img2fmri: a python package for predicting group-level fMRI responses to visual stimuli using deep neural networks

Maxwell Bennett

Department of Computer Science,
Columbia University

Christopher Baldassano

Department of Psychology,
Columbia University

Abstract

Here we introduce a new python package, img2fmri, to predict group-level fMRI responses to individual images. This prediction model uses an artificial deep neural network (DNN), as DNNs have been successful at predicting cortical responses in the human visual cortex when trained on real world visual categorization tasks. To validate our model, we predict fMRI responses to images our model has not previously seen from a new dataset. We then show how our frame-by-frame prediction model can be extended to a continuous visual stimulus by predicting an fMRI response to Pixar Animation Studio's short film Partly Cloudy. In analyzing the timepoint-timepoint similarity of our predicted fMRI response around human-annotated event boundaries in the movie, we find that our model outperforms the baseline model in describing the dynamics of the real fMRI response around these event boundaries, particularly in the timepoints just before and at an event. These analyses suggest that in visual areas of the brain, at least some of the temporal dynamics we see in the brain's processing of continuous, naturalistic stimuli can be explained by dynamics in the stimulus itself, since they can be predicted from our frame-by-frame model. All code, analyses, tutorials, and installation instructions can be found at https://github.com/dpmlab/img2fmri.

Recent research has shown that continuous and naturalistic visual stimuli, such as narrative movies, can evoke brain responses in high-level visual regions that are stable for timescales on the order of seconds to tens of seconds (Baldassano et al. 2017). What drives these temporal dynamics in the brain? One possible explanation is that high-level sensory regions have inherently slower dynamics, which are present even in the absence of a stimulus (Stephens et al. 2013) and in infants (Yates et al. 2022). Having slow-changing representations allows for the accumulation of information over time, providing context for interpreting the current stimulus based on information from the recent past (Honey et al. 2012; Hasson et al. 2015).

An alternative explanation for the long-timescale stability of high-level sensory regions could be that these temporal dynamics are, at least in part, due to dynamics inherent to the stimulus itself. Recent work by Heusser et al. (2021) sought to quantify the dynamics of content in a narrative movie by using topic modeling and hidden Markov models (HMMs) to discretize continuous stimulus into events characterized by their trajectories through semantic space. They found that the movie itself exhibited stable semantic events, suggesting that the event structure in brain responses could be "inherited" from the temporal structure of the stimulus, rather than arising from slow cortical dynamics. However, this approach relies on human annotations of stimulus content, rather than being directly derived from the stimulus alone.

Our approach to better understand this long-timescale stability in high-level sensory regions is to characterize the dynamics in brain responses driven purely by stimulus, using a frame-by-frame prediction model for visual cortex. Here we introduce a new python package, `img2fmri`, to predict group-level fMRI responses to individual images. This prediction model uses an artificial deep neural network (DNN), as DNNs have been successful at predicting cortical responses in the human visual cortex when trained on real world visual categorization tasks (Cichy et al. 2016). These neural networks learn to extract features (e.g. shapes, textures, eyes) from naturalistic visual data that allow them to accurately classify objects, animals, and scenes in the images they process (Olah et al. 2017), and can also be used to extract those predominant features from input to subsequently be

used in predicting cortical responses (Eickenberg et al. 2017). Research has also shown that the hierarchy of layers in a trained DNN can predict along a hierarchy of processing in the brain, where deeper, or higher, layers in a DNN best predict higher levels of cortical processing (Kell et al. 2018; Schrimpf et al. 2020). Our model is built by combining a pretrained ResNet-18 DNN with a linear regression model to predict fMRI responses to individual images. The mapping from DNN to the brain is fit using data from the open source BOLD5000 project (Chang et al. 2019), which includes fMRI responses for three subjects viewing 4916 unique images drawn from ImageNet (Deng et al. 2009), COCO (Lin et al. 2014), and SUN (Xiao et al. 2010). For each image, we predict activity patterns for five visual regions of interest (ROIs) for each subject's brain (defined by the BOLD5000 project): an early visual region of voxels near the calcarine sulcus sensitive to visual stimuli (EarlyVis), the lateral occipital complex (LOC), an object-selective region (Malach et al. 1995), as well as the occipital place area (OPA), retrosplenial cortex (RSC), and parahippocampal place area (PPA), three scene-selective regions (Epstein 2008).

To validate our model, we predict fMRI responses to images our model has not previously seen from the Twinset dataset (Mohsenzadeh et al. 2019). Our model can significantly predict fMRI responses at the group level for most categories in this dataset (Fig 1), and can significantly predict responses in every individual subject (Fig 2) († $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$, etc.).

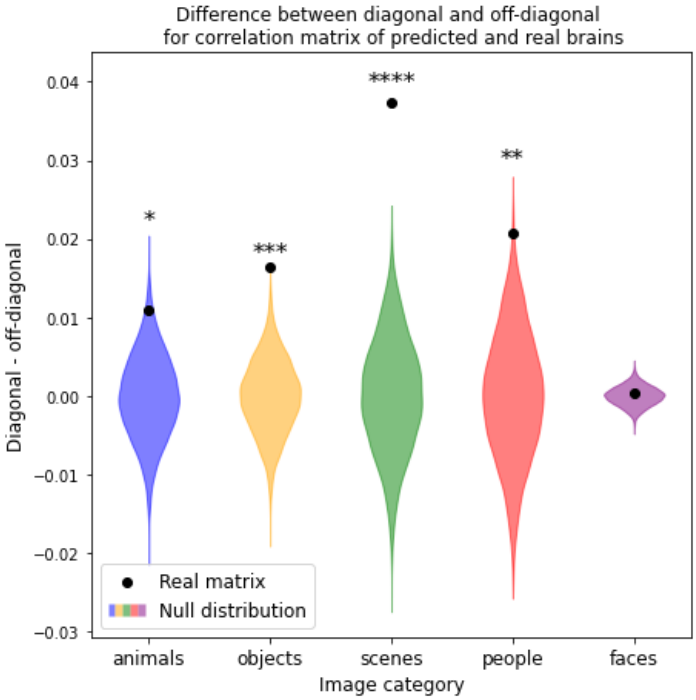

Figure 1: Per-category fMRI prediction permutation analysis. We predict fMRI responses for Twinset images split across five categories, and compare these predictions to averaged fMRI responses of 15 human subjects. For each category, we generate a null distribution by shuffling the rows of our correlation matrix of predicted and real fMRI responses and calculating the difference between the diagonal and off-diagonal values of each shuffle. We compare this null distribution to our aligned, unshuffled matrix, and see that our predictions perform well above chance in four of the five categories: animals (* p<0.05), objects (*** p<0.001), scenes (**** p<0.0001), and people (** p<0.01).

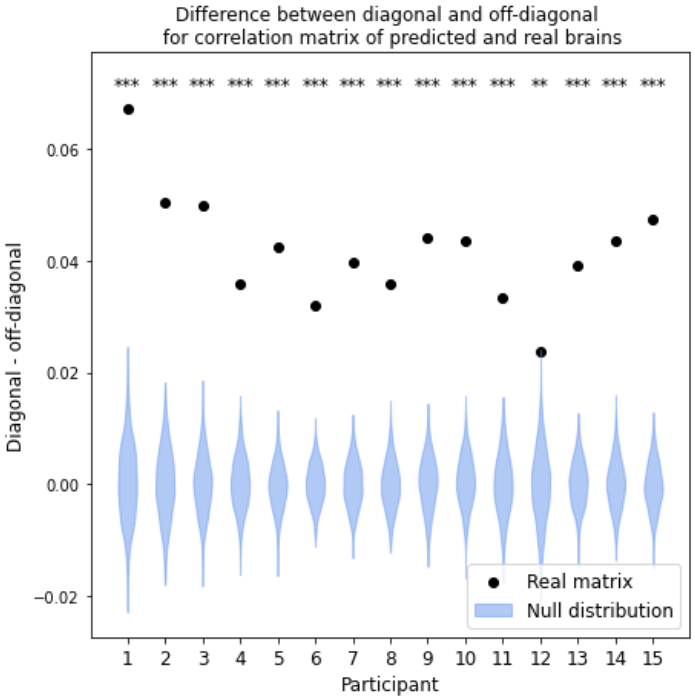

Figure 2: Per-participant fMRI prediction permutation analysis. We predict fMRI responses for Twinset images, and compare these predictions to fMRI responses of 15 human subjects. For each human subject, we generate a null distribution by shuffling the rows of our correlation matrix of predicted and real fMRI responses and calculating the difference between the diagonal and off-diagonal values of each shuffle. We compare this null distribution to our aligned, unshuffled matrix, and see that our predictions perform well above chance in all subjects (** p<0.01, *** p<0.001).

We then show how our frame-by-frame prediction model can be extended to a continuous visual stimulus by predicting an fMRI response to Pixar Animation Studio's short film *Partly Cloudy*. Here we compare the timepoint-timepoint similarity from our predicted frame-by-frame response to the timepoint-timepoint similarity in the actual group-averaged fMRI response, using data from the 33 adults in Richardson et al. (2018). For comparison, we also attempted to predict brain dynamics using a baseline model based only on the low-level luminance of the visual stimulus. In analyzing the

timepoint-timepoint similarity of our predicted fMRI response around human-annotated event boundaries in the movie (Fig 3), we find that our model outperforms the luminance model in describing the dynamics of the real fMRI response around these event boundaries, particularly in the timepoints just before an event, and at an event itself (* p<0.05, ** p<0.01). These analyses suggest that in visual areas of the brain, at least some of the temporal dynamics we see in the brain's processing of continuous, naturalistic stimuli can be explained by dynamics in the stimulus itself, since they can be predicted from our frame-by-frame model.

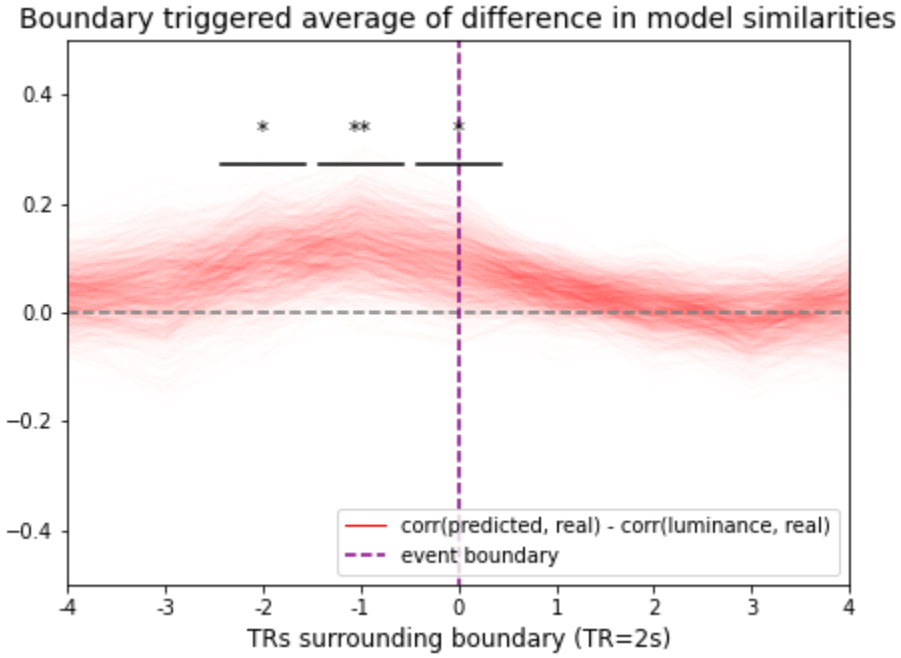

Figure 3: Boundary triggered average analysis for *Partly Cloudy*. We compare both the dynamics of our predicted response to *Partly Cloudy* and also the dynamics of a low-level luminance model to the dynamics of the averaged human fMRI response to the film, and take the difference around human-annotated boundaries to evaluate where our fMRI predictions better model the temporal dynamics seen in the human response to the film. We see that in the two TRs prior to an event boundary, and at the event boundary itself, our predicted brain response more closely models the event structure we see in the real fMRI response than the model based purely on the luminance of image frames (* p<0.05, ** p<0.01). For full explanation, code, and additional analyses, please see our accompanying Jupyter notebook: `overview.ipynb`.

All analyses, notebooks, and code can be found at https://github.com/dpmlab/img2fmri. The README in this repository outlines installation steps of background software, with the primary requirements being Python 3 or higher, PyTorch, and neuroimaging softwares AFNI and FSL. We also include a Dockerfile and docker image (via Docker Hub) for a pre-installed container. Our package is being released under the MIT license, and is also released as a pip/PyPI package, with API documentation available on ReadTheDocs.

To use the model and view these analyses, we encourage readers to explore our `overview.ipynb` in the previously linked github repository. For more information on the training of our model using the open source BOLD5000 dataset and pretrained ResNet-18 DNN, we have included a notebook `model_training.ipynb` within our `model_training` folder that outlines the model training process and offers suggestions for extending the model to predict fMRI responses from other feature-detecting models, and to other brain ROIs. Users can report any issues via GitHub issues, as outlined in `CONTRIBUTING.rst`, or by emailing the authors at mbb2176@columbia.edu.

References

1. Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. Discovering Event Structure in Continuous Narrative Perception and Memory [Internet]. Vol. 95, Neuron. Elsevier BV; 2017. p. 709-721.e5. Available from: http://dx.doi.org/10.1016/j.neuron.2017.06.041
2. Stephens GJ, Honey CJ, Hasson U. A place for time: the spatiotemporal structure of neural dynamics during natural audition [Internet]. Vol. 110, Journal of Neurophysiology. American Physiological Society; 2013. p. 2019–26. Available from: http://dx.doi.org/10.1152/jn.00268.2013
3. Yates TS, Skalaban LJ, Ellis CT, Bracher AJ, Baldassano C, Turk-Browne NB. Neural event segmentation of continuous experience in human infants [Internet]. Cold Spring Harbor Laboratory; 2021. Available from: http://dx.doi.org/10.1101/2021.06.16.448755
4. Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, Devinsky O, et al. Slow Cortical Dynamics and the Accumulation of Information over Long Timescales [Internet]. Vol. 76, Neuron. Elsevier BV; 2012. p. 423–34. Available from: http://dx.doi.org/10.1016/j.neuron.2012.08.011
5. Hasson U, Chen J, Honey CJ. Hierarchical process memory: memory as an integral component of information processing [Internet]. Vol. 19, Trends in Cognitive Sciences. Elsevier BV; 2015. p. 304–13. Available from: http://dx.doi.org/10.1016/j.tics.2015.04.006
6. Heusser AC, Fitzpatrick PC, Manning JR. Geometric models reveal behavioural and neural signatures of transforming experiences into memories [Internet]. Vol. 5, Nature Human Behaviour. Springer Science and Business Media LLC; 2021. p. 905–19. Available from: http://dx.doi.org/10.1038/s41562-021-01051-6
7. 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 [Internet]. Vol. 6, Scientific Reports. Springer Science and Business Media LLC; 2016. Available from: http://dx.doi.org/10.1038/srep27755
8. Olah C, Mordvintsev A, Schubert L. Feature Visualization [Internet]. Vol. 2, Distill. Distill Working Group; 2017. Available from: http://dx.doi.org/10.23915/distill.00007
9. Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system [Internet]. Vol. 152, NeuroImage. Elsevier BV; 2017. p. 184–94. Available from: http://dx.doi.org/10.1016/j.neuroimage.2016.10.001
10. Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy [Internet]. Vol. 98, Neuron. Elsevier BV; 2018. p. 630-644.e16. Available from: http://dx.doi.org/10.1016/j.neuron.2018.03.044
11. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? [Internet]. Cold Spring Harbor Laboratory; 2018. Available from: http://dx.doi.org/10.1101/407007
12. Chang N, Pyles JA, Marcus A, Gupta A, Tarr MJ, Aminoff EM. BOLD5000, a public fMRI dataset while viewing 5000 visual images [Internet]. Vol. 6, Scientific Data. Springer

Science and Business Media LLC; 2019. Available from: http://dx.doi.org/10.1038/s41597-019-0052-3
13. Deng J, Dong W, Socher R, Li LJ, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database [Internet]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. Available from: http://dx.doi.org/10.1109/CVPR.2009.5206848
14. Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, et al. Microsoft COCO: Common Objects in Context [Internet]. arXiv; 2014. Available from: https://arxiv.org/abs/1405.0312
15. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo [Internet]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2010. Available from: http://dx.doi.org/10.1109/CVPR.2010.5539970
16. Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, et al. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. [Internet]. Vol. 92, Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences; 1995. p. 8135–9. Available from: http://dx.doi.org/10.1073/pnas.92.18.8135
17. Epstein RA. Parahippocampal and retrosplenial contributions to human spatial navigation [Internet]. Vol. 12, Trends in Cognitive Sciences. Elsevier BV; 2008. p. 388–96. Available from: http://dx.doi.org/10.1016/j.tics.2008.07.004
18. Mohsenzadeh Y, Mullin C, Lahner B, Cichy R, Oliva A. Reliability and Generalizability of Similarity-Based Fusion of MEG and fMRI Data in Human Ventral and Dorsal Visual Streams [Internet]. Vol. 3, Vision. MDPI AG; 2019. p. 8. Available from: http://dx.doi.org/10.3390/vision3010008
19. Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. Development of the social brain from age three to twelve years [Internet]. Vol. 9, Nature Communications. Springer Science and Business Media LLC; 2018. Available from: http://dx.doi.org/10.1038/s41467-018-03399-2