Voxel to voxel encoding models reveal unexpected structure in unexplained variance

Maggie Mae Mell (mell@musc.edu)
Medical University of South Carolina Department of Neuroscience, 173 Ashley Avenue
Charleston, SC 29425 USA

Thomas Naselaris (tnaselar@musc.edu)
Medical University of South Carolina Department of Neuroscience, 173 Ashley Avenue
Charleston, SC 29425 USA

Encoding models based on feedforward convolutional neural networks (CNN) accurately predict BOLD responses to natural scenes in many visual cortical areas. However, for a fraction of voxels in all visual areas CNN-based models fail. Is the unexplained variance in these voxels just noise? We investigated this using voxel-to-voxel (vox2vox) encoding models that predict activity in a target voxel given activity in a population of source voxels. We found that linear vox2vox models increased prediction accuracy over CNN-based models for any pair of source/target visual areas, and recovered receptive field location even in voxels for which the CNN-based model failed. Vox2vox model prediction accuracy depended critically on the source/target pair: for feedforward models (source area lower in the visual hierarchy than target area) prediction accuracy decreased with hierarchical distance between source and target. It did not decrease for feedback models. In contrast, the same analysis applied across layers of a CNN did not reveal this feedforward/feedback asymmetry. We conclude that the variance unexplained by CNN-based encoding models is shared across visual areas, encodes meaningful information about the stimulus, and may be related to feedback connections that are present in the brain but absent in the neural network.

Keywords: fMRI; encoding models; neural networks; feedback; top-down; connectivity

Introduction

A critical measure of understanding in visual neuroscience is the ability to predict how the brain will respond to arbitrary, complex stimuli. Models that predict brain activity in response to visual stimuli are known as encoding models (St-Yves and Naselaris, 2011). Currently, the most accurate encoding models for predicting responses in visual cortical areas to natural scene stimuli are based upon convolutional neural networks (CNNs) that have been trained on object recognition tasks (Krizhevsky, 2012). However, CNN-based encoding models leave much of the variance in brain activity in response to natural scenes unexplained, particularly in voxels with peri-foveal and foveal receptive field locations. What is it about this unexplained activity that makes it difficult for CNN-based encoding models to predict?

It is possible that variance unexplained by CNN-based models simply reflects a noise ceiling. That is, unexplained variance is activity driven by unmeasured (and therefore unmodelable) nuisance sources or fMRI-related artifacts. Alternatively, unexplained variance may reflect a common, potentially stimulus-driven source of activity that the feature maps of performance-optimized CNNs fail to adequately model. If so, we would expect that unexplained variance would be correlated across voxels, and would encode meaningful and potentially recoverable information about the stimulus.

To investigate the nature of this unexplained variance, we developed a voxel-to-voxel (vox2vox) encoding model approach (Figure 1). Unlike stimulus-to-voxel (stim2vox) encoding models (e.g., the CNN-based encoding model), vox2vox encoding models use activity in a population of source voxels to predict activity in a target voxel. Vox2vox encoding models can leverage stimulus-driven activity in source voxels to explain activity in target voxels, even if this activity cannot be explained by extant stim2vox models. Vox2vox encoding models can thus be used to rapidly and easily mine the variance unexplained by any stim2vox model for meaningful structure and information content.

Here we show that even simple linear vox2vox encoding models account for considerably more variance in fMRI BOLD responses to complex natural scenes than CNN-based stim2vox encoding models, regardless of the particular pairing of source and target visual areas. In addition, we show that receptive location of the target voxels inferred from the weights of vox2vox models are consistent with retinotopy derived from independent mapping experiments. We then present evidence that the inability of CCN-based encoding models to predict such widely shared and retinotopically mapped activity may be related to a mismatch between the purely feedforward architecture of the CNN and bidirectional architecture of the human visual system.

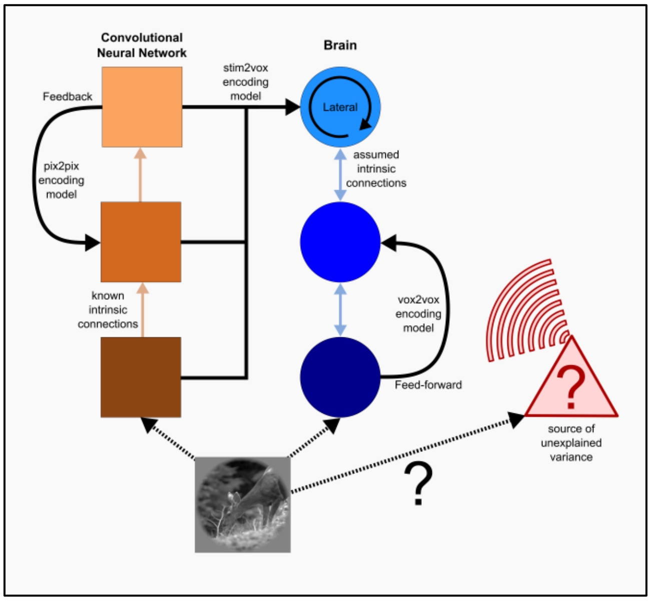
Convolutional Neural Network
Brain
stim2vox encoding model
Feedback
Lateral
assumed intrinsic connections
pix2pix encoding model
known intrinsic connections
vox2vox encoding model
Feed-forward
?
source of unexplained variance
?

Figure 1: Schematic of models. Squares (left) represent layers of a CNN. Circles (middle) represent distinct visual brain areas (e.g., V1, V2, V3, etc). CNN-based encoding models (middle arrow) map stimuli (picture at bottom) to brain activity; we refer to this as a "stim2vox" encoding model. Variance in brain activity that cannot be explained by the stim2vox model (red triangle at right) may or may not be common to all visual areas, and may or not may not be related to stimulus. We explore properties of this unexplained activity using voxel-to-voxel (vox2vox) encoding models that use activity from a source area to predict activity in a target voxel. Examples of feed-forward and lateral models (black arrows, middle) are shown. Pixel-to-pixel (pix2pix) models that linearly combine activity in pixels of a source layer to predict activity in a target pixel. The example here depicts a feedback pix2pix model (left black arrow).

Methods

Data We analyzed data from two fMRI experiments: a standard retinotopic mapping experiment, and a public ally available natural scenes dataset (vim-1). The vim-1 dataset includes BOLD responses to 1,890 natural scene photographs for two subjects. Voxels were localized to several regions of interest, including V1, V2, V3, V4, V3a, V3b, & LO (for our analyses we combined V3a and V3b into one area, V3ab). The retinotopic mapping experiment featured standard rotating wedge, expanding ring, and drifting bar stimuli. Coverage included all areas named above.

Stimulus-to-voxel models Two "stim2vox" models were used to generate 'ground truth' receptive field information and predict voxel activity to natural scene stimuli. For all voxels in the vim-1 dataset, a feature-weighted receptive field model (fwRF; St. Yves and Naselaris, 2017) was applied to the feature maps of a deep convolutional neural network. For retinotopic mapping experiments a population receptive field (pRF) analysis was used (Dumoulin et al., 2008).

Voxel-to-voxel models Voxel-to-voxel encoding models linearly combine activity from one brain area to predict activity in one voxel. We used ridge regression to determine the weights assigned to each voxel in a source area. We fit a separate vox2vox encoding model for each pair of visual areas named above. Thus, for each voxel we fit seven distinct vox2vox encoding models corresponding to the seven ROIs named above.

Pixel-to-pixel models Pixel-to-pixel models linearly combine activity in the pixels of one CNN layer to predict activity of a single target pixel. These models ignore the connection weight learned when the CNN was trained to recognize objects in natural scenes. As with vox2vox models we fit a pix2pix model for every possible pair of source/target layers.

Prediction accuracy and cross-validation: All encoding models were trained on 1750 responses to natural scene photographs and cross-validated on the remaining 120. Prediction accuracy is the Pearson correlation between model predictions and measured activity (in the brain or in the CNN).

Results

To determine if the unexplained variance was shared across voxels or just random and independent noise, we first compared our vox2vox encoding models to CNN-based stim2vox encoding model for each source-target pairing (Figure 2). Vox2vox models consistently outperform stim2vox models. For nearly every target voxel in every source/target pairing, vox2vox models have higher cross-validated prediction accuracy. Importantly, the vox2vox model accurately predicts activity (Pearson correlation > 0.2; permutation test) in many of the voxels for which the CCN-based model has almost no prediction accuracy (Pearson correlation near 0). Thus, the vox2vox model must leverage a widely shared source of activity to which this stim2vox model is almost entirely blind.

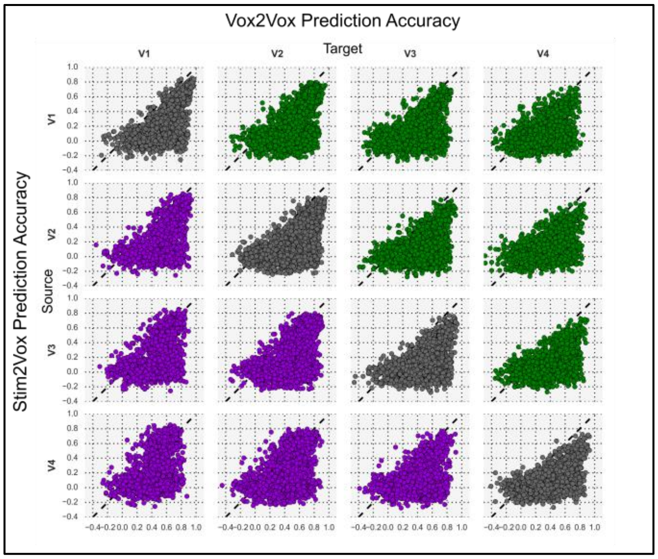

Figure 2: Matrix of source-target pairings for V1, V2, V3, and V4. Row indicates source area, columns represent target areas. Squares of matrix are color-coded according to direction of vox2vox model. Green for feed-forward, grey for lateral, purple for feedback. In each square is a scatter plot with each point representing one target voxel in that model. Prediction accuracy for vox2vox model is on x-axis and prediction accuracy for stim2vox model on y-axis.

Does this shared yet unexplained activity encode any meaningful stimulus information? We estimated receptive field location for each voxel by applying a pRF model to data obtained from a retinotopic mapping experiment. We then estimated vox2vox encoding models using these data. We extracted weights assigned to all source voxels for a target voxel and projected them as points in visual space using the RF center estimates from the pRF model (Figure 3). Source voxels with the largest weights clustered on or near the RF center of the target voxel, while source voxels with large negative weights surrounded it. Next, we binned and summed weights according to RF center and identified the visual field location with largest summed weights. We refer to this as the vox2vox receptive field location. The vox2vox receptive field locations are highly consistent with the receptive field locations obtained from the pRF models. Similar results were obtained for the natural scenes data (not shown here), even for voxels whose activity was accurately predicted by only the vox2vox model. Thus, variance unexplained by the CCN-based stim2vox model appears to be retinotopically mapped. This finding suggests that this unexplained activity does indeed encode meaningful stimulus information.

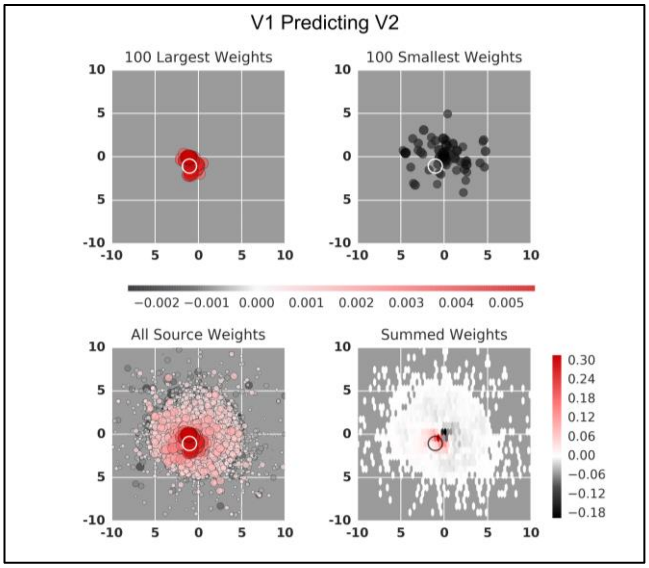

Figure 3: Vox2vox encoding models are consistent with cortical retinotopy. Grey squares represent visual field, degree of visual angle on both the x and y axes. Bubbles represent the RF locations of voxels in V1 (source area) used to predict one V2 (target area) voxel. Color indicates weight assigned to each V1 voxel by vox2vox model, red is positive and black is negative. Size of bubble is proportional to magnitude of assigned weight. White circle indicates ground truth RF location for target V2 voxel. Top plots show 100 largest weights on left and 100 smallest on right. Bottom left, all weights. Bottom right, smoothed plot shows sum of all weights in each hexagon. Hexagon with max summed weight outlined in yellow.

CNNs are purely feedforward, whereas the human visual system has extensive lateral and feedback connections. This difference in architecture might be one of the many possible reasons that CNN-based encoding models fail to predict activity that is accurately predicted by vox2vox encoding models. To test this hypothesis, we examined median prediction accuracy of the vox2vox models for each pairing of source and target visual area as a function of hierarchical distance and direction (Figure 4). Median prediction accuracy was highest in lateral models; as distance between a source and target area increased in the feedforward direction, prediction accuracy declined. For feedback models prediction accuracy did not decrease with distance. For most source/target pairs, the feedback model had higher median prediction accuracy than the feedforward model.

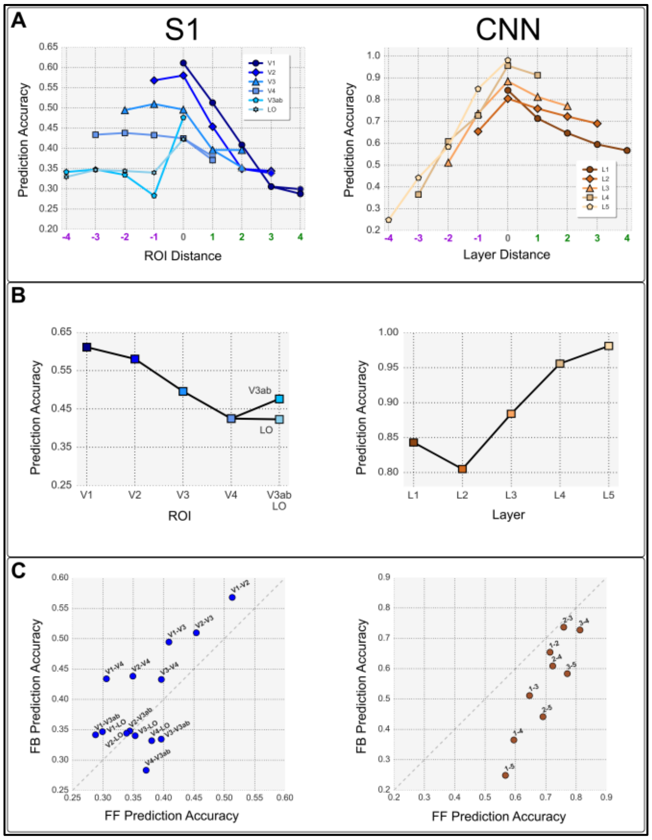

Figure 4: A: Median prediction accuracy as a function of model direction & pair distance. Hierarchical distance between source-target pairs on x-axis. 0 indicates lateral model for area/layer. Positive distances for feedforward direction; negative for feedback direction. Median prediction accuracy of vox2vox/pix2pix model for each source/target pair on y-axis. B: Prediction accuracy of lateral vox2vox and pix2pix models. Each brain area or layer is on x-axis, listed in hierarchal order. Median prediction accuracy for that area/layer's lateral model is on y-axis. C: Feed-forward versus feedback prediction accuracy for each source-target pairing. Median feedback prediction accuracy appears on y-axis, feed-forward on x-axis. Each dot represents one source-target pairing (e.g. V1 <-> V2 or L1 <-> L2).

We compared patterns of prediction accuracy obtained from vox2vox models to patterns of prediction accuracy obtained from pix2pix models. Pix2pix models were applied to each pairing of source and target layer in the CNN. The pattern of differences in prediction accuracy between lateral, feedforward, and feedback models observed in the brain were not observed in the CNN. Qualitatively, the feedforward pattern is similar between CNN layers and brain areas. In both cases, median prediction accuracy declines as distance between source and target increases. In contrast, while prediction accuracy of lateral models in in the brain declines with ascension of the visual hierarchy, it increases for deeper layers in the CNN. Furthermore, median prediction accuracy of feedback models in CNN layers decline sharply, unlike in the brain.

Discussion

The relationships between patterns of activity in distinct visual areas in the brain are nonlinear; as such, linear vox2vox models cannot possibly provide an adequate characterization of inter-area relationships in the brain. The surprisingly large prediction accuracy of linear vox2vox models is therefore highly instructive. It suggests that either (1) much of the BOLD activity we measure with fMRI encodes stimulus information in a format that is missed by the best extant (and highly nonlinear) stim2vox models (i.e., the CCN-based encoding model), or (2) much of the activity is driven by a stimulus-independent source of variance that is nonetheless widely shared across distinct visual areas. The fact that this unexplained variance is retinotopically mapped suggests that it is indeed stimulus-related (option 1): it is unlikely that a measurement artifact would be spatially structured in this way. It is therefore important to consider how we might improve upon stim2vox encoding models. Our results provide a hint. Linear approximations to the relationships between visual areas (vox2vox models) in the brain and layers (pix2pix models) in the CNN fail in very different ways depending on where the source and target are positioned along the processing hierarchy. In the feedforward direction, the CNN appears to capture the increasingly nonlinear relationship imposed by hierarchical distance between visual areas; in the feedback direction the vox2vox model is largely unaffected by hierarchical distance. This asymmetry is not present in the CNN. Exploration of network architectures that exhibit this interesting, if unintuitive asymmetry may be a promising direction in the search for a better model of the brain.

Acknowledgments

NEI R01 EY023384

References

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, *39(2)*, 647-660.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097-1105). Lake Tahoe, NV: Curran Associates, Inc.

St-Yves, G., & Naselaris, T. (2017). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*.