optimize

$$\sigma_k(\mathbf{y}) = \frac{\exp(y_k)}{\sum_i \exp(y_i)} \ . \tag{8}$$

and

$$\gamma_k(\mathbf{y}) = y_k \ . \tag{9}$$

That is, we modify the input such as to maximize the activity of the kth voxel y_k, while suppressing the responses of all other voxels in the same ROI y_v using a softmax nonlinearity. This leads to an high amplitude both in absolute value and relative to the other voxels within a ROI.

We further regularize the input using an ℓ_1 loss on all components (pixel values) of I. The ℓ_1 leads to the suppression of noise in the image, which otherwise easily occurs in this optimization process.

The objective is thus to minimize

$$-\log(\sigma_k(\mathbf{y})) - \gamma_k(\mathbf{y}) + \lambda \ell_1 \ , \tag{10}$$

with $\lambda = 10^{-7}$ for FFA and MT and $\lambda = 10^{-6}$ in other ROIs.

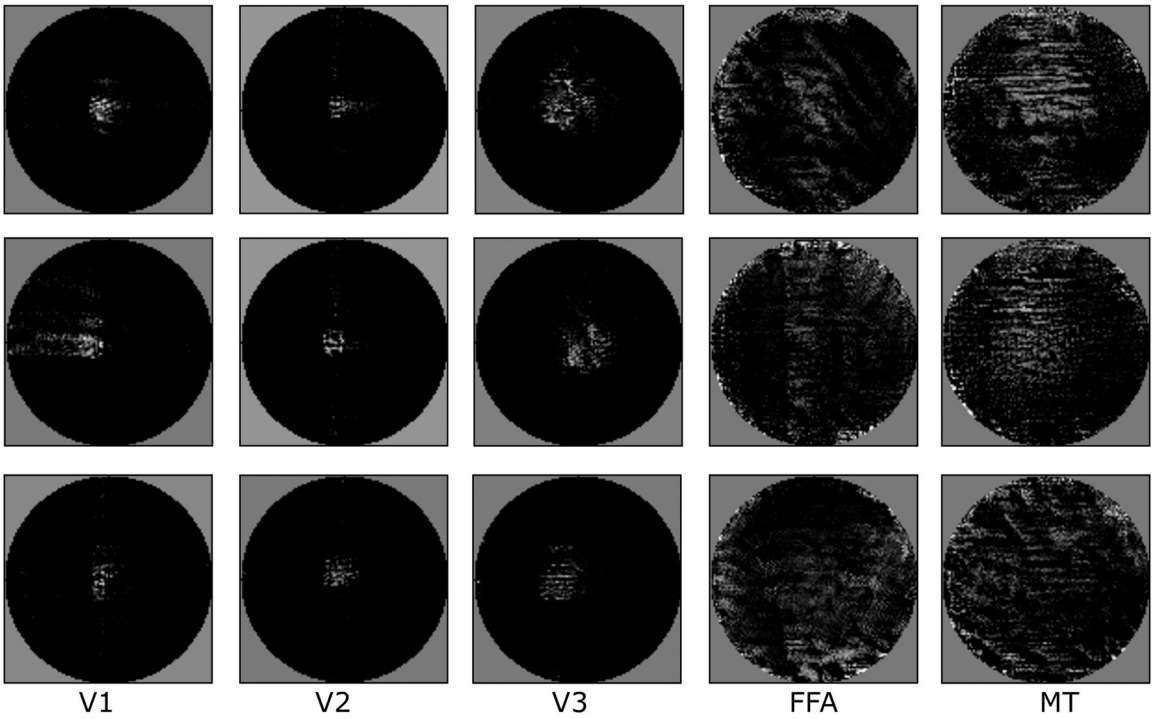

Fig 5. Examples of preferred inputs that maximize simulated voxel responses in different brain regions. Static frames from preferred inputs for three different voxels in the modeled ROIs. See S2 Video for observing the behaviour of these preferred inputs over time.

https://doi.org/10.1371/journal.pcbi.1008558.g005

A standard SGD optimizer was used together with an adaptive learning rate (starting value $\eta = 10^7$, reduction factor 0.8 after 5 iterations with no change) to optimize the stimuli. The iteration was stopped when no pixel changed more than 10^{-3} within 50 optimization steps.

As our video stimuli were square one way this optimization structure could exploit the objective was to cover the whole image with 45˚ oriented moving bars, as diagonals across the image would be the optimal way to create most energy within the input. We could work around this issue by retraining the NIF model with a circular aperture superimposed on the input videos. During preferred input optimization the aperture region was excluded by setting its gradients to 0. A similar effect could occur at small frequencies due to standard convolutional filters in current neural networks operating within square receptive fields. This can only be solved by adopting non-squared convolutional filters.

The results for different areas can be seen in Fig 5. All preferred inputs show superimposed moving wavelets at different orientations and frequencies. For V1, V2 and V3 they are constrained to their receptive fields. MT shows large circular fields of superimposed frequencies. FFA also shows larger regions of superimposed frequencies with circular dropouts.

The preferred inputs of V1, V2 and V3 are plausible, while the derivations for the higher order regions are difficult to interpret. Note that our example architecture is not biologically plausible, so this analysis should be read as a demonstration of the option of deriving preferred inputs of voxels rather than as a new insight into our cognition.

As stated at the beginning of this section, a different approach for visualizing what has been learned from the ROI data would be deriving what the higher order convolutional neural network channels represent, rather than observing what individual voxels prefer, i.e. a

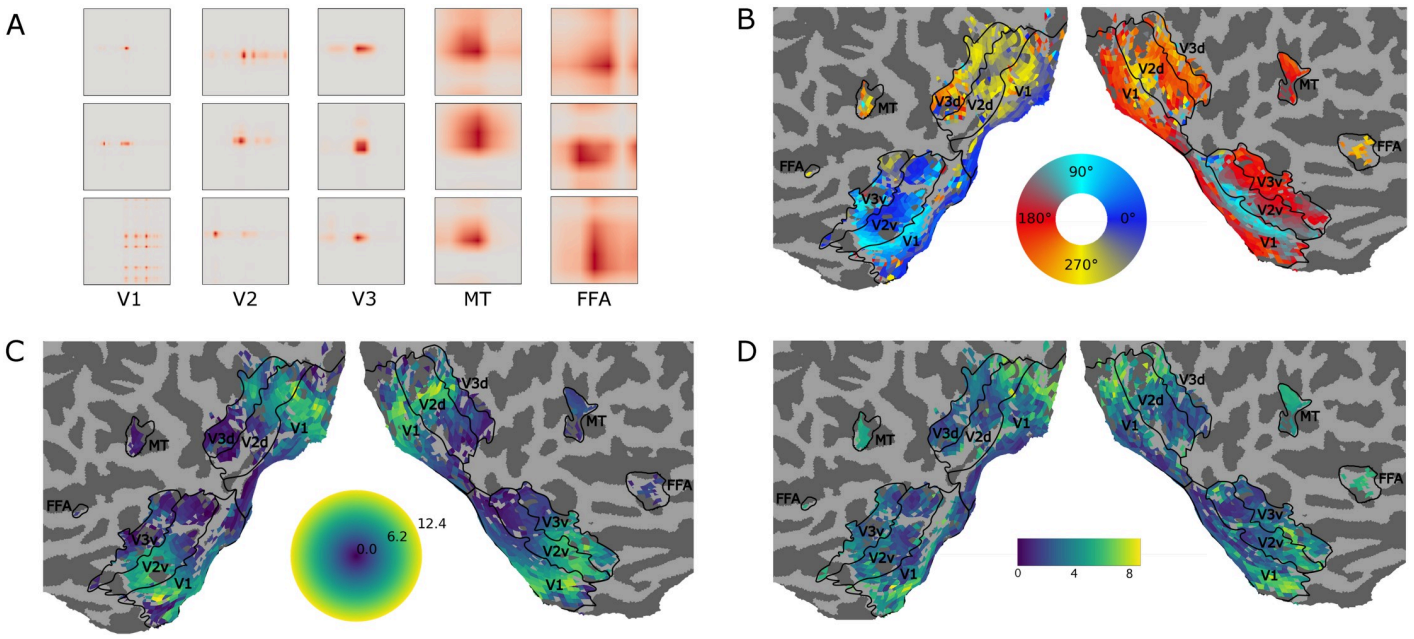

Fig 6. Receptive field maps. A. Various spatial receptive fields in video pixel space $\mathbf{U}_s$ learned for different ROIs within our framework. Most estimated spatial receptive fields are unipolar. B-D. Basic retinotopy that arose in the voxel-specific spatial observation matrix $\mathbf{U}_s$ within the NIF model. B. Polar angle. C. Eccentricity. D. Receptive field size.

https://doi.org/10.1371/journal.pcbi.1008558.g006

visualization of channels akin to Fig 4, but for higher order regions. This would avoid the superimposing nature of the voxelwise preferred images. This is a topic of research currently investigated by convolutional neural network interpretability, and not satisfactorily solved yet [46, 47].

Receptive field mapping

We examined whether the retinotopic organization of the visual cortex can be recovered from the spatial observation models [48]. Here, $\mathbf{U}_s$ represents spatial receptive field estimates for every voxel. Some of these voxel-specific receptive fields are shown in Fig 6A. The model has primarily learned classical local unimodal population receptive fields, but also more complex non-classical response profiles. This matches the expectation that population responses as inferred from neuroimaging data are not necessarily restricted to unipolar receptive fields. The model can be further constrained in case unipolar responses are expected (see [43] for a possible approach).

To check that the NIF model has indeed captured sensible retinotopic properties, we determined the center of mass of the spatial receptive fields and transformed these centers to polar coordinates using the central fixation point as origin. Sizes of the receptive fields were estimated as the standard deviation across $\mathbf{U}_s$, using the centers of mass as mean. Due to the pooling operations and convolutional processing, the $\mathbf{U}_s$ for each voxel had to be rescaled to the original input size to perform this operation. Voxels whose responses could not be significantly predicted were excluded from this analysis. Fig 6 shows polar angle (B), eccentricity (C) and receptive field size (D) for early visual system areas observed by our model. Maps were generated with `pycortex` [49]. Note that the boundaries between visual areas V1, V2 and V3 have been estimated with data from a classical wedge and ring retinotopy session. As can be seen, reversal boundaries align well with the traditionally estimated ROI boundaries. The larger

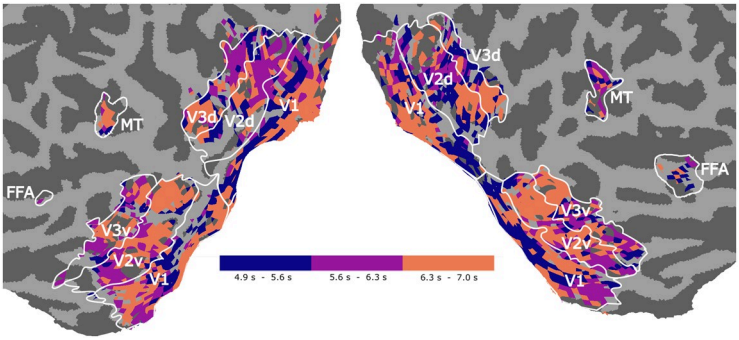

Fig 7. Differences in hemodynamic delay extracted from $\mathbf{U}_t$. For every voxel k we see the delay encoded in $\mathbf{U}_t[t, k]$ that has the maximal weight.

https://doi.org/10.1371/journal.pcbi.1008558.g007

eccentricity and increase in receptive field size (C) matches the expected fovea-periphery organization as well. Our results thus indicate that the NIF framework allows the estimation of accurate retinotopic maps from naturalistic videos.

Further properties of observational models

Recall that our model aims to predict the observed BOLD response from a spatiotemporal stimulus. We can obtain a rough estimate of the peak of the BOLD response by determining for each voxel the delay t that has the maximal weight $\mathbf{U}_t[t, k]$ assigned. Fig 7 shows the distribution of these delays across cortex, providing an insight into spatial differences in the hemodynamic response function. Results show a consistent slowing of the HRF for downstream areas [50].

Finally, we can investigate how stimulus features are encoded by investigating $\mathbf{U}_c$. In Fig 8 we show the feature weights for three different features in V1. We observe that different areas of early visual cortex show inhibition or excitation for the selected features. This provides insight into how stimulus features are represented across cortex.

It is of interest to examine whether these $\mathbf{U}_c$ weight distributions remain stable under different runs. We have run the same model five times, collecting the spatiotemporal channel weights and their associated $\mathbf{U}_c$ maps. Pairwise min-max-normalized mean-squared errors (MSE) were computed between these 5×64 channels to identify similar ones (low MSE implies similar channels, see Fig 9B for examples). The temporal dimension of the channels has been omitted by averaging over it as features appearing a few frames apart would have a

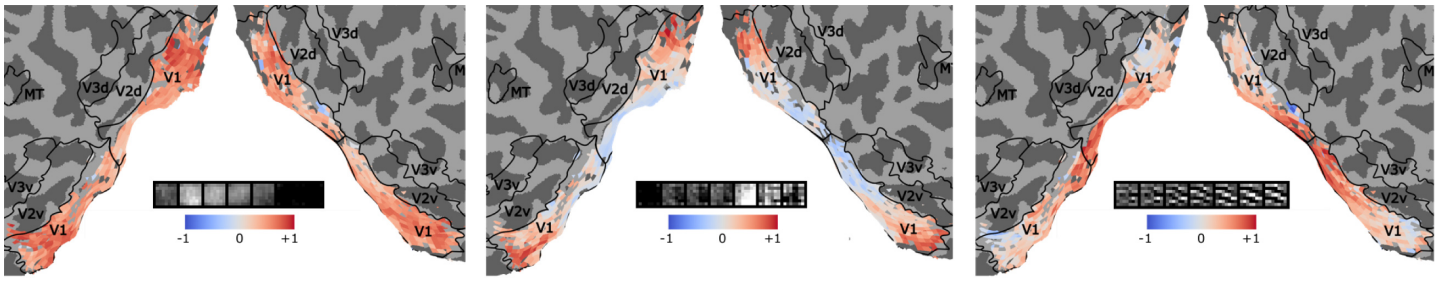

Fig 8. Projected $\mathbf{U}_c$ weight values for three different features in V1. Weight values were normalized between -1 and 1 by dividing them by the absolute maximum. The figure shows that the features are not evenly distributed across different cortical locations. The $\mathbf{U}_c$ matrix makes their analysis accessible.

https://doi.org/10.1371/journal.pcbi.1008558.g008

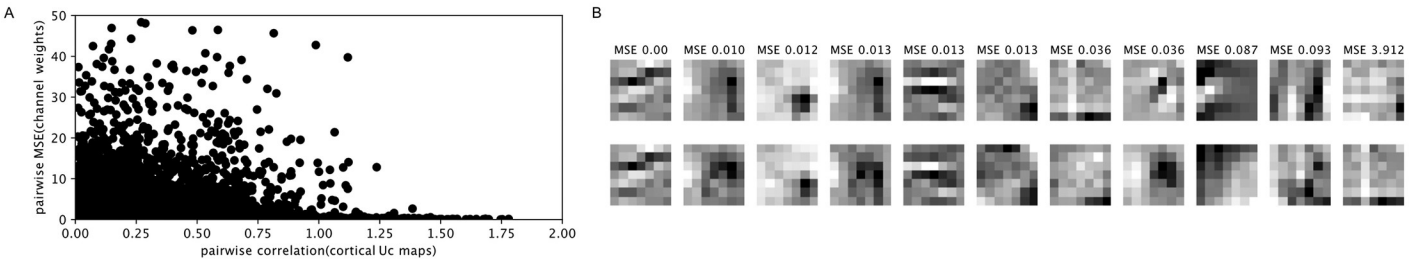

Fig 9. Relation between channel similarity and U_c map similarity. A. Relation between channel similarity and Uc map similarity in V1. Correlations are corrected with a fisher z-transform, and correlation signs are omitted. Highly similar Uc maps (high correlations) only occur for highly similar (small MSE) channels. However channels similar under MSE do not imply a highly similar Uc map. B. Examples of mean-squared error as a channel similarity measure.

https://doi.org/10.1371/journal.pcbi.1008558.g009

large influence on the MSE, but little influence on Uc due to temporal pooling to TR. Likewise, we took pairwise Pearson correlations between the Uc weight maps (only significantly predictable voxels) for each channel, leading to 5×64 comparisons between approximately 1000 voxel-wise weights in V1. Signs of correlations were omitted as negative correlations between maps point at inverted weight maps which may occur as $\mathbf{U}_c$ is not constrained to be positive. Fig 9A shows the relation between both measures. While we do see that highly similar Uc maps only occur for highly similar channels, highly similar channels do not necessarily have highly correlated Uc maps. This analysis has been restricted to V1 as similar image-based comparison of higher order convolutional features is not possible.

Processing of high-level semantic properties

So far, we have investigated characteristics of the NIF model that pertain to neural computations and representations and how these drive voxel responses. In this final analysis we investigate to what extent different neural populations are able to uncover high-level semantic content from the input stimulus. We focus on face detection since the processing of visual features pertaining to the discrimination of human faces is extremely well studied in the cognitive neuroscience literature [51]. In particular, FFA is known to play a central role in the visual processing of human faces [52]. Consequently, we expect that the representations learned by the FFA component of our model are related to human face processing.

We test this hypothesis using an *in silico* experiment closely resembling standard fMRI experimental procedures in cognitive neuroscience. We passed 90 video segments of the regular input length of 3 TR, taken from the test set, through the trained NIF model. These videos were divided into two classes, one containing frontal views of human faces and the other not containing faces (45 videos per class). We analyzed the predicted BOLD responses of the models in the two experimental conditions using a mass univariate approach. For each voxel, we computed the t-statistic of the face minus no-face contrast and the associated p-values. We corrected for multiple comparisons using the false discovery rate (FDR) with alpha equal to 10^{-4}. The left panel of Fig 10A shows the fraction of significant voxels in each brain region. The results show that FFA is the only region that is significantly activated by the contrast. The right panel shows that the voxels which are significantly activated also tend to be significantly predicted by the model. Fig 10B shows the significant (absolute) t-scores on the cortex.

We complemented these results with a multivariate decoding analysis [53]. We trained a logistic regression model on the predicted voxel responses of each ROI in order to predict if the input contained faces. We also performed this logistic regression analysis directly on the channel responses of the model (max-pooled across the spatio-temporal feature map). In the analysis we also included direct predictions from the pixel values of the input images. We

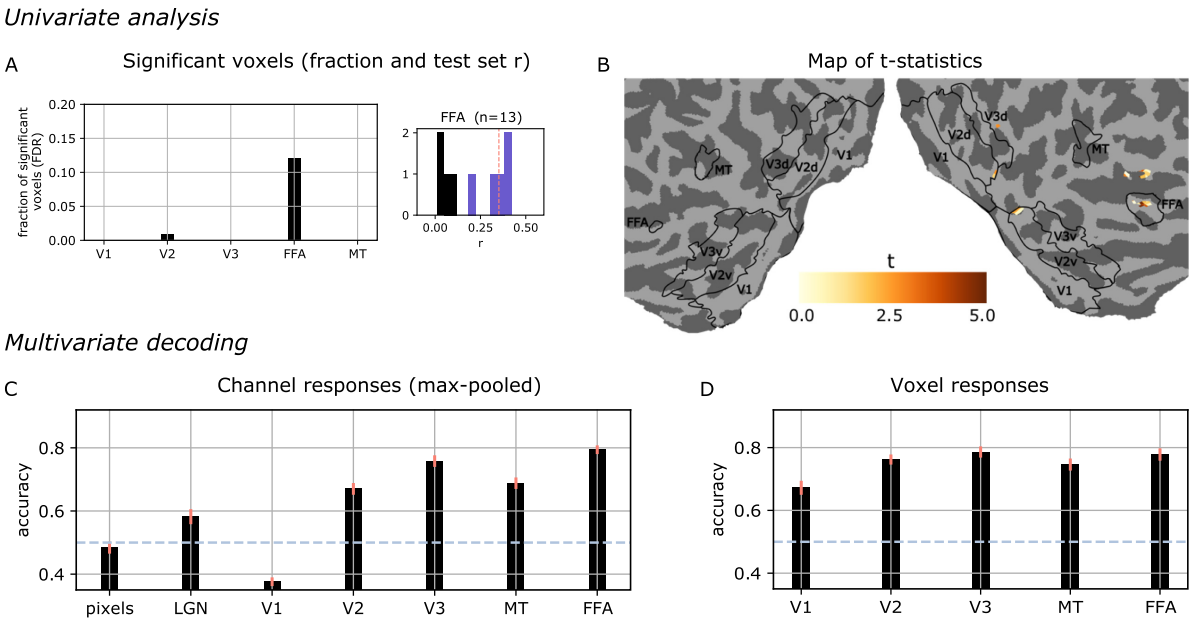

Fig 10. Results of an in-silico experiment. The trained network was presented with video segments from the test set showing either faces or no faces. A., B. Univariate analysis. A. Significant voxels in each ROI. Correlations between predicted and observed voxel responses on the test set. B. Cortical map of the t-statistic for univariate analysis. C., D. Multivariate logistic regression. C. Decoding from ROI-wise tensor activations (channel responses max-pooled across the whole feature map) or raw input values (*pixels*, *LGN*). D. Decoding from predicted voxel responses. Overall, we see that FFA is the most discriminative area for the face recognition experiment.

https://doi.org/10.1371/journal.pcbi.1008558.g010

estimated the mean accuracy and its standard error by repeating the training 50 times with random splits into 35 training and 10 test examples respectively. As shown in Fig 10, the highest classification performance is achieved for FFA, both at the channel level and at the voxel level. This confirms our expectation that the model FFA has learned higher-order semantic properties that match its functional role in the brain. Furthermore, we see that multivariate data from increasingly downstream regions are more suitable to dissociate faces from non-faces. This indicates the prospect of studying *in silico* what behavioural goals higher-order sensory areas are optimized for. This also hints at the possibility of using neural information processing systems estimated from brain data to support the solution of pattern recognition tasks.

Data requirements

The training of modern convolutional neural networks is known to require large amounts of data. The modeling framework described here likewise has data requirements that are not fulfilled by the large majority of current neuroscientific experiments. The required amount of data for a saturating model is unclear however. Fig 11 describes the data requirements for the specific experiment presented here. The example model we present saturates around the 12 hour mark. As several factors influence the required amount of training data this should neither be understood as a lower nor a higher bound on the amount of data required for applying this method. In general, we recommend to record single runs until test performance saturates.

The upper bars show the ROI-wise median of the voxel-wise noise ceiling of the correlation. It is an estimate of the upper limit on any model's predictability attainable on the repeated test data set, given the noise in the data. An early description of the idea behind the noise ceiling can be found in [54]. We have used the Monte Carlo noise ceiling (MCnc) method mentioned in [55] and [56], and described in more detail in [57]. We have used Kendrick Kay's public

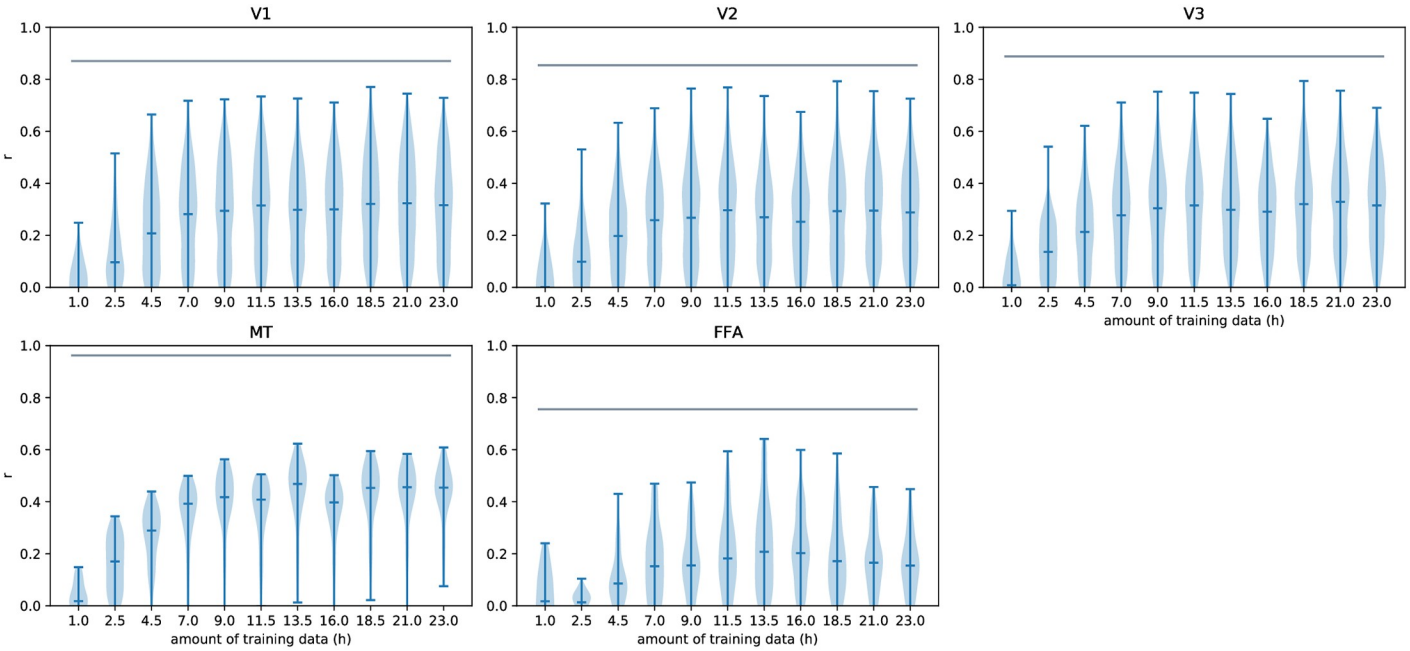

Fig 11. Test set performance over different amounts of training data. The example model was trained with increasing amounts of data, starting from the initial session. Voxel-wise correlations were determined on the test set for different areas, their distributions shown here. The performance of the example model saturates around the 12 hour mark. This result is likely specific for the stimulus modality, recording parameters, the model architecture and our particular participant.

https://doi.org/10.1371/journal.pcbi.1008558.g011

implementation. In the MCnc method, for every individual voxel, median correlations between simulated measurements and signals are estimated in a Monte Carlo simulation setting. Here a *measurement* is the sum between a signal and a noise component. *Signal* and *noise* are assumed to follow Gaussian distributions, for which mean and variance parameters are estimated from the z-scored data. The signal mean is the mean across the averaged test data time course. The noise mean is assumed to be 0. The noise variance is estimated across all test data repetitions. The signal variance is the rectified (non-negative) difference between the variance across the averaged test data time course and the noise variance. Using these parameters we have performed 500 signal simulations with 22 measurements (same signal, different noise) each. The figure shows the median of the voxel-wise noise ceilings within individual ROIs.

Comparing to the task-driven approach

The currently most used technique for describing visual and auditory hierarchies is task-driven modeling with convolutional neural networks. A hypothesized convolutional neural network architecture is trained on a dataset with a specific objective function. Then experimental stimuli are passed through this pretrained architecture to obtain layer-wise activities in response to these stimuli, and the activity tensors are compared to brain activity under the same stimuli with encoding models or RSA. With these methods, layer distributions are identified across cortex. Many correspondences between modern convolutional neural networks and the visual system could be uncovered using the task-driven method.

Our aim with this paper is not to rival the currently best models in this area of visual modeling, but to propose a new approach to computational modeling of neural processing systems with a simplified visual system architecture as an example. Nevertheless we would like to attempt comparing quantitative performance between the task-driven and a data-driven approach using greedy readout models in the human visual information processing system.

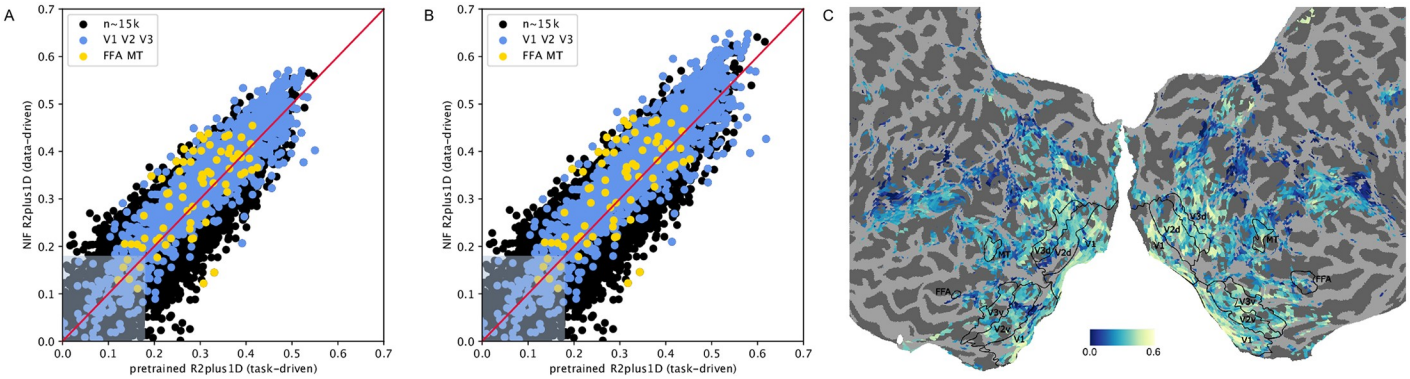

Fig 12. Comparison between task-driven and data-driven approach on our dataset. A. Correlations for early visual system and higher order areas. B. Correlations for early visual system and higher order areas (fisher-z corrected for linear comparability). Shaded areas cover non-significant voxels. C. Areas analyzed in this comparison, and their projected correlations.

https://doi.org/10.1371/journal.pcbi.1008558.g012

For video stimuli experiments it is common to use convolutional neural networks trained on video action classification [10]. We chose the `R(2+1)D` architecture [58], a well-performing network developed for action recognition on the Kinetics data set [59], based on ResNet [60]. It is a modern neural network architecture, including typical modern model choices like skip connections, batch normalization, ReLU units; and utilizing complex convolutional blocks with separated temporal and spatial convolutions. The network, originally trained on 15 Hz Kinetics data was fine-tuned on converted 22.86 Hz Kinetics data to align the learned temporal dynamics with our own data. The original model classified on cropped spatial windows inside the $112 \times 112 \times 16$ data, which we omitted during fine-tuning to keep the input fixated around the fovea as in our NIF example model. The other training settings were kept identical to the description in the original paper and in the code, with the pretrained model published in `pytorch torchvision` [61].

Approximately 15.000 voxels with highest variance during the test set recordings were selected for this analysis, a number chosen in order to cover most of the visual system (see Fig 12C). We compared the task-driven case, using features pretrained on Kinetics; and the purely data-driven case, training all network parameters (convolutional features) on the objective function of predicting brain activity as in the NIF framework (thus denoted NIF in the figure). In both cases activity of all voxels was predicted based on the activity tensors `conv1` to `conv5` separately and in the same model. In the task-driven case, $\mathbf{U}_s$ and $\mathbf{U}_c$ readout parameters were learned for every voxel and layer, while the fixed pretrained features acted as a basis. In the data-driven case, both readout parameters and all convolutional block features were learned. RGB input was used, and the z-standardisation normalization used during pretraining was applied in the task-driven case as otherwise its performance would have been lower. The temporal dimension was omitted as `R(2+1)D` expects 16 frames. At 22.86 Hz this matched the number of frames shown in 1 TR of our data, so this merely restricted the model to predicting voxel-wise activity from video covering 1 TR instead of 3 TR, and not learning $\mathbf{U}_t$ parameters. To obtain voxel-wise correlations to estimate model performance, after model training for every voxel we chose the top-performing layer on the test set.

Results are shown in Fig 12. The task-driven and the data-driven approach are similar in performance, but the data-driven NIF-based approach outperforms the task-driven one using pretrained features especially in the early visual system and in higher order ROIs.

As the correlations achieved by the task-driven model are still relatively high and similar to the purely data-driven model our result only slightly contradicts the result of [13], where the

predictive power of the pretrained features performed slightly better in V1. Potential explanations for these differences include that the `R(2+1)D` convolutional neural network does not match brain hierarchies well, however we do see a visible improvement in the data-driven case. Another explanation for these differences is that the cranial window in [13] has been on an area where pretrained DNN features indeed match V1 feature detectors well. Another different explanation is that the higher resolution of electrophysiological recordings leads to more accurate results than our functional MRI data.

This model comparison will not rule out the possibility that the pretrained features can be improved upon by using newer model developments from the machine learning community, or a more brain-like task. This numerical performance comparison should not distract the reader from recognizing the fundamental difference between the task-driven and our suggested data-driven modeling approach. By imposing an architecture of ROIs instead of taking the greedy approach, implemented as separate convolutional layers; we expect to learn the information processing between ROIs. A numerical performance comparison for this idea of training end-to-end models representing visual system architectures does not exist yet. Also, for sensory systems we believe it is worth exploring whether the data-driven approach leads to more accurate ROI representations, especially in higher order areas which divide into specialized areas solving different tasks important for human cognition—not all of which are known, and some of which may not be describable by neural network objective functions.

Discussion

This paper proposes neural information flow for neural system identification. The approach relies on neural architectures described in terms of interacting brain regions, each performing nonlinear computations on their input. By coupling each brain region with associated measurements of neural activity, we can estimate neural information processing systems end-to-end. Using fMRI data collected during prolonged naturalistic stimulation we showed that we can successfully predict BOLD responses across different brain regions. Furthermore, meaningful spatial, temporal and feature receptive fields emerged after model estimation. The learned receptive fields are specific to each brain region but collectively explain all of the observed measurements. To the best of our knowledge, these results demonstrate for the first time that biologically interpretable information processing systems consisting of multiple interconnected brain regions can be directly estimated end-to-end from neural data.

As explained in the introduction, NIF generalizes current encoding models. For example, basic population receptive field models [62] and more advanced neural network models [5] are special cases of NIF that assume no interactions between brain regions and make specific choices for the nonlinear transformations that capture neuronal processing.

The researcher can specify alternative NIF models and then use explained variance as a model selection criterion. This is similar in spirit to dynamic causal modeling (DCM) [63]. However, NIF models can identify changes in neural computation that are not detectable in approaches that only focus on estimating effective connectivity. For example, they can be used to investigate in detail the changes in neural information processing under different conditions.

NIF can be naturally extended in several directions. The employed convolutional layer to model neural computation can be replaced by neural networks that have a more complex architecture. For example, recurrent neural networks can be trained in the same way as the feed-forward architecture presented here. Furthermore, lateral and feedback processing is easily included by adding additional links between brain regions and unrolling the backpropagation procedure over time. NIF models can also be extended to handle other data modalities.

Alternative observation models can be formulated that allow inferring neural computations from other measures of neural activity (e.g., single- and multi-unit recordings, local field potentials, calcium imaging, EEG, MEG). Moreover, NIF models can be trained on multiple heterogeneous datasets at the same time, providing a solution for multimodal data fusion. The framework can also be applied to other sensory inputs. For example, auditory areas can be trained on auditory input (see e.g. [64]). If this is combined with visual input then we may be able to uncover new properties of multimodal integration [65].

Note that we are not restricted to using neural data as the sole source of training signal. We may instead (or additionally) condition these models on behavioral data, such as motor responses or eye movements [23]. The resulting models should then show the same behavioral responses as the system under study. We can also teach NIF models to perceive and act upon the task at hand directly using reinforcement neural network training [66]. In this way, NIF models provide a starting point for creating brain-inspired AI systems that more closely model how real brains solve cognitive tasks.

Finally, we can use NIF models as *in silico* models to examine changes in neural computation. For example, we can examine how neural representations change during learning or as a consequence of virtual lesions in the network [67]. This can provide insights into cognitive development and decline. We can also test what happens to neural computations when we directly drive individual brain regions with external input. This provides new ways for understanding how brain stimulation modulates neural information processing, guiding the development of future neurotechnology [68].

Summarizing, we view NIF as a way to construct biologically-inspired computational models that capture neural information processing in biological systems. As such, it provides a blend of computational and experimental neuroscience [69]. This gives us a principled approach to make sense of the high-resolution datasets produced by continuing advances in neurotechnology [70]. We expect that NIF models will deliver exciting new insights into the principles and mechanisms that determine neural information processing in biological systems.

Code accessibility

A basic implementation of the NIF method on a smaller data set [71, 72] can be found at github.com/kateiyas/basicNIF.

Supporting information

S1 Video. Features (weights) learned inside the neural network layer for V1.
(GIF)

S2 Video. Animated preferred inputs for voxels in specific ROIs.
(GIF)

Acknowledgments

We would like to thank Kendrick Kay who helped us with estimating the noise ceiling on our data, and Martin Hebart for advice and discussions.

Author Contributions

Conceptualization: L. Ambrogioni, U. Güçlü, M. A. J. van Gerven.

Data curation: K. Seeliger.

Formal analysis: K. Seeliger, L. M. van den Bulk.

Funding acquisition: M. A. J. van Gerven.

Investigation: K. Seeliger, U. Güçlü, M. A. J. van Gerven.

Methodology: K. Seeliger, L. Ambrogioni, M. A. J. van Gerven.

Project administration: L. Ambrogioni, M. A. J. van Gerven.

Resources: M. A. J. van Gerven.

Software: K. Seeliger, Y. Güçlütürk, L. M. van den Bulk.

Supervision: L. Ambrogioni, M. A. J. van Gerven.

Validation: K. Seeliger.

Visualization: K. Seeliger, Y. Güçlütürk, U. Güçlü.

Writing – original draft: K. Seeliger, L. Ambrogioni, M. A. J. van Gerven.

Writing – review & editing: K. Seeliger, Y. Güçlütürk, M. A. J. van Gerven.

References

1. Churchland PS, Sejnowski TJ. The Computational Brain. MIT Press; 1992.
2. Stanley GB. Neural system identification. In: Bioelectric Engineering. Springer; 2005. p. 367–388.
3. Wu MCK, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. Annual Review of Neuroscience. 2006; 29(1):477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024 PMID: 16776594
4. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage. 2011; 56(2):400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073 PMID: 20691790
5. van Gerven MAJ. A primer on encoding models in sensory neuroscience. Journal of Mathematical Psychology. 2017; 76:172–183. https://doi.org/10.1016/j.jmp.2016.06.009
6. Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience. 2016; 19(3):356–365. https://doi.org/10.1038/nn.4244 PMID: 26906502
7. Kriegeskorte N. Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science. 2015; 1:417–446. https://doi.org/10.1146/annurev-vision-082114-035447 PMID: 28532370
8. Lindsay GW. Convolutional neural networks as a model of the visual system: Past, present, and future. arXiv preprint arXiv:200107092. 2020.
9. Güçlü U, van Gerven MAJ. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. The Journal of Neuroscience. 2015; 35(27):10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015 PMID: 26157000
10. Güçlü U, van Gerven MAJ. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage. 2015; 145:329–336. http://dx.doi.org/10.1016/j.neuroimage.2015.12.036. PMID: 26724778
11. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports. 2016; 6(1). https://doi.org/10.1038/srep27755 PMID: 27282108
12. Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nature Communications. 2017; 8(15037). https://doi.org/10.1038/ncomms15037 PMID: 28530228
13. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Computational Biology. 2019; 15(4):e1006897. https://doi.org/10.1371/journal.pcbi.1006897 PMID: 31013278
14. Cadena SA, Sinz FH, Muhammad T, Froudarakis E, Cobos E, Walker EY, et al. How well do deep neural networks trained on object recognition characterize the mouse visual system? In: Future directions at the intersection of neuroscience and artificial intelligence (Neuro-AI) workshop during Advances in Neural Information Processing Systems (NeurIPS); 2019. Available from: https://openreview.net/forum?id=rkxcXmtUUS.

15. Shi J, Shea-Brown E, Buice M. Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. In: Advances in Neural Information Processing Systems (NeurIPS); 2019. p. 5765–5775.
16. Joukes J, Hartmann TS, Krekelberg B. Motion detection based on recurrent network dynamics. Frontiers in Systems Neuroscience. 2014; 8:239. https://doi.org/10.3389/fnsys.2014.00239 PMID: 25565992
17. Klindt DA, Ecker AS, Euler T, Bethge M. Neural system identification for large populations separating "what" and "where". In: Advances in Neural Information Processing Systems (NeurIPS); 2017. p. 3506–3516.
18. St-Yves G, Naselaris T. The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. NeuroImage. 2018; 180:188–202. https://doi.org/10.1016/j.neuroimage.2017.06.035 PMID: 28645845
19. Antolík J, Hofer SB, Bednar JA, Mrsic-Flogel TD. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. PLoS Computational Biology. 2016; 12(6):1–22. https://doi.org/10.1371/journal.pcbi.1004927 PMID: 27348548
20. McIntosh LT, Maheswaranathan N, Nayebi A, Ganguli S, Baccus SA. Deep learning models of the retinal response to natural scenes. In: Advances in Neural Information Processing Systems (NeurIPS); 2016. p. 1369–1377.
21. Batty E, Merel J, Brackbill N, Heitman A, Sher A, Litke AM, et al. Multilayer Recurrent Network Models of Primate Retinal Ganglion Cell Responses. In: International Conference on Learning Representations (ICLR); 2017. p. 1–12. Available from: https://openreview.net/forum?id=HkEI22jeg.
22. Kindel WF, Christensen ED, Zylberberg J. Using deep learning to reveal the neural code for images in primary visual cortex. arXiv preprint arXiv:200107092. 2017.
23. Sinz F, Ecker AS, Fahey P, Walker E, Cobos E, Froudarakis E, et al. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In: Advances in Neural Information Processing Systems (2018); 2018. p. 7199–7210.
24. Ecker AS, Sinz FH, Froudarakis E, Fahey PG, Cadena SA, Walker EY, et al. A rotation-equivariant convolutional neural network model of primary visual cortex. arXiv preprint arXiv:180910504. 2018.
25. Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N. Recurrence is required to capture the representational dynamics of the human visual system; 116(43):21854–21863.
26. McClure P, Kriegeskorte N. Representational distance learning for deep neural networks. Frontiers in Computational Neuroscience. 2016; 10:131. https://doi.org/10.3389/fncom.2016.00131 PMID: 28082889
27. Güçlü U, van Gerven MAJ. Modeling the Dynamics of Human Brain Activity with Recurrent Neural Networks. Frontiers in Computational Neuroscience. 2017; 11:7. https://doi.org/10.3389/fncom.2017.00007 PMID: 28232797
28. Tripp B. Approximating the architecture of visual cortex in a convolutional network. Neural Computation. 2019; 31:1–41. https://doi.org/10.1162/neco_a_01211
29. Seeliger K, Sommers RP, Güçlü U, Bosch SE, van Gerven MAJ. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. bioRxiv preprint. 2019; https://doi.org/10.1101/687681.
30. Uludağ K, Roebroeck A. General overview on the merits of multimodal neuroimaging data fusion. NeuroImage. 2014; 102:3–10. https://doi.org/10.1016/j.neuroimage.2014.05.018 PMID: 24845622
31. Tokui S, Oono K, Hido S, Clayton J. Chainer: A next-generation open source framework for deep learning. In: Proceedings of Workshop on Machine Learning Systems (LearningSys) during Advances in Neural Information Processing Systems (NeurIPS). vol. 5; 2015. p. 1–6.
32. Davies RT, Gardner J, Moffat S, Young M, Collinson P. Doctor Who; 2005.
33. ITU-R. Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. International Telecommunications Union. 2011.
34. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
35. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015. p. 1026–1034.
36. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies; 21:1–6.
37. Graham DJ, Chandler DM, Field DJ. Can the theory of "whitening" explain the center-surround properties of retinal ganglion cell receptive fields? Vision Research. 2006; 46(18):2901–2913. https://doi.org/10.1016/j.visres.2006.03.008 PMID: 16782164

38. Dan Y, Atick JJ, Reid RC. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. Journal of Neuroscience. 1996; 16(10):3351–3362. https://doi.org/10.1523/JNEUROSCI.16-10-03351.1996 PMID: 8627371

39. Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M. Independence of luminance and contrast in natural scenes and in the early visual system. Nature Neuroscience. 2005; 8(12):1690. https://doi.org/10.1038/nn1556 PMID: 16286933

40. Jones JP, Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology. 1987; 58:1233–1258. PMID: 3437332

41. Bashivan P, Kar K, DiCarlo JJ. Neural population control via deep image synthesis. Science. 2019; 364 (6439). https://doi.org/10.1126/science.aav9436 PMID: 31048462

42. Lehky SR, Sejnowski TJ, Desimone R. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns.; 12(9):3568–81.

43. Walker EY, Sinz FH, Cobos E, Muhammad T, Froudarakis E, Fahey PG, et al. Inception loops discover what excites neurons most using deep predictive models; 22(12):2060–2065.

44. Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences; 177 (4):999–1009.e10.

45. Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. University of Montreal; 2009.

46. Xie N, Ras G, van Gerven MAJ, Doran D Explainable Deep Learning: A Field Guide for the Uninitiated. arXiv preprint. 2020; arXiv:2004.14545.

47. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A. The Building Blocks of Interpretability. Distill. 2018. http://dx.doi.org/10.23915/distill.00010

48. Wandell BA, Winawer J. Computational neuroimaging and population receptive fields. Trends in Cognitive Sciences. 2015; 19(6):349–357. https://doi.org/10.1016/j.tics.2015.03.009 PMID: 25850730

49. Gao JS, Huth AG, Lescroart MD, Gallant JL. Pycortex: An interactive surface visualizer for fMRI. Frontiers in Neuroinformatics. 2015; 9:23. https://doi.org/10.3389/fninf.2015.00023 PMID: 26483666

50. Calhoun VD, Adali T, Kraut M, Rivkin P, Pearlson G. Visualizing spatially distributed hemodynamic lag times in event-related functional MRI: Estimation of a characteristic visual "impulse response". In: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 1998. p. 2124–2127.

51. Haxby JV, Hoffman EA, Gobbini MI. The distributed human neural system for face perception. Trends in Cognitive Sciences. 2000; 4(6):223–233. https://doi.org/10.1016/S1364-6613(00)01482-0 PMID: 10827445

52. Kanwisher N, McDermott J, Chun MM. The fusiform face area: A module in human extrastriate cortex specialized for face perception. Journal of Neuroscience. 1997; 17(11):4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997 PMID: 9151747

53. Haxby JV, Connolly AC, Guntupalli JS. Decoding neural representational spaces using multivariate pattern analysis. Annual Review of Neuroscience. 2014; 37:435–456. https://doi.org/10.1146/annurev-neuro-062012-170325 PMID: 25002277

54. David SV, Gallant JL. Predicting neuronal responses during natural vision. Network: Computation in Neural Systems. 2005; 16(2-3):239–260. https://doi.org/10.1080/09548980500464030 PMID: 16411498

55. Lage-Castellanos A, Valente G, Formisano E, De Martino F. Methods for computing the maximum performance of computational models of fMRI responses. PLoS Computational Biology. 2019; 15(3): e1006397. https://doi.org/10.1371/journal.pcbi.1006397 PMID: 30849071

56. Han K, Wen H, Shi J, Lu KH, Zhang Y, Fu D, et al. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. NeuroImage. 2019; 198:125–136. https://doi.org/10.1016/j.neuroimage.2019.05.039 PMID: 31103784

57. Kay KN, Winawer J, Mezer A, Wandell BA. Compressive spatial summation in human visual cortex. Journal of Neurophysiology. 2013; 110(2):481–494. https://doi.org/10.1152/jn.00105.2013 PMID: 23615546

58. Tran D, Wang H, Torresani L, Ray J, LeCun Y and Paluri M. A closer look at spatiotemporal convolutions for action recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018; 6450–6459.

59. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A. The Kinetics Human Action Video Dataset. arXiv preprint. 2017; arXiv:1705.06950.

60. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015; 770-–778.

61. Marcel S, Rodriguez Y. He K, Zhang X, Ren S, Sun J. Torchvision: The machine-vision package of torch. Proceedings of the 18th ACM International Conference on Multimedia. 2010; 1485–1488.

62. Dumoulin SO, Wandell BA. Population receptive field estimates in human visual cortex. NeuroImage. 2008; 39(2):647–660. https://doi.org/10.1016/j.neuroimage.2007.09.034 PMID: 17977024

63. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003; 19(4):1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7 PMID: 12948688

64. Güçlü U, Thielen J, Hanke M, van Gerven MAJ. Brains on beats. In: Advances in Neural Information Processing Systems (NeurIPS); 2016. p. 2101–2109.

65. Simanova I, Hagoort P, Oostenveld R, van Gerven MA. Modality-independent decoding of semantic information from the human brain. Cerebral Cortex. 2014; 24:426–434. PMID: 23064107

66. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Boston, MA: The MIT Press; 2017.

67. Graziano MSA, Aflalo TN. Mapping behavioral repertoire onto the cortex. Neuron. 2007; 56:239–251. https://doi.org/10.1016/j.neuron.2007.09.013 PMID: 17964243

68. Roelfsema PR, Denys D, Klink PC. Mind reading and writing: The future of neurotechnology. Trends in Cognitive Sciences. 2018; 22(7):1–13. https://doi.org/10.1016/j.tics.2018.04.001 PMID: 29729902

69. Churchland PS, Sejnowski TJ. Blending computational and experimental neuroscience. Nature Reviews Neuroscience. 2016; 17(11):667–668. https://doi.org/10.1038/nrn.2016.114 PMID: 30283241

70. Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nature Neuroscience. 2011; 14(2):139–142. https://doi.org/10.1038/nn.2731 PMID: 21270781

71. Schoenmakers S, Güçlü U, van Gerven MAJ, Heskes T. Gaussian mixture models and semantic gating improve reconstructions from human brain activity. Frontiers in Computational Neuroscience. 2015; 8:173. https://doi.org/10.3389/fncom.2014.00173 PMID: 25688202

72. Schoenmakers S, Barth M, Heskes T, van Gerven MAJ. Linear reconstruction of perceived images from human brain activity. NeuroImage. 2013; 83:951–961. https://doi.org/10.1016/j.neuroimage.2013.07.043 PMID: 23886984