Image-computable encoding models of human broadband iEEG responses to natural images reveal time-dependent representations

Ghislain St-Yves (gstyves@umn.edu)
Department of Neuroscience, University of Minnesota
Minneapolis, MN 55455 USA

Harvey Huang (Huang.Harvey@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Zeeshan Qadir (Qadir.Zeeshan@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Morgan Montoya (Montoya.Morgan@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Greg Worrell (Worrell.Gregory@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Kai J Miller (Miller.Kai@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Kendrick Kay (kay@umn.edu)
Center for Magnetic Resonance Research
Department of Radiology, University of Minnesota
Minneapolis, MN 55455 USA

Dora Hermes (Hermes.Dora@mayo.edu)
Department of Physiology and Biomedical Engineering, Mayo Clinic
Rochester, MN 55905 USA

Thomas Naselaris (nase0005@umn.edu)
Department of Neuroscience, University of Minnesota
Minneapolis, MN 55455 USA

Abstract

Image-computable models designed for prediction of fMRI BOLD signals have been shown to generalize well to iEEG broadband for simple stimuli. We show that models with complex feature spaces (DNNs) that have been used to predict natural image responses in fMRI signals can also be trained to predict iEEG broadband responses. The high temporal sampling afforded by iEEG signal enable us to precisely locate the onset of activity and characterize the temporal evolution of representational tuning at various recording sites. We show that the temporal onset is strongly correlated with network tuning depth across all subjects. Furthermore, we show that tuning properties in several channels vary over time after onset, with retinotopic representations subsiding and feature tuning drifting toward more semantic-like representations.

Keywords: Vision; Human iEEG; Encoding model, Natural scenes, Time-dependent representations

Introduction

Intracranial EEG (iEEG) recordings offer both high temporal sampling and reasonably high spatial resolution. It is customary to perform a time-frequency transformation of these signals in order to isolate specific spatiotemporally located recurrent brain processes. We refer to the power of the spectrogram region between 70 and 170 Hz as broadband (BB). BB responses are of particular interest as they have been related to local neuronal (input) firing rates (Miller, Sorensen, Ojemann, & den Nijs, 2009; Ray & Maunsell, 2011; Manning, Jacobs, Fried, & Kahana, 2009)

Most iEEG (eCoG and sEEG) studies focus on decoding few categories or specific stimulus properties and their change over time. While one can infer a lot about brains from the discriminative aspect of these signals, a predictive encoding model of the signal offers a possible account of the representations that produce that discrimination (Naselaris & Kay, 2015). When encoding models have been used, they have been limited to properties of simple stimuli (Hermes, Petridou, Kay, & Winawer, 2019). Large repertoires of natural image responses are required to train accurate complex models with large feature spaces, like NSD for fMRI (Allen et al., 2022). Natural images are required to stimulate the putative interactions between different representation stages in the visual cortex present during normal vision.

iEEG BB responses to simple stimuli have been shown to be well-predicted by encoding models designed for fMRI (Hermes et al., 2019). These models leverage the familiar concepts of receptive field (RF) and feature tuning (e.g. spatial frequency and orientation tuning) in order to characterize activity recorded at individual channels. However, this characterization is reflective of both sensory and cognitive signals that may have very different dynamics and may encode very different features. As a result, the representation of stimulus features at any one iEEG site may drift over the timecourse of a single stimulus presentation. Unlike fMRI, iEEG has the temporal resolution needed to investigate this possibility.

We designed a set of interpretable linear readout heads between a set of feature maps (from a deep neural network trained to categorize images) that permit smooth time-varying feature tuning. This allowed us to consider the question of how feature tuning vary over the timecourse of a single stimulus presentation. We report on two evidence for time-varying representations: time-varying retinotopic tuning and time-varying feature tuning.

Methods

A new iEEG natural image response dataset (NSD-iEEG, here partially presented) was recorded with subjects undergoing epilepsy treatment at Mayo Clinic. The data was preprocessed (re-referencing, time-frequency transformed) according to current best standards for iEEG data (Mercier et al., 2022). The subjects were presented with 1,000 distinct images (NSD's shared1000 (Allen et al., 2022)). 900 images were shown once while 100 images where shown 6 times. In the following, our encoding models were trained and tested with these 1,500 trial responses.

Our first goal is to predict the stimulus-onset-aligned BB log-power, $r_v(x_s,t)$, where v refers to one specific channel, x_s is the fixed stimulus presented during this recording epoch s (duration $T = 2.4s$) and $-0.8s < t < 1.6s$ is the time around stimulus onset at $t = 0$ whose presentation duration is $\tau = 0.8s$. In the following, we shall consider the **spatial**, **temporal** and **feature** parametrization of the linear model

$$\bar{r}_v(x_s,t) = \sum_{kji} W_{vkji}(t)\phi_{kji}(x_s) \tag{1}$$

Where the temporal dimension will be discretized at a later stage, but should be viewed as just another dimension on which model tuning can vary. $\phi_{kji}(x_s)$ is a generic features maps extractor e.g. one of the intermediary step along a DNN backbone (kji indexes features and 2d space, respectively). Here we used the pytorch implementation of Alexnet as a first and most common candidate feature extractor due to its convenient and manageable feature maps sizes. The linear tuning weights W are separated into a spatial RF and a low-rank (rank R) representation of feature and temporal tuning.

$$W_{vkji}(t) = s_v g_{vji} \sum_r^R w_{vk}^r H_v^r(t) \tag{2}$$

where the tuning functions g, w and H are all normalized, explaining the need for an additional scaling factor s_v specified for each modelled unit. g and H are further constrained to be positive valued via reparametrization.

We have also adapted the model to complex, multilayer, feature space along the lines of the feature-weighted receptive field (fwRF) model by rescaling and applying a consistent spatial RF at the multiple (different spatial resolution) layers of the DNN (St-Yves & Naselaris, 2018). While this complicates the description (which for the sake of brevity is not shown here), the resulting model still follows the same general principles described above.

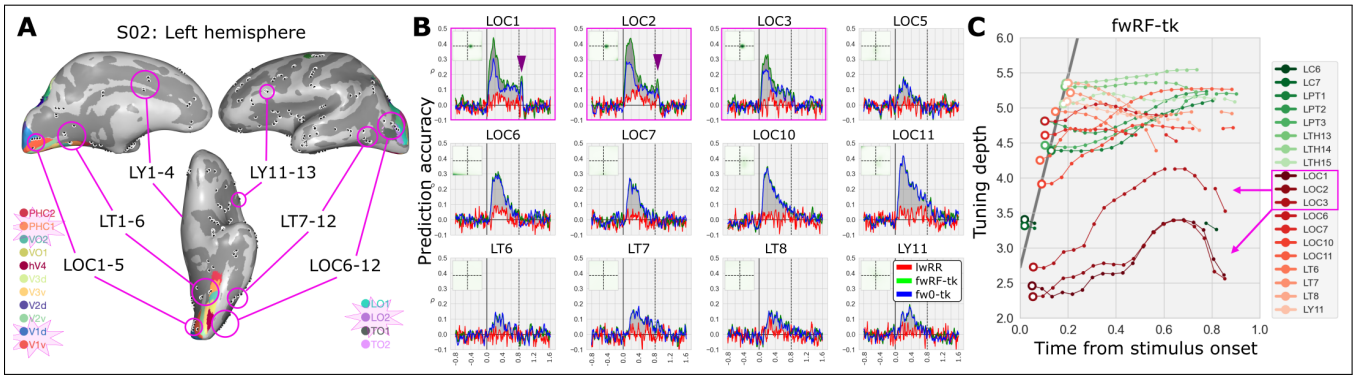

Figure 1: Distinct phases in representations in iEEG BB for various channels. A) Location of the electrodes in the left hemisphere for subject 2. We highlighted the locations of the most predictable electrodes described in B and C. Relevant visual ROI for this subject have also been highlighted. **B)** Retinotopically active channels (magenta highlight) show a clear difference in prediction accuracy (y-axis) between a fixed, flat, RF model (fw0-tk) and a tuned RF model (fwRF-tk, RF shown in inset) in the initial peak of predictability, but not in the latter tail present during stimulus presentation. Stimulus offset also induce a second phase of retinotopically specific activation (e.g. purple triangles in LOC1 and LOC2 channels) in some channels. Model accuracy is also compared to a reference layerwise ridge regression model (lwRR) were every time point is predicted independently. **C)** Feature tuning also vary over time. First, we note the clear positive correlation (gray line) between tuning depth (center of "mass" of the tuning weights on the network backbone of the encoding model) and channel response onset (empty colored circle symbols) relative to stimulus onset ($t = 0$). The chain of smaller dot attached to each onset circle shows the evolution of tuning depth in their respective regions of significant validation accuracy. Green and red indicate channels from subjects 1 and 2.

Model training (via stochastic gradient descent with $L2$ loss function) and validation is performed in a cross-validated manner using a k-fold procedure over the whole dataset. For each fold, a validation set (1-in-k of the overall data, $k = 15$ here) is chosen and the remaining data are randomly sampled ($N = 8$ times) into a training set for gradient estimation and a holdout set (20% of data) for early stopping criterion. Model parameters are frozen when the holdout set reaches minimum holdout loss. A prediction is then made for the validation set and averaged over all N training samples for that fold. Other interpretable properties of the model were averaged over all folds and samples.

Results

The trained model (fwRF-tk) significantly predicted BB responses in several channels (channel locations and validation accuracy for one subject shown in Fig 1A and B, respectively), and to a greater degree than a model with a fixed flat RF over the same feature maps (fw0-tk) in specific circumstances and channels and much greater than a model (lwRR) that predicted every time point independently (i.e. without time kernels) in all cases. Every predictable channel showed a tendency for relatively high predictability after onset latency followed by a gradual decrease, and sometime plateau.

Retinotopically active channels are identified by their robust and highly localized spatial tuning (Fig 1B, inset). These channels also showed very strong offset activity. Interestingly, our encoding model methodology reveals clear phases in the BB representation for retinotopically localized channels. Model differences between fwRF-tk and fw0-tk are highest after response onset for a period of roughly 400ms (and after stimulus offset in some cases) but virtually disappear afterward during the plateau of predictability. This indicates a clear change in the single-channel representation over time. This is, as far as we are aware, a new observation in the characterization of iEEG BB responses.

Our encoding model approach also reveals changes in the BB feature representation for non-retinotopically active channels (which were the majority here). This tuning drift can be illustrated by the concept of tuning depth. Tuning depth is, in effect, the center of mass of neural network layer depth weighted by the normalized encoding model feature tuning weights. Small depth values mean that the balance of feature tuning is on early network layers, and vice-versa. Tuning depth at response onset correlated very strongly with response onset latency (Fig 1C). That is, short latency was preferentially tuned to shallow network layers, while high latency tuned preferentially to deep layers. This is consistent with the typically conceived visual processing scaffold. However, we noticed that tuning depth varied over time after response onset, with a tendency for feature tuning to drift toward deeper layers.

This representational drift can occur for several reasons like the progressive recruitment of local neural circuitry or feedback effects from higher cortical areas. The result of this drift is that visual representations in all channels appear to be attracted over time toward more semantic-like representations. This kind of drift toward more semantic-like representations following a backward flow had been observed between V4 and IT in monkey electrophysiology and explain the ubiquity of semantic-like representation everywhere in cortex observed in fMRI (Sexton & Love, 2022).

It would be interesting to see if this change in representations can be accounted by dynamical model (like RNNs) with fixed tuning or what normative assumptions are required for such a model to express the type of representational changes observed here.

Acknowledgements

This work was supported by NIH R01EY023384 (T.N., K.K., D.H.). Collection of the dataset was supported by R01 EY035533 (K.K., D.H.)

Author contributions

T.N., K.K. and D.H. conceived of the project and designed the experiment. D.H., H.H., Z.Q., M.M., K.J.M. and G.W. were involved in data collection and curation and D.H. preprocessed the data. G.S.-Y. designed the encoding model, performed the data analysis and wrote the paper. T.N. and D.H. edited the paper and directed the overall project.

Competing interests

We declare no competing interests.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Hermes, D., Petridou, N., Kay, K. N., & Winawer, J. (2019). An image-computable model for the stimulus selectivity of gamma oscillations. *Elife*, *8*, e47035.

Manning, J. R., Jacobs, J., Fried, I., & Kahana, M. J. (2009). Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, *29*(43), 13613–13620.

Mercier, M. R., Dubarry, A.-S., Tadel, F., Avanzini, P., Axmacher, N., Cellier, D., ... others (2022). Advances in human intracranial electroencephalography research, guidelines and good practices. *Neuroimage*, *260*, 119438.

Miller, K. J., Sorensen, L. B., Ojemann, J. G., & den Nijs, M. (2009, 12). Power-law scaling in the brain surface electric potential. *PLOS Computational Biology*, *5*(12), 1-10.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of mvpa using explicit models of representation. *Trends in Cognitive Sciences*, *19*(10), 551 - 554.

Ray, S., & Maunsell, J. H. R. (2011, 04). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLOS Biology*, *9*(4), 1-15.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188 - 202. (New advances in encoding and decoding of brain signals)