Neural Information Flow: Learning neural information processing systems from brain activity

K. Seeliger* **(kseeliger@posteo.jp)**

L. Ambrogioni* **(luca.ambrogioni@gmail.com)**

U. Güçlü (umuguc@gmail.com)

M. A. J. van Gerven (m.vangerven@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour
Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Abstract

Neural information flow (NIF) is a new framework for system identification in neuroscience. NIF models represent neural information processing systems as coupled brain regions that each embody neural computations. These brain regions are coupled to observed data specific to that region via linear observation models. NIF models are trained via backpropagation, directly leveraging the neural signal as the loss. Trained NIF models are accessible for *in silico* analyses. Using a large-scale fMRI video stimulation dataset and a feed-forward convolutional neural network-based NIF model as an example we show that, in this manner, we can estimate models that learn meaningful neural computations and representations. Our framework is general in the sense that it can be used in conjunction with any neural recording techniques. It is also scalable, providing neuroscientists with a principled approach to make sense of high-dimensional neural datasets.

Keywords: system identification ; neural networks ; vision

*: Equal contribution

Introduction

Uncovering the neural computations that subserve behaviour and cognition is a major goal in neuroscience (Churchland & Sejnowski, 1992). Arguably, true understanding of the brain requires replicating biological neural information processing *in silico*. The goal of *neural system identification* is to uncover neural information processing systems from observed measurements in response to environmental changes (Stanley, 2005; Wu, David, & Gallant, 2006). However, a generally accepted method for deriving and recovering neural computations from observed brain data has not been proposed so far. For sensory systems, relating hypothesized representations to hierarchical processing observed in brain activity – with representation similarity analysis (RSA) or encoding models (Naselaris, Kay, Nishimoto, & Gallant, 2011; van Gerven, 2017) – is the currently most widely adapted method for studying the hierarchy of neural information processing. Remarkably, the current best representation hierarchies explaining information processing along the ventral visual stream can be extracted from externally trained convolutional neural networks (Kriegeskorte, 2015; Güçlü & van Gerven, 2015; Yamins et al., 2014). However, the hypothesized feature representations used in representation studies have been estimated with artificially defined goals, such as discrete classification on the ImageNet competition. Furthermore, within these commonly used analysis methods representations remain uncoupled and can not represent causal bottom-up and top-down cognitive information flow between brain regions. That is, they are lacking the coupling of distinct neural regions commonly studied with effective connectivity methods. Established techniques for uncovering effective connectivity such as Dynamic Causal Modeling (DCM) are able to uncover this causal coupling (Friston, Harrison, & Penny, 2003), but do not capture the nature of information processing that drives the interaction between brain regions.

We propose a new modeling framework combining the power of neural network-based representations and effective connectivity methods. The parameters of Neural Information Flow (NIF) models are directly estimated on brain data. They learn to generate neural data in regions of interest in response to interactions between neural regions and/or sensory or behavioural input. They can be interpreted as synthetic (*in silico*) brains that learn to capture the nonlinear computations taking place in a real brain. Instead of modeling brain regions in an isolated fashion, they take afferent information inputs into account (Haak et al., 2013).

In the following we outline the basic methodology of NIF modeling. Using a large dataset acquired under naturalistic video stimulation we demonstrate that the model is capable of generating realistic brain measurements and that the computations captured in the model are biologically meaningful.

Methodology

The example architecture presented here is formulated within a state-of-the-art neural network framework, using backpropagation for estimating its parameters. We represent the information contained in an individual neural population as tensor $\mathbf{N}$ that embodies activity in spatio-temporal receptive fields. Voxel activity in the regarded region of interest is coupled to $\mathbf{N}$ with spatial, temporal and channel observation models, and drives the parameter learning in the convolutional lay-

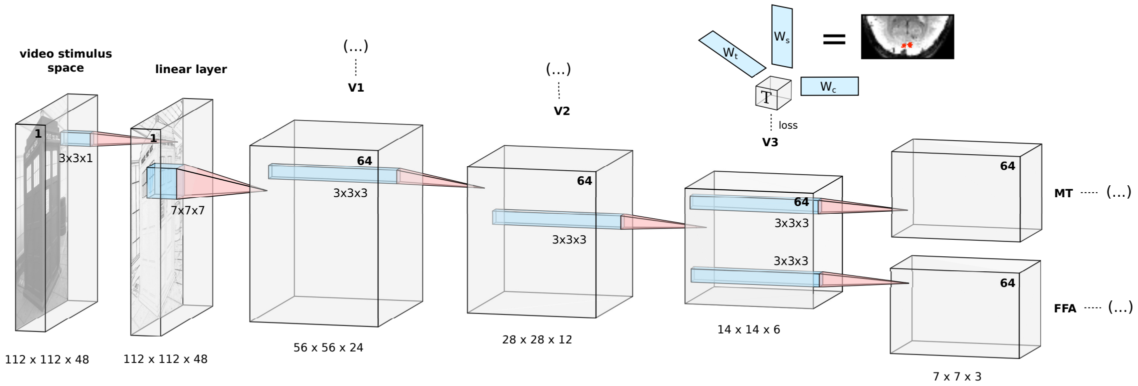


Figure 1: Architecture of the example NIF model of the early visual system. Underneath the tensors resulting from the 3D convolution operations (with 3D grayscale stimulus video patches being the input to the network) we state the size of each input space ($x \times y \times t$) to the next layer. We use 3D video patches consisting of 3×16 frames (covering three TRs of 700 ms each), pushed forwards in time by $\Delta_t = 6$ TR (4.2 s) to align them with the hemodynamic response. Multiple video patches are used to allow for voxel-specific differences in the hemodynamic response peak (learned within $\boldsymbol{W}_t$). The number of feature maps (channels) in each input space is printed in boldface, with the stimulus (input) space consisting of a single channel (grayscale). Before any of the observed tensors we pass the input video patch through a single-channel ($3 \times 3 \times 1$) convolutional layer without a nonlinearity, serving as a learnable linear preprocessing step that accounts for retinal and LGN modulations. The voxel- and region-specific observation models consisting of the spatio-temporal weight vectors $\boldsymbol{W}_x$, $\boldsymbol{W}_y$ and $\boldsymbol{W}_t$ tap into their specific tensor, with FFA and MT having their own tensors from V3 input for analysis purposes. Convolutional kernel sizes are $7 \times 7 \times 7$ in the first convolutional layer (leading to the V1 tensor), and $3 \times 3 \times 3$ for all other layers. After every convolution operation we apply a sigmoid nonlinearity and spatial average pooling with $2 \times 2 \times 2$ kernels. Before entering the $\boldsymbol{W}_t$ observation models the temporal dimension is average pooled so that each point t covers 1 TR. All weights in this model (colored blue) are learned by backpropagating the mean squared error losses from univariately predicting the BOLD activity of the observed voxels influenced by the weight.

ers that produce $\mathbf{N}$. $\mathbf{N}[i,:,:,:]$, a single spatio-temporal feature map encodes the responsiveness to a local characteristic of the input, such as a specifically oriented edge or coherent motion. Such specific local features arise within the NIF model during training it on brain activity. 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 storing the activations of all the columns that respond to the same spatial location. As in DCM, the coupling of regions is a choice of the experimenter, and can be neural or behavioural in nature. Specifics of the neural network architecture are a design choice, allowing feed-forward (only modeling nonlinear transformations of the input), or recurrent (modeling lateral and top-down interactions) regions. There is a natural limit on the complexity (number of free parameters) of the neural architecture as more complex models have increasing difficulties to learn accurate generative models for the neural data.

Modeling information flow

NIF models assume that neural computations in a region operate on its afferent inputs that reflect either sensory input or neural activity in other brain regions. Effective connectivity from a source region a to a sink region b is expressed as a convolution between neural tensor $\mathbf{N}_a$ and a tensor of synaptic weights $\mathbf{W}_{a \to b}$. In other words, here the flow of cognitive information is modeled using ND convolutions on afferent inputs. Equation (1) shows an example afferent input model with two spatial coordinates x and y, one temporal coordinate t and the channel index c_{in}. $\mathbf{W}$ thus contains spatio-temporal receptive fields of the succeeding cortical columns. A 3D convolution is performed on the spatial dimensions and the temporal dimension as follows:

$$\begin{aligned}(\mathbf{N}_a \star \mathbf{W}_{a \to b})[c_{\text{out}}, x, y, t] = \\ \sum_{c_{\text{in}}, dx, dy, dt} \mathbf{N}_a[c_{\text{in}}, x - dx, y - dy, t - dt] \\ \mathbf{W}_{a \to b}[c_{\text{in}}, dx, dy, dt, c_{\text{out}}] \; .\end{aligned} \tag{1}$$

The information processing representation (activation) encoded by the j-th brain area is then a function of its afferent input from brain area i:

$$\mathbf{N}_j = \boldsymbol{f}\left(\sum_{i=1}^{N} \mathbf{N}_i \star \mathbf{W}_{i \to j} + \mathbf{B}_j\right). \tag{2}$$

Here, $\boldsymbol{f}$ is a nonlinear neural network activation function (applied element-wisely) and $\mathbf{B}_j$ determines the bias. Using this setup, we can model how neural populations respond to sensory input, as well as to each other.

Observation models

To estimate the parameters of NIF models they are linked to (predicting) brain activity measurements or behaviour (e.g. motor behaviour). This prediction is done across a set of learned weight vectors reading out the tensor $\mathbf{N}$ dimensions in the framework of tensor decomposition (or, to be computationally feasible, its low rank approximation). In the NIF modeling example outlined here we learn to predict brain activity from functional magnetic resonance imaging (fMRI) in response to video stimuli. Let $\mathbf{Y} \in \mathbb{R}^{K \times T}$ denote region-specific BOLD responses of K voxels acquired over T time points. Our observation model is defined by

$$\mathbf{Y}[k,t+\Delta_t] = \sum_{c,x,y,t} \mathbf{N}[c,x,y,t]\boldsymbol{W}[c,x,y,t,k] + \varepsilon[k]\,, \quad (3)$$

where $\varepsilon[k]$ is normally distributed measurement noise and Δ_t is a fixed temporal shift that sets a minimum BOLD signal delay. To reduce the computational load of parameter estimation, we use a factorized low-rank decomposition of $\boldsymbol{W}$. That is,

$$\boldsymbol{W}[c,x,y,t,k] \approx \boldsymbol{W}_c[c,k]\boldsymbol{W}_t[t,k] \sum_{r=1}^{R} \boldsymbol{W}_{x,r}[x,k]\boldsymbol{W}_{y,r}[y,k]\,. \quad (4)$$

Here, $\boldsymbol{W}_c[c,k]$ is the channel receptive field, $\boldsymbol{W}_{r,x}[x,k]$ and $\boldsymbol{W}_{r,y}[y,k]$ are r spatial receptive fields (where R is the target rank) and $\boldsymbol{W}_t[t,k]$ is the temporal receptive field of the k-th voxel. In our example the rank R is set to be 4 to allow for more complex (e.g. diagonal) receptive fields. $\boldsymbol{W}_{r,x}[x,k]$, $\boldsymbol{W}_{r,y}[y,k]$ and $\boldsymbol{W}_t[t,k]$ are constrained to be positive by applying a softmax nonlinearity across the voxel-specific vector. The voxel-specific channel observation model $\boldsymbol{W}_c[c,k]$ learns sensitivity of a voxel to specific feature channels, and can weight each channel's contribution positively or negatively. The estimated voxel-specific observation models are interpretable. In our example, the spatial weight vectors can be interpreted as the population receptive field of a voxel. The temporal weight vector can model voxel-wise differences in the hemodynamic response function.

The NIF example model presented here is illustrated and described in Figure 1. To demonstrate the NIF framework within reasonable computational load we downsampled the spatial video size to 112×112 pixels for training and transformed them to grayscale. The model trained for 8 epochs.

Functional MRI data

The example model is trained on 3T whole-brain functional magnetic resonance imaging (fMRI) data from a single human participant (male, age 27.5 years) exposed to 23.3 h of spatiotemporal and auditory naturalistic stimuli (episodes of BBC's *Doctor Who* (Davies, Gardner, Moffat, Young, & Collinson, 2005)). We collected data in a Siemens 3T MAGNETOM Prisma system inside a 32-channel head coil (Siemens, Erlangen, Germany). A T2*-weighted echo planar imaging pulse sequence at a TR of 700 ms 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). We used a multiband-multi-echo protocol with multiband acceleration factor of 8, TE of 39 ms and a flip angle of 75 degrees. Following common procedures in visual neural encoding studies we use a large set of fMRI data from stimuli that have been presented once for training (119.225 volumes), and a short resampled test set on which we estimate quantitative predictive model performance (1.031 volumes). The video episodes were presented on a rear-projection screen with the Presentation software package, cropped to 698×732 pixels squares so that they covered 20° 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. 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.

Results and observations

The trained NIF model can be analysed in multiple ways. We focus on showing that it has learned meaningful and known properties of the visual system.

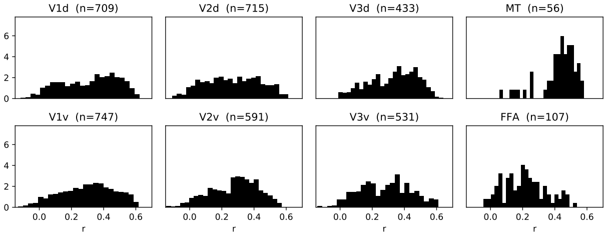


Figure 2: Histograms (normalized) of voxel-wise correlations between predicted and observed BOLD responses on the test set in different observed brain regions.

Figure 2 shows the correlations between predicted and observed brain activity on the test set for the voxels of all ROIs. We can see that the NIF model indeed generates realistic brain activity in response to unseen input stimuli (out of sample prediction), i.e. the objective of the model training has been reached. We now focus on what the model has learned within the weight matrices.

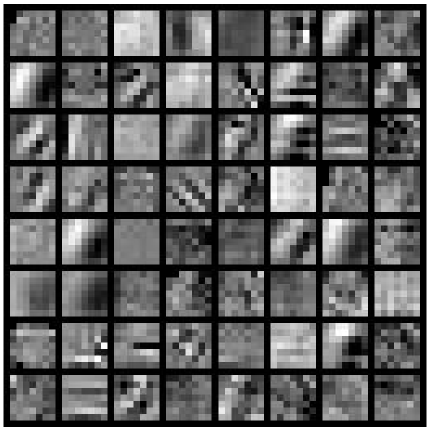


Figure 3: Frame 3 of the spatio-temporal channel weights of the V1 convolutional layer operating on the video input stimuli, learned on neural data alone. For visualization, weights are clipped at the extremes and rescaled between 0 and 1.

Figure 3 shows the 64 channels (feature detectors) learned within the tensor connected to V1 voxels. We see that within

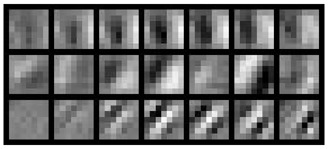


Figure 4: Examples of learned feature detectors over time. Many of the channels learned from brain activity show temporal variety.

this principled and data driven method typically known feature detection mechanisms of V1 arise, namely Gabor feature detectors. Alongside with them many feature detectors that can not be formulated as Gabors (but are reminiscent of those arising within object recognition convolutional networks) are learned. Several of these feature detectors show temporal variance. Figure 4 shows all frames of 3 of the temporally variant feature detectors.

Retinotopy

The voxel-specific matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ learn the spatial receptive fields of every voxel. Their outer product shows the location of this receptive field on the current input space (e.g. video space, feature map). A sensibly trained NIF model should be expected to have learned basic retinotopy.

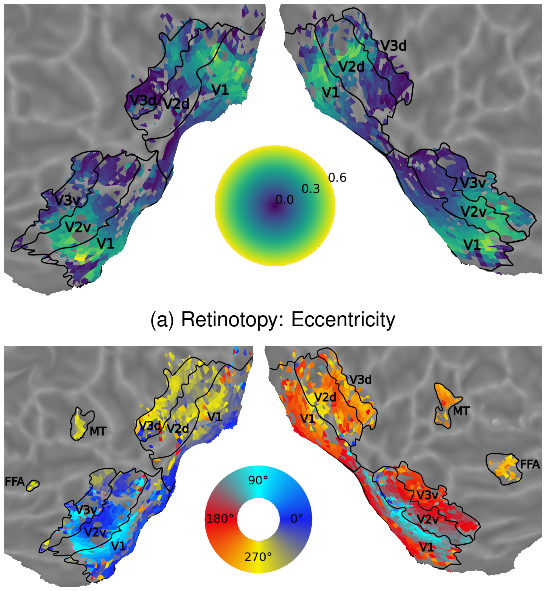
We determined the center of mass of the voxel-wise spatial receptive fields and transformed them to polar coordinates with the fixation point used as the fovea center. Eccentricity and polar angle are projected onto cortical flat maps, as generated by `pycortex` (Gao, Huth, Lescroart, & Gallant, 2015), in Figure 5. The annotated ROIs for V1, V2 and V3 have been estimated with classical wedge and ring retinotopy. It becomes clear that reversal boundaries align well with the traditionally estimated ROI boundaries. The eccentricity matches the expected fovea-periphery organization as well. The NIF model thus learned sensible retinotopic characteristics of the visual system.

Discussion and conclusions

To the best of our knowledge we have demonstrated for the first time that biologically meaningful neural information processing systems can be estimated directly from neural data from naturalistic stimulation. NIF modeling provides us with a principled approach to make sense of the high-resolution large datasets that will be produced by continuing advances in neurotechnology (Stevenson & Kording, 2011). NIF modeling allows neuroscientists to specify hypotheses about neuronal interactions and test these by quantifying how well the resulting models explain observed measurements. We expect that (variants of) NIF models will provide new insights into the principles and mechanisms that dictate neural information processing in biological systems.

Acknowledgments

This research was supported by VIDI grant number 639.072.513 of The Netherlands Organization for Scientific Research (NWO).



(a) Retinotopy: Eccentricity

(b) Retinotopy: Polar angle

Figure 5: Retinotopy of significantly predictable voxels that arises in the voxel-specific spatial observation matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ within the NIF model training example. There is a good fit between the classical retinotopy and retinotopy estimated directly from naturalistic spatiotemporal stimuli using the NIF model.

References

Churchland, P. S., & Sejnowski, T. J. (1992). *The Computational Brain*. MIT Press.

Davies, R. T., Gardner, J., Moffat, S., Young, M., & Collinson, P. (2005). *Doctor Who.* BBC.

Friston, K., Harrison, L., & Penny, W. (2003, aug). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, *9*, 23.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005–10014.

Haak, K. V., Winawer, J., Harvey, B. M., Renken, R., Dumoulin, S. O., Wandell, B. A., & Cornelissen, F. W. (2013). Connective field modeling. *NeuroImage*, *66*, 376–384.

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

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Stanley, G. B. (2005). Neural System Identification. In *Neural engineering* (pp. 367–388). Kluwer Academic/Plenum Publishers.

Stevenson, I. H., & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature Neuroscience*, *14*(2), 139–142.

van Gerven, M. A. J. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, *76*, 172–183.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*(1), 477–505.

Yamins, D. L. K., 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. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.