Neural System Identification with Cortical Information Flow

L. Ambrogioni*, K. Seeliger*, U. Güçlü, M. A. J. van Gerven

Donders Institute for Brain, Cognition and Behaviour
Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Abstract

Cortical information flow (CIF) is a new framework for system identification in neuroscience. CIF models represent neural systems as coupled brain regions that each embody neural computations. These brain regions are coupled to observed data specific to that region. Neural computations are estimated via stochastic gradient descent. We show using a large-scale fMRI dataset that, in this manner, we can estimate models that learn meaningful neural computations. Our framework is general in the sense that it can be used in conjunction with any (combination of) neural recording techniques. It is also scalable, providing neuroscientists with a principled approach to make sense of the high-dimensional neural datasets.

1 Introduction

A major goal in neuroscience is to uncover the neural computations that subserve behaviour and cognition (Churchland and Sejnowski, 1992). Arguably, a true understanding of the brain demands the development of computational models that replicate neural information processing in biological systems. After all, as Feynman famously stated: "What I cannot create, I do not understand." One way to approach this goal is to estimate neural information processing systems from observed measurements, which we refer to as neural system identification (Stanley, 2005; Wu et al., 2006). However, so far, frameworks to accurately recover neural computations from observed data remain wanting.

Encoding models provide part of the solution as they attempt to uncover which stimulus features drive neural responses in specific brain regions (Naselaris et al., 2011; van Gerven, 2017). However, current encoding models ignore the information flow between brain regions. Furthermore, they are often only indirectly linked to neural data. E.g., by using the features estimated by neural networks trained on some pattern recognition task as the basis set with which to predict neural responses (Güçlü and van Gerven, 2015; Yamins et al., 2014). On the other hand, techniques like dynamic causal modelling (DCM) aim to uncover effective connectivity between brain regions by estimating interactions from neural data (Friston et al., 2003). However, DCM does not address what kind of computations drive the interaction between brain regions.

In this paper, we combine the virtues of encoding models and effective connectivity methods. We define neural information processing system whose parameters can be estimated directly from neural data. The resulting generative models, which we refer to as cortical information flow (CIF) models, can be interpreted as synthetic brains that capture the computations that take place in real brains. CIF estimates nonlinear computations in individual brain regions in an end-to-end manner. We show that CIF models estimated from large neural datasets collected under naturalistic stimulation explain observed brain measurements. Moreover, the emerging neural computations are biologically meaningful, as demonstrated by qualitative and quantitative analyses.

The present paper outlines the basic principles of CIF models. However, the framework is general in the sense that the experimenter is free to choose the neural architecture of individual brain regions and how these regions map onto observed measurements which can be either neural or behavioural in nature. The

*Equal contribution.

philosphy of cortical information flow is outlined in Figure 1. Given the generality of our framework, we expect that they and their offspring will guide the development of a new family of generative models that allow us to uncover the principles of neural computations in biological systems.

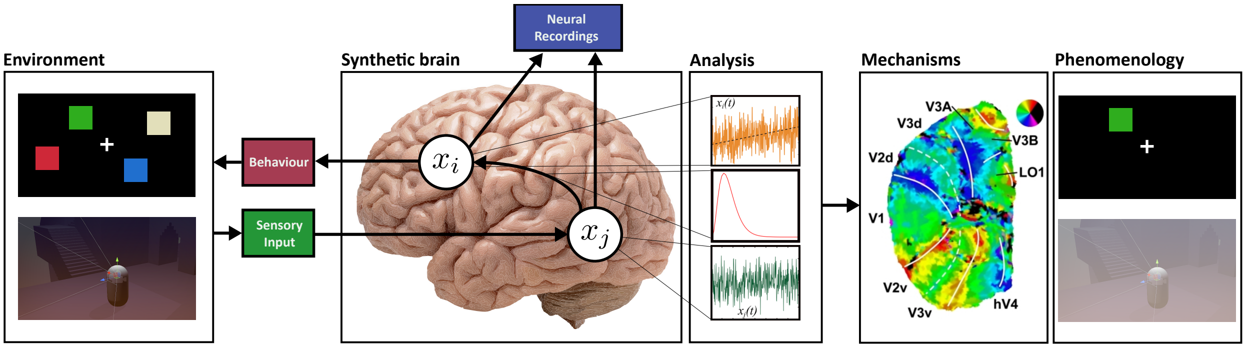

Figure 1: The philosophy underlying cortical information flow. CIF models define synthetic brains that model information processing in real brains. They are specified in terms of mutually interacting neuronal populations (white discs) that receive sensory input (green) and give rise to measurements of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

2 Cortical information flow

This section describes the technical details of cortical information flow. The core of CIF models is given by tensors that encode the activity of a cortical area. Each of these tensors is coupled to the observed neural response of a cortical area through a low-rank tensor operator that is factorized into the tensor product of a spatial receptive field and a feature map. These operators encode the population receptive fields and the feature extraction properties of each unit of observation (voxels in our case). Conversely, the activity of the entries of the underlying tensor represent the population activity in a cortical area.

2.1 Tensor representation of cortical areas

The basic building blocks of CIF are tensors that represent the activity of cortical areas. Tensors are mathematical objects that generalize matrices to an arbitrary number of array dimensions. Like matrices, tensors can be used both for storing data and encoding (linear) transformations. The activity of a cortical area is encoded in a four-dimensional (neural) tensor $\boldsymbol{N}$, whose array dimensions represent respectively channels, two retinal spatial coordinates and time. The resulting feature maps encode the responsiveness to local characteristics of the input such as oriented edges. 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.

2.2 Modeling information flow

The mammalian cortex is organized in terms of several functional regions. These regions implement nonlinear transformation of their input, which can be either sensations generated by the environment or information received from other brain regions. This implies that the responses of brain regions should be modelled by taking the afferent inputs to that brain region into account (Haak et al., 2013).

CIF models express effective connectivity from region a to region b as a convolution between the neural tensor $\boldsymbol{N}_b$ and a tensor of synaptic weights $\boldsymbol{W}_{a\to b}$:

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

As shown in Equation (1), the tensor $\boldsymbol{W}$ has five array dimensions: input channels, two spatial coordinates, one temporal coordinate and output channel. $\boldsymbol{W}$ represents the spatiotemporal receptive fields of individual cortical coolumns. The convolution is performed along the two spatial dimensions and the temporal dimension. The output feature maps are obtained as weighted averages of the input feature maps. For example, an output feature encoding a corner can be obtained by averaging the response of two edge features. Using this notation, we can define the activation of the j-th brain area as a function of the afferent input:

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

where $\mathbf{f}$ is a sigmoid activation function (applied element-wise) and $\boldsymbol{B}_j$ is a bias term.

2.3 Observation models

Our model of cortical information flow can be coupled to an arbitrary number of heterogeneous measurements using tensor operators. In the present paper, we measure brain activity using functional magnetic resonance imaging (fMRI). Consequently, the measurement operators map each tensor to the blood-oxygenation-level dependent (BOLD) response of each corresponding brain region. We use a factorized rank-one decomposition to represent how neural activity maps to observed measurements. Let $\mathbf{b}$ be a vector of BOLD responses. Our observation model can be written as follows:

$$b[j] = \sum_{c,x,y,t} \boldsymbol{N}[c, x, y, t]\mathbf{W}_c[c, j]\mathbf{W}_x[x, j]\mathbf{W}_y[y, j]\mathbf{W}_t[t, j] + \epsilon[j] \tag{3}$$

where $\mathbf{W}_c[c, j]$ is the channel receptive field, $\mathbf{W}_x[x, j]$ and $\mathbf{W}_y[y, j]$ are the spatial receptive fields and $\mathbf{W}_t[t, j]$ is the temporal receptive field of the j-th voxel. Finally, $\epsilon[j]$ is normally distributed measurement noise.

2.4 Model estimation

Once the structure of the CIF model is defined, its parameters (synaptic weights and observation model) can be estimated using stochastic gradient descent (SGD), also known as backpropagation (Mandt et al., 2017). Losses are estimated within brain regions as mean squared error across voxels. The individual loss terms for every brain region are summed to obtain a final loss, which is minimized using SGD. CIF was implemented in the `chainer` framework for neural networks (Tokui et al., 2015).

3 Experimental validation

The cortical flow approach will be illustrated with an example in visual neuroscience. We use dorsal and ventral visual stream data from a long-term exposure of a single participant to spatiotemporal (video) stimuli in the MRI scanner.

3.1 Model specification

For the purpose of demonstration, we used a purely feed-forward architecture modeling the ventral (V1v, V2v, V3v) and dorsal (V1d, V2d, V3d) visual streams. The two streams are connected to their sensory input

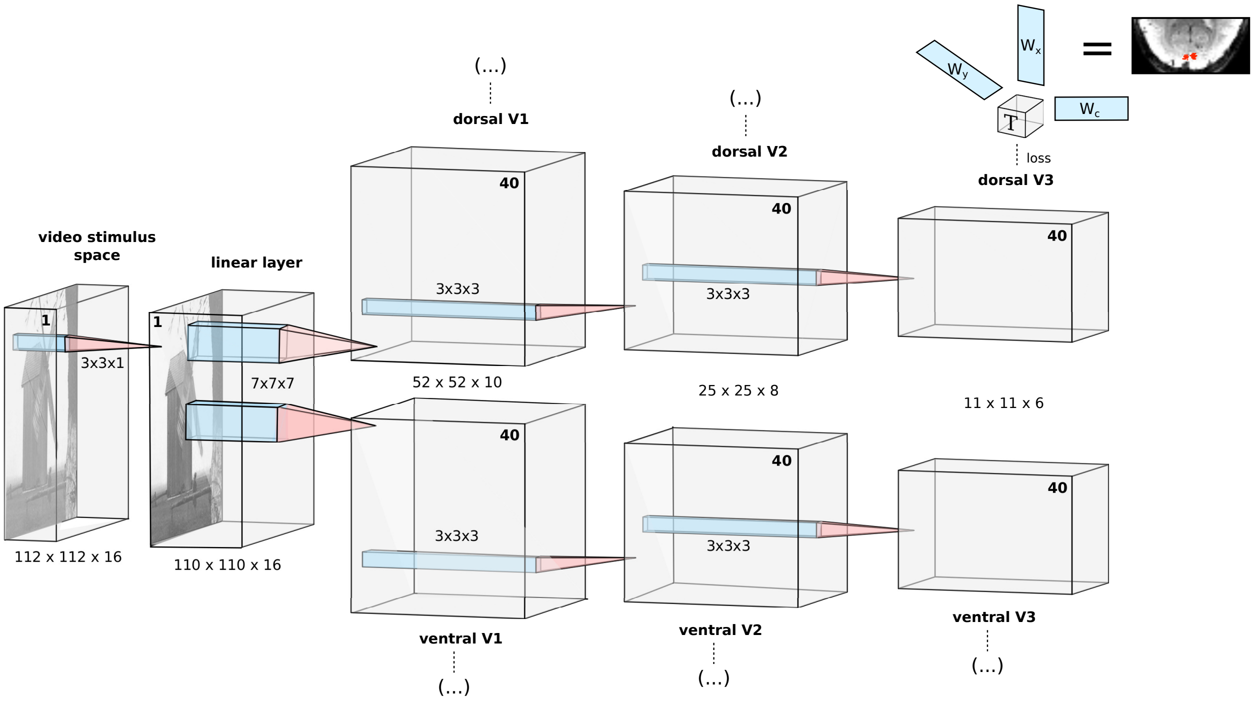

Figure 2: Architecture of the example model built around the dorsal and ventral parts of the early visual system. Between the separate tensors resulting from the 3D convolution operations we state the size of each input space ($x \times y \times t$) to the next layer. We use 3D video patches consisting of 16 frames since this covers one TR of 700 ms at 23.86 Hz. 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. This layer serves as learnable linear preprocessing step. The result of this layer then passes into the separate dorsal and ventral visual streams of our example model. Convolutional kernel sizes are $7 \times 7 \times 7$ in the first convolutional layer (leading to the dorsal or ventral V1 tensor), and $3 \times 3 \times 3$ for all other layers. The temporal dimension t is preserved across the model and only diminished by the temporal convolutions. It is collapsed with an average pooling operation before the resulting tensor enters the observation model (i.e., we do not make use of $\mathbf{W}_t$ here). After every convolution operation we apply a sigmoid nonlinearity and spatial average pooling with 2×2 kernels. Model losses for updating all model weights (colored blue) are directly estimated on the measured BOLD brain activity.

using a linear layer consisting of a single 3×3 convolutional kernel, representing processing in the lateral geniculate nucleus. The architectural details of our CIF model are described in Figure 2. Three-dimensional video patches serve as the input to the network. The objective is to predict the corresponding voxel activity for all voxels. For the video stimulus domain, our model should learn spatiotemporal receptive fields for each brain region that jointly explain the voxel responses within that region.

The model was trained for 15 epochs with a batch size of 4, using the Adam optimizer (Kingma and Ba, 2014) with adjusted learning rate $\alpha = 0.0005$. Weights were initialized with Gaussian distributions scaled by the number of feature maps in every layer (He et al., 2015).

3.2 Stimulus material

The participant watched episodes from seasons 2 to 4 after the 2005 relaunch of the series (*Tenth Doctor*). These episodes were novel to the participant and presented once. Episodes were split into 12 min chunks (with each last one having varying length) and presented with a short break after two runs. After most full episodes we additionally recorded the same randomly permuted selection of 7 short movies (*Pond Life* and *Space / Time*, taken from the series' next incarnation to avoid overlap), leaving us with 22 (*Space / Time*) and 26 (*Pond Life*) repetitions of the same test stimulus. This data was averaged to form a resampled, thus noise-reduced test set, which is a common setup in representation or encoding model experiments. In our analysis we omitted season 2 episode 7 (`run005`) as the effect of video buffering problems occurring in the second half of its presentation (see full data description) on the stimulus is unclear. Thus we used data from 30 episodes. We also omitted the BOLD fade-out volumes from a static black blank screen shown for 16 s at the end of half the training and all test set acquisition runs. This left us with 119.225 whole-brain volumes of single-presentation data, forming our training set; and 1.031 volumes of resampled data, forming our test set.

As there were small differences between frame rates in the train and test sets we re-encoded the stimulus videos to a uniform frame rate of 23.86 Hz for training the example model (original training set: 23.98 Hz, test set: 25 Hz except for one video with 30 Hz). This specific frame rate was chosen in order to be able to extract exactly 16 frames for every TR of 700 ms. To increase the capacity of the model we reduced the video size to 112×112. Along with 16 frames this leads to an input size of $112 \times 112 \times 16$. These choices allow us to demonstrate the framework within reasonable computational load. To further increase the expressiveness of the model (avoiding that the model has to balance or omit colour information) we transformed all videos to gray-scale. Otherwise the stimulus video was left in its original state. E.g., we did not apply other preprocessing common for visual recognition networks, such as demeaning.

3.3 Data acquisition

We used data from a single human participant (male, age 27.5 years) exposed to 23.3 h of spatio-temporal and auditory naturalistic stimuli (episodes of BBC's *Doctor Who* (Davies et al., 2005)) over a period of six months. The data set, its recording procedure and a data quality analysis have been described in detail in a separate manuscript. Here we focus on the necessary information to understand the model example. The BOLD data will be made publicly available through `data.donders.ru.nl`[1]. Data collection was approved by the local ethical review board (CMO regio Arnhem-Nijmegen, The Netherlands, CMO code 2014-288) and was carried out in accordance with the approved guidelines. An amendment of the general ethical approval was acquired due to the extraordinary length of this experiment (NL45659.091.14). Informed consent was given by the participant for every separate session.

Data was collected in a Siemens 3T MAGNETOM Prisma system inside a Siemens 32-channel head coil (Siemens, Erlangen, Germany). We used a T2*-weighted echo planar imaging pulse sequence at a TR of 700 ms. Functional scans comprised 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.

Stimuli were presented on a rear-projection screen using an Eiki projector with a resolution of 1024×768 pixels, placed outside the shielded room. A stimulus PC equipped with the Presentation software package was used to display the stimuli. Stimulus videos were resized and cropped to squares so that they covered 20° of the visual field, leading to a video presentation size of 698×732 pixels. The participant's head position was stabilized within and across sessions with a custom-made MRI-compatible headcast, along with further measures. The participant had to fixate on a fixation cross displayed at the center of the presentation screen. During pilot experiments, the subject practiced maintaining his position and attention fixed for a prolonged time while fixating on a highly salient stimulus.

3.4 Data preprocessing

Minimal BOLD data preprocessing was applied using `FSL v5.0`. We first realigned all volumes within each 12 min run to their middle volume (reference volume). We then estimated the transformation of each reference volume to the middle (reference) volume from the very first run in the series (the first episode).

[1] `DOI: 11633/di.dcc.DSC_2018.00082_134`

This estimated transformation was applied to all training and test volumes, leading to all volumes realigned to the middle volume of the very first run. Note that we did not apply further preprocessing steps. We omitted slice time correction due to our fast multiband protocol.

We applied further common linear preprocessing steps to the individual voxel time series to facilitate learning. The signal of every voxel used in the model was linearly detrended and then standardized (demeaned and moved to unit variance); with statistics collected voxel-wisely within each train or test run of approximately 1000 volumes. BOLD data for the test set was then averaged, with a final standardisation step after the averaging.

To keep the architecture as barebones as possible, we use a fixed delay of 5 TR to associate stimulus video chunks with responses. The peak of the BOLD signal corresponding to a stimulus is thus expected to occur within a time window of 3500 ms to 4200 ms after the signal.

3.5 Functional localizers

To isolate regions of interest, we recorded functional localizer data in additional sessions. We used polar wedges to perform retinotopy, using `FSL v5.0` for analysis. We mapped V1, V2 and V3 for left and right hemisphere and separately for the dorsal and ventral streams (separated by the calcarine sulcus). A number of functional localizers were estimated with specific experiments, analyzed as contrasts in `FSL`. The functional localizer results were verified by comparing with previous experimental findings (Yarkoni et al., 2011).

4 Results

We now analyze the performance and characteristics of the estimated CIF model.

4.1 Response prediction

After estimating the CIF model, we observed that BOLD responses in each brain region could be predicted well by the model, as illustrated in Figure 3. This implies that neural computations along the visual hierarchy can be directly estimated in an end-to-end manner from observed neural data. This is fundamentally different from earlier approaches, that use predefined feature representations to predict region-specific neural responses.

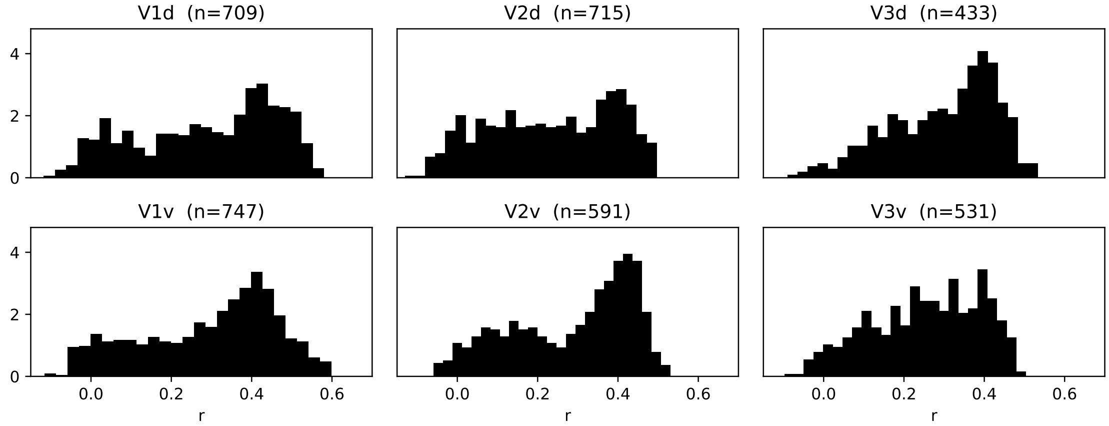

Figure 3: Histograms (normalized) of the correlations between predicted and observed BOLD responses on the test set in different observed brain regions.

4.2 Characterizing neural computations

Ultimately, we are interested in understanding the neural computations that are being performed in individual brain regions. Here we resorted to a basic analysis by visualizing the spatial receptive fields (weights) of channels in areas V1v and V1d. Figure 4 shows the receptive fields that have been learnt. Basic features reminiscent of Gabor wavelets seem to emerge as a function of learning. Gabor wavelets have been used to model the properties of neurons in early visual cortex (Daugman, 1985; Jones and Palmer, 1987). The representations encoded in downstream areas could in principle be visualized using basic decoding or deconvolution techniques (Schoenmakers et al., 2013; Zeiler and Fergus, 2012).

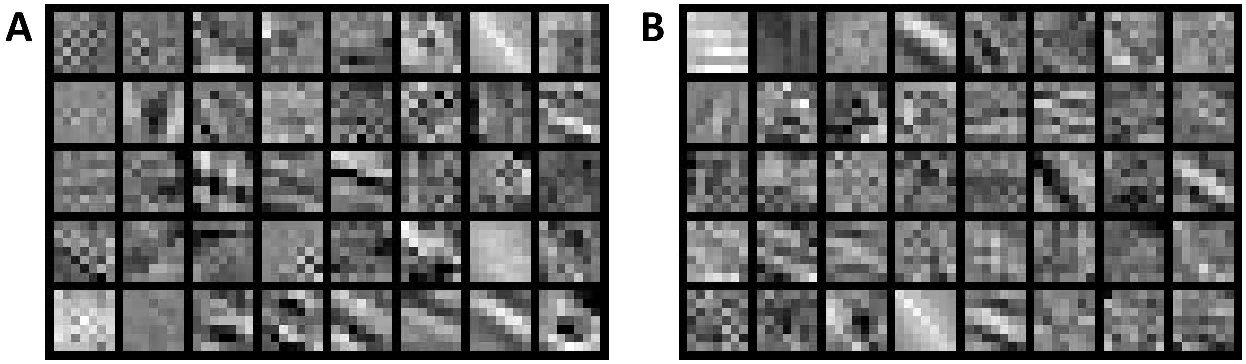

Figure 4: Spatio-temporal channel weights directly learned on neural data in the ventral (A) and dorsal (B) stream within our proposed framework. For visualization purposes, weights are clipped at the extremes and all weights are then globally moved between 0 and 1.

If we consider the variation of some of these receptive fields over time then it becomes evident that they pick up translation of contrast shift across the visual field (Figure 5). This makes sense from a functional point of view since these receptive fields also drive the subsequent computations in other downstream areas, including motion-sensitive areas such as area MT.

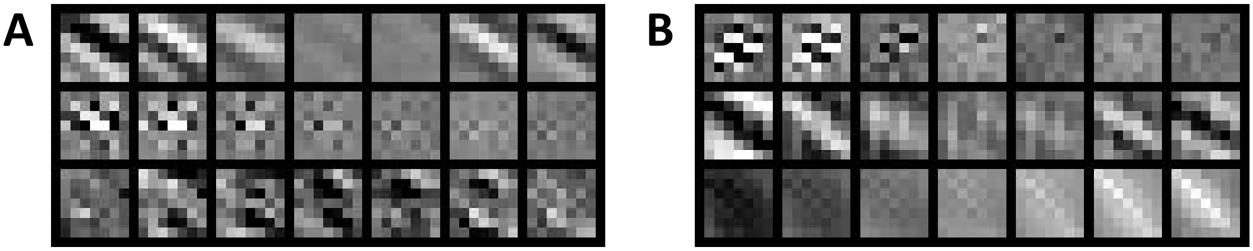

Figure 5: A subset of the channels learned from brain activity show regular translations in time in the ventral (A) and dorsal (B) stream.

We can also analyze the learned weight vectors from the tensor decomposition to check further properties of the model. For instance, taking the outer product between $\mathbf{W}_x$ and $\mathbf{W}_y$ will provide us with simple population receptive fields for every voxel, some of which are visualized in Figure 6. The prolonged symmetries are an artifact of the rank-1 decomposition and can be avoided with a more complex (e.g. higher rank) spatial observation model. Looking at some of these maps it becomes obvious that the model has learned localized spatial receptive fields.

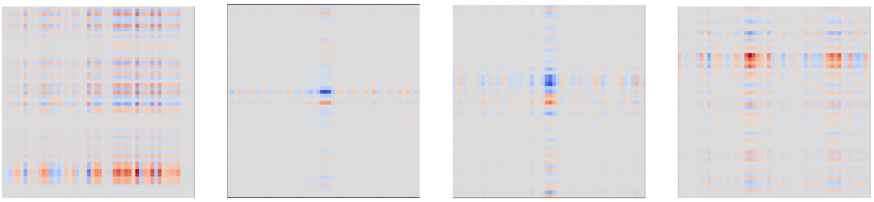

Figure 6: Dorsal V1 spatial receptive fields for four different voxels in video pixel space learned within our framework. Red regions utilize positive weights and blue regions negative ones.

5 Discussion

This paper introduces cortical information flow as a new approach for neural system identification. The approach relies on a neural architecture specified in terms of interacting brain regions that each embody nonlinear computations. By conditioning each brain region on associated measurements of neural activity, we can estimate neural information processing systems end-to-end. By allowing interactions between brain regions, each brain region will act as a regularizer for the neural computations that emerge in other brain regions. After all, the estimated neural computations must jointly explain observed responses across all brain regions.

We show using fMRI data collected during prolonged naturalistic stimulation that we can successfully predict BOLD responses across different brain regions. Furthermore, meaningful receptive fields emerged after model estimation. Importantly, the learnt 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 meaningful neural information processing systems can be estimated directly from neural data. CIF allows neuroscientists to specify hypotheses about neuronal interactions and test these by quantifying how well the resulting models explain observed measurements.

CIF generalizes current encoding models. For example, basic population receptive field models (Dumoulin and Wandell, 2008) and more advanced neural network models (van Gerven, 2017) are special cases of CIF that assume no interactions between brain regions and make specific choices for the nonlinear transformations that capture neuronal processing. Furthermore, current approaches mainly rely on using neural networks that are trained to solve another task such as object recognition (Güçlü and van Gerven, 2015; Yamins et al., 2014). An exception is, e.g. (Güçlü and van Gerven, 2016), but the employed models required the transformation of sensory input to lower-dimensional stimulus features and did not allow the explicit recovery of neural computations in individual regions of interest. Here we show that biologically-interpretable models can be estimated directly from neural data using cortical information flow.

The present work also provides a new approach to effective connectivity analysis. The researcher can specify alternative CIF models and then use explained variance as a model selection criterion. This is similar in spirit to dynamic causal modelling. However, instead of using changes in neural *dynamics* to estimate effective connectivity, we can embraces changes in neural *computation* to estimate causal interactions.

CIF can be naturally extended in several ways. 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 used to explicitly model the changes in neural dynamics that are now captured by 3D convolutions. Furthermore, lateral and feedback processes are easily added by adding additional links between brain regions. Also, the employed BOLD model can be easily replaced with more sophisticated (differentiable) models of the BOLD response such as a double gamma hemodynamic response function.

CIF models can also be extended to handle other data modalities. Alternative observation models can be formulated that allow us to infer neural computations from other measures of neural activity (e.g., single- and multi-unit recordings, local field potentials, calcium imaging, EEG, MEG). Moreover, CIF models can be conditioned on multiple heterogeneous datasets at the same time, providing an elegant solution for multimodal data fusion. Cortical flow can also be easily applied to other sensory inputs. For example, auditory areas can be conditioned on auditory input (see e.g. (Güçlü et al., 2016)). If this is combined with visual input then we may be able to uncover new properties of multimodal integration (Simanova et al., 2014).

Note that we are not restricted to conditioning CIF models on neural data. We may instead (or also) condition these models on behavioural data, such as motor responses or eye movements. The resulting models should then show the same behavioural responses as the system under study. We can even teach CIF models to solve the task at hand directly using reinforcement learning (Sutton and Barto, 2017). In this sense, CIF models provide a starting point for creating brain-inspired AI systems that more closely model how real brains solve cognitive tasks.

Estimated CIF models can be interpreted as synthetic brains that model their biological counterparts. This implies that we can subject them to any approach which can also be used to probe neural information processing in real brains. For instance, we can apply any method for neural data analysis to the neural time series that result from driving the model with external input. Recently developed nonlinear decoding techniques can shed further light on the neural representations that are encoded by different brain regions (Güçlütürk et al., 2017), providing insight into the phenomenological experience of synthetic brains. Here we restricted ourselves to demonstrating the virtues of our approach using basic receptive field analyses.

Finally, we can use CIF models as in-silico models to examine changes in neural computation. For example, we can examine how neural representations change during learning or by virtual lesioning of the network (Graziano and Aflalo, 2007). 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 approaches for understanding how brain stimulation modulates neural information processing, guiding the development of future neurotechnology (Roelfsema et al., 2018).

Summarizing, we view cortical information flow as a starting point for building a new family of rich, general, biologically-inspired computational models that capture neural information processing in biological systems. As such it provides a perfect blend of computational and experimental neuroscience (Churchland and Sejnowski, 2016). CIF models are also scalable since they make use of efficient stochastic gradient methods, as developed by the artificial intelligence community. This provides us with a principled approach to make sense of the high-resolution datasets produced by continuing advances in neurotechnology (Stevenson and Kording, 2011). We expect that (variants of) CIF models will provide exciting new insights into the principles and mechanisms that dictate neural information processing in biological systems.

Acknowledgements

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

References

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

Churchland, P. S. and Sejnowski, T. J. (2016). Blending computational and experimental neuroscience. *Nature Reviews Neuroscience*, 17(11):667–668.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169.

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

Dumoulin, S. O. and Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660.

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

Graziano, M. S. A. and Aflalo, T. N. (2007). Mapping Behavioral Repertoire onto the Cortex. *Neuron*, 56:239–251.

Güçlü, U., Thielen, J., Hanke, M., and van Gerven, M. A. J. (2016). Brains on beats. In *Neural Information Processing Systems*, pages 1–12.

Güçlü, U. and 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.

Güçlü, U. and van Gerven, M. A. J. (2016). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience*, pages 1–19.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., and van Gerven, M. A. J. (2017). Deep adversarial neural decoding. *Neural Information Processing Systems*, pages 1–12.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 1026–1034.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1233–1258.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mandt, S., Hoffman, M. D., and Blei, D. M. (2017). Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18(134):1–35.

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

Roelfsema, P. R., Denys, D., and Klink, P. C. (2018). Mind Reading and Writing: The Future of Neurotechnology. *Trends in Cognitive Sciences*, 22(7):1–13.

Schoenmakers, S., Barth, M., Heskes, T., and van Gerven, M. A. J. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961.

Simanova, I., Hagoort, P., Oostenveld, R., and van Gerven, M. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*, 24:426–434.

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

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

Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. The MIT Press, Boston, MA.

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: A next-generation open source framework for deep learning. In *Workshop on Machine Learning Systems (LearningSys) during Advances in Neural Information Processing Systems (NIPS) 2015*.

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

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

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and 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.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.

Zeiler, M. D. and Fergus, R. (2012). Visualizing and understanding convolutional networks. *arXiv preprint arXiv:1311.2901*.