Compression-enabled interpretability of voxel-wise encoding models

Fatemeh Kamali[1], Amir Abolfazl Suratgar[1,*], Mohammadbagher Menhaj[1], and Reza Abbasi-Asl[2,3,*]

[1]Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
[2]Department of Neurology, Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA
[3]UCSF Weill Institute for Neurosciences
[*]Corresponding authors: Reza Abbasi-Asl (Reza.AbbasiAsl@ucsf.edu) and Amir Abolfazl Suratgar (a-suratgar@aut.ac.ir)

Abstract

Voxel-wise encoding models based on convolutional neural networks (CNNs) have emerged as state-of-the-art predictive models of brain activity evoked by natural movies. Despite the superior predictive performance of CNN-based models, the huge number of parameters in these models have made them difficult to interpret for domain experts. Here, we investigate the role of model compression in building more interpretable and more stable CNN-based voxel-wise models. We used (1) structural compression techniques to prune less important CNN filters and connections, (2) a receptive field compression method to choose the model receptive fields with optimal center and size, and (3) principal component analysis to reduce the dimensionality of the model. We demonstrate that the compressed models offer a more stable interpretation of voxel-wise pattern selectivity compared to uncompressed models. Furthermore, our receptive field compressed models reveal that the optimal model receptive fields become larger along the ventral visual pathway, as the receptive fields become more centralized. Overall, our findings unveil the role of model compression in building more interpretable voxel-wise models.

1. Introduction

A prominent question in computational neuroscience is how sensory information emerges in the visual cortex [1, 2]. Various computational models have been developed to predict the sensory representations and brain activity during sensory tasks. Constructing accurate and data-driven models requires large-scale data collected from the brain capturing brain's high-dimensional and complex activity. In the past decade, functional magnetic resonance imaging (fMRI) has emerged as a standard technique to record the brain activity during natural visual tasks [1–7]. Voxel-wise encoding models of fMRI blood oxygen level-dependent (BOLD) signal take sensory stimuli as input to predict brain activity during various visual tasks. On the other hand, the decoding models use brain activity to reconstruct and categorize visual stimuli [2, 6, 8]. These encoding and decoding models provide functional description of cortical areas.

Voxel-wise encoding models of BOLD signals are often composed of two components: (1) a feature extraction module to construct a rich feature set from visual stimulus, and (2) a response prediction module to accurately predict BOLD signal from stimulus features. The feature extraction module has been historically built based on classical machine learning techniques such as local binary pattern (LBP) [9], fisher vector [3], word to vector [10], or Gabor wavelets [2, 11]. With the recent developments in the field of deep learning, deep networks and specifically convolutional neural networks (CNNs) have emerged as the state-of-the-art encoding models of BOLD signal [8, 12–18]. These networks extract both high-level and low-level features through their hierarchical structure. For example, Agrawal et al. proposed a deep CNN to predict BOLD signals in the visual cortex and showed superior performance of CNN-based models compared to SIFT-based models [3]. Other types of deep neural networks including recurrent neural networks [19], autoencoders [5], deep residual networks [20], image captioning models [4], and capsule networks [21] have also offered accurate predictive models of responses in human visual cortex. For the response prediction module in voxel-wise models, regularized linear regression has been the standard approach [3, 7, 8, 15, 19, 20]. A linear model provides a simple map between non-linear features and the BOLD signal and therefore is more interpretable [7]. Figure 1.A illustrates the two modules of the voxel-wise encoding models often used to predict BOLD responses.

Despite their promising performance, models based on deep CNNs are often extremely hard to interpret [22–24]. Specifically, millions of parameters and the highly non-linear transformations in these models make them impossible to understand for human observers and domain experts. In many scientific applications such as computational neuroscience, this form of post-hoc interpretation is essential in understanding the scientific phenomena underlying the model. Recently, model compression has emerged as an efficient technique to interpret CNN-based models [25, 26]. Compression removes redundant components of the model while preserving the accuracy of the model, therefore, a compressed model is easier to understand for a domain expert. Additionally, a compressed model requires considerably fewer computational operations, and therefore is faster than the uncompressed model. This reduced computational cost could facilitate applications of compressed models in real-time prediction.

In this paper, we explore the role of model compression in building more interpretable and stable CNN-based models of BOLD signal. We use multiple compression techniques including (1) a recently established structural compression [25] method to prune less important CNN filters, (2) Deep Compression [27] to remove less important connections, (3) a receptive field compression [16] to choose the receptive fields with the optimal center and size, and (4) principal component analysis [28] to reduce the dimensionality of the model. Using two separate fMRI dataset collected during natural vision from 6 participants, we first show that the compressed encoding models are as accurate as the uncompressed models in predicting the BOLD signal. Furthermore, we demonstrate that compression reduces the model size and computational cost. We then establish that the compressed encoding models reveal increased category-selectivity along the ventral visual pathway with higher stability compared to uncompressed models. Finally, we leverage our compressed models to quantitatively compare the model receptive field sizes and locations along different visual pathways.

2. Results

2.1. Compressed CNN-based encoding models accurately predict BOLD responses

While our goal is to improve the interpretability of the CNN-based voxel-wise models, this cannot come at significant costs to the prediction accuracy. As such, it is necessary to determine the prediction accuracy of compressed models. To quantify the accuracy, we computed the voxel-wise Pearson correlation coefficient between the predicted and measured BOLD signal. We then reported the average correlation coefficient for the following compression techniques: (1) structural compression (SC) where redundant filters are removed from the CNN-base model, (2) deep compression (DC) where CNN model weights that are close to zero are removed from the model, and (3) receptive field compression (RC) where the optimal receptive field size and location is determined for each voxel (see Methods for the detailed description of each method). We have systematically compared the prediction accuracy of these compressed models with the uncompressed

CNN-based model for each visual areas. The visual areas that we considered are V1, V2, V3, V4, Lateral Occipital (LO), Middle Temporal (MT), Fusiform Face Area (FFA), Parahippocampal Place Area (PPA), Extrastriate Body Area (EBA), and Retrosplenial Cortex (RSC).

Figure 1 illustrates the average prediction accuracy of the encoding models based on the individual CNN layers (as opposed to the entire set of CNN features from all layers which is discussed in Figure 2). Panel B in Figure 1 presents the accuracies for the best performing models among 8 distinct models that are built based on each layers of the CNN. The average correlation coefficient is reported across all the voxels within each visual area. Our results indicate that both structurally compressed and Deep-compressed models perform comparably to the uncompressed model. For FFA, the Deep-compressed model underperforms other models by 3%, most likely due to higher sensitivity of face features to pruning CNN weights. The receptive field compressed model has a slightly lower accuracy in the early and intermediate visual areas compared to other models (10% in V1, 5% in V2 and V3, and 2% in V4), but achieves a similar accuracy to the uncompressed model for the higher visual areas. Figure 1.C illustrates the predictive accuracy of the models based on each individual CNN layer for structurally compressed models (as opposed to panel A where the best performing model is selected). The result is similar for other compression techniques and therefore are not shown here to avoid redundancy. Our findings suggest that early to middle layers of the CNN are better predictors of the responses in early and intermediate visual areas, while responses in higher visual areas are better predicted by deeper CNN layers. For areas V1, V2, V3, and V4, CNN layers 2 to 5 achieve the highest accuracies. For areas LO, MT, FFA, PPA, and EBA, models based on layers 6 to 8 perform better compared to those based on other layers. Surprisingly, for the RSC, the best performing models are based on the middle layers of the CNN. One possible explanation for this observation is that RSC is known to be a scene-selective area [3, 29]. However, the CNN used in our models (AlexNet) is trained on a dataset that largely contains objects rather than scenes.

We further used the features extracted from the entire CNN layers in one single encoding model to estimate the BOLD responses. In addition to considering each individual compression technique, here we also present two combined compression methods: (1) structural and receptive field compression (SRC) and (2) deep and receptive field compression (DRC). For both SRC and DRC, the receptive field compression is used after the structural or deep compression. The prediction accuracies reported in Figure 2.A suggest that using the entire feature map to predict the BOLD signal results in higher accuracies ($9\pm2\%$ on average) compared to regressing single layers of CNNs. This is not surprising because a larger number of features with complimentary image statistics (from different layers of CNN) are used in these models compared to the single-layer models. Among the visual areas, FFA voxels exhibit a considerably larger improvement when the entire CNN features are used (correlation coefficient of 0.52 with SE of 0.01) compared to the best performing individual-layer model (correlation coefficient of 0.30 with SE of 0.01) for receptive field compression. This suggests that FFA models require a diverse set of features including both low-level and high-level image statistics perhaps due to the complexity of the face images.

For the best performing models in Figure 2.A., we also quantified the contribution of each layer to the prediction accuracy for each visual area (Figure 2.B). Overall, our results suggest that the average contribution of the extracted feature maps to predicting BOLD signal in each visual area largely depends on the position of that area in the visual hierarchy. This is consistent with observations from previous studies [13] indicating that features from lower CNN layers have higher contribution to the prediction of responses in lower visual areas, while higher visual areas are better predicted by the higher CNN layers.

We further visualize the voxel-wise accuracies based on the entire CNN layers on the cortical map in one of the Subject (see Supplementary Materials, Figures A1 to A7 for the accuracies in each individual Subject). To create these maps we computed the Pearson correlation coefficients between predicted and measured responses for each voxel. The inflated and flattened cortical maps for the uncompressed and compressed models are shown in Figure 3. The structurally compressed model outperforms the the uncompressed model in ProS, DVT and the lateral part of the V1, suggesting these regions are better modeled using less CNN filters. The deep-compressed model has a higher predictive accuracy compared to the uncompressed model in parts of V1, V2, V3 and the lateral part of V4. For the receptive field compressed model, the lateral parts of

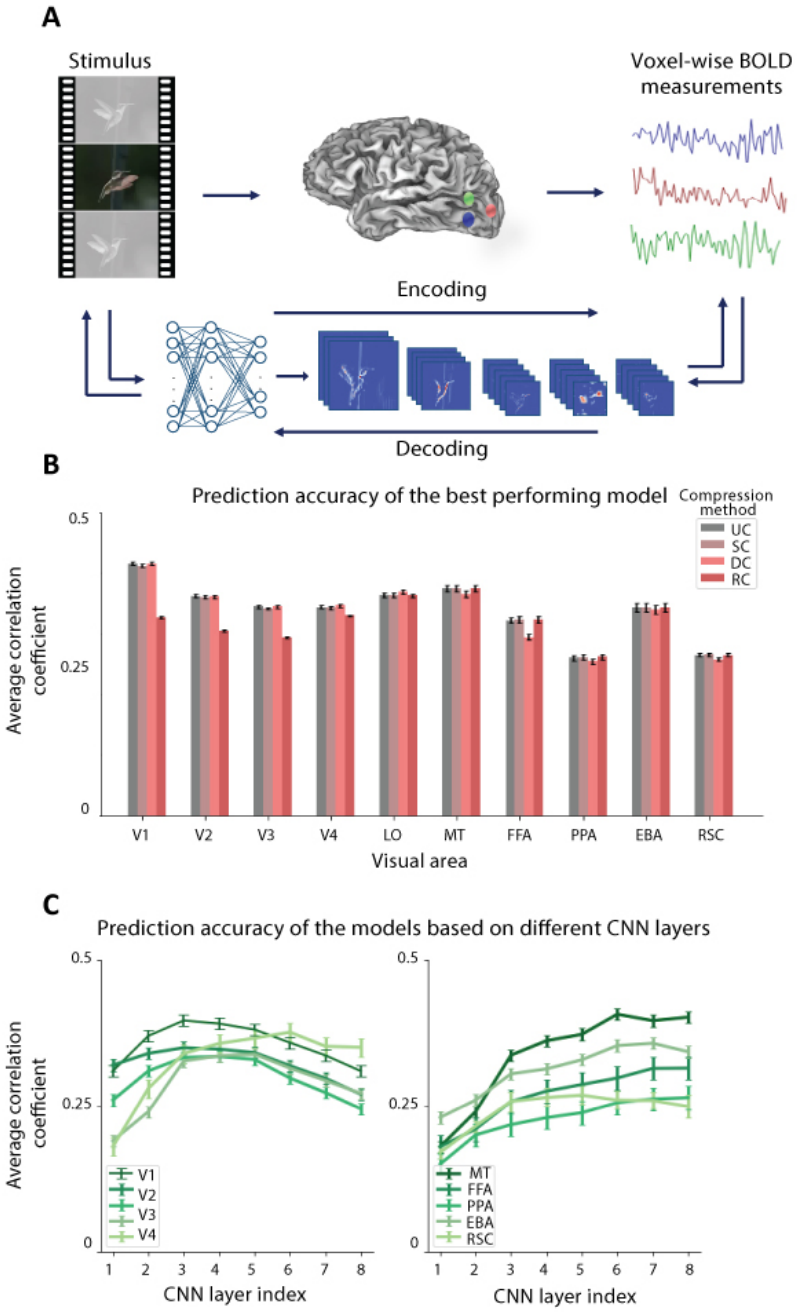

Figure 1: Compressed voxel-wise encoding models accurately predict BOLD responses. A. An encoding model predicts the fMRI BOLD responses from the visual stimulus features. The decoding model predicts the optimal stimulus from the BOLD responses. **B.** The Pearson correlation coefficient between estimated response on best CNN layer and measured fMRI averaged over all voxels in each visual area across all Subjects ± standard error for uncompressed (UC), structurally compressed (SC), deep-compressed (DC), and receptive field compressed (RC) model. Compression does not significantly affect the accuracy. **C.** The Pearson correlation between estimated and measured fMRI averaged over all voxels in each visual area for compressed models for different CNN layers ± standard deviation. The contribution of the higher CNN layers is attenuated for early visual areas, while the reverse trend occurs for higher visual areas.

V1, V2, V3, V4 and also TPOJ and FST areas are modeled more accurately compared to the uncompressed model. Other Subjects display similar results (Supplementary Materials Figure A1).

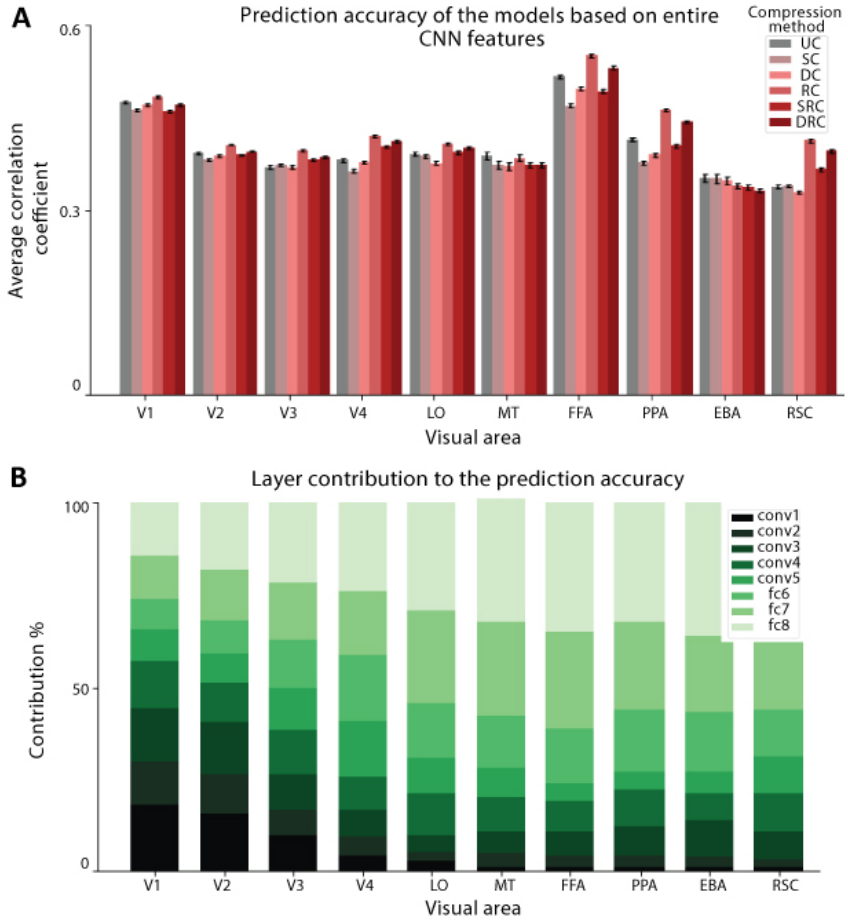

Figure 2: Prediction accuracy of the encoding model based on the entire CNN layers. A. Each bar indicates the pearson correlation between estimated and measured fMRI averaged over all voxels in each visual area and across all Subjects ± standard error for uncompressed (UC), structurally compressed (SC), deep-compressed (DC) and receptive field compressed (RC), structurally receptive field (SRC) compressed and deep receptive field compressed model (DRC). Compression does not significantly affect the accuracy. **B.** Each column indicates the contribution of each CNN layer to prediction accuracy. Feature maps extracted from lower CNN layers have higher contribution to lower visual areas and Feature maps extracted from higher CNN layers have higher contribution to higher visual areas.

2.2. Compression reduces the model size and computational cost while preserving the accuracy

So far, we demonstrated that the encoding models that are compressed following our guidelines (see Methods) retain a satisfactory prediction accuracy. Here, we examine the compression ratio, the model size, and the computational cost of these compressed models. The model size is quantified by the number of weights (or parameters) in the model. To quantify the computational cost, we tracked the number of floating point operations per second (FLOPS) and the number of trainable parameters. For a convolutional layer in a neural network, the number of FLOPS refers to the number of floating point operations in that layer to extract all of the feature maps, without accounting for the regression overhead. Table 1 summarizes the number of FLOPS

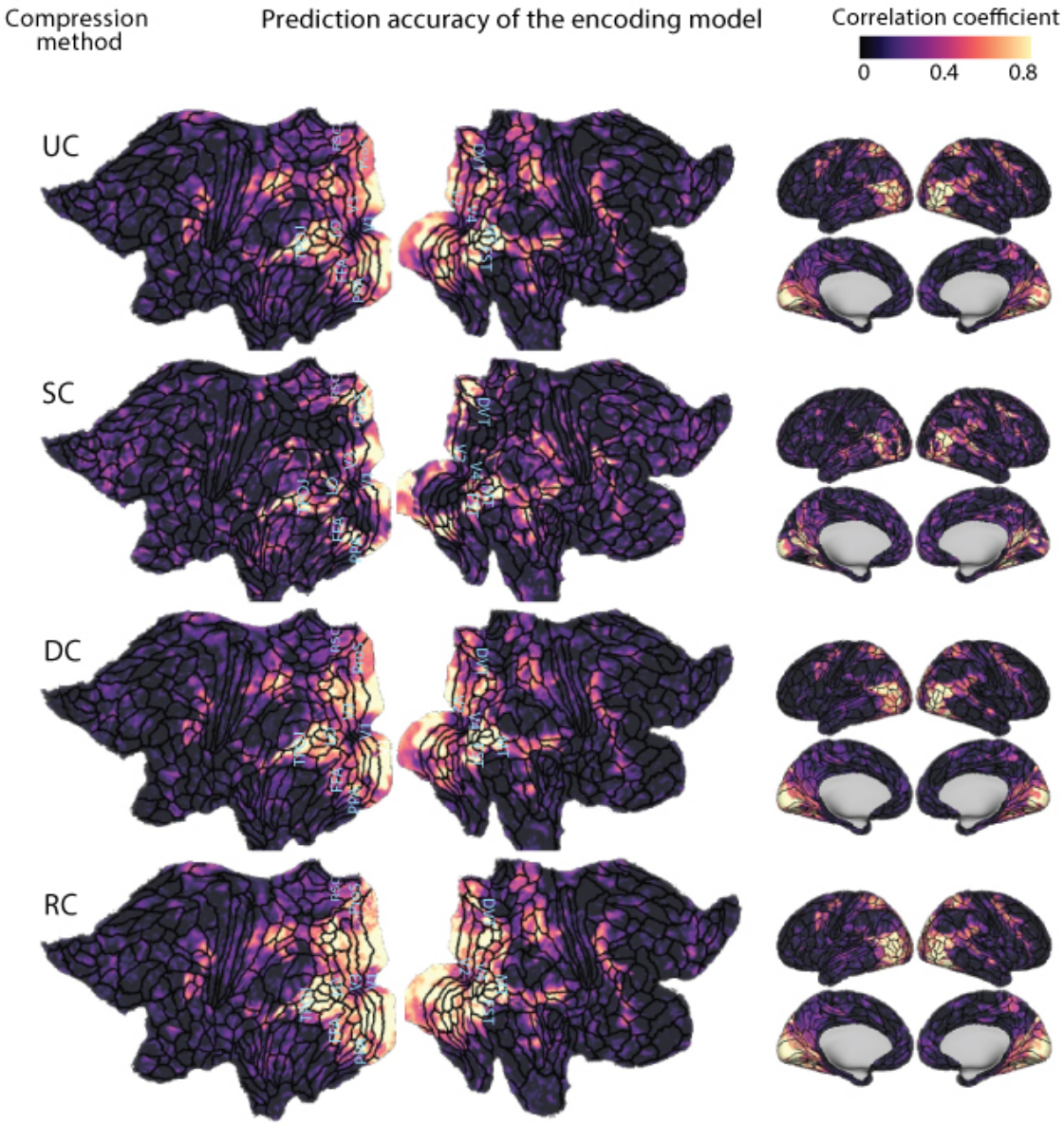

Figure 3: Prediction accuracy of the compressed encoding models on cortical maps. The maps show the correlations between estimated and measured response for uncompressed model (UC), structurally compressed model (SC), deep-compressed model (DC), receptive field compressed model (RC). Compared to uncompressed models, structurally compressed models better estimate fMRI responses in ProS, DVT and the lateral part of V1. deep-compressed model is better in central part of V1, V2, V3 and the lateral part of V4. The receptive field compressed model better estimate lateral part of V1, V2, V3, V4 and also TPOJ and FST areas.

and the number of weights for the uncompressed, deep-compressed and structurally compressed models (see Methods for the details of each compression method). We also present the compression ratios based on both the number of FLOPS and weights in this table. The compression ratio is computed by dividing the number of FLOPS (or weights) required for the uncompressed model by that of the compressed model. The PCA and the receptive field compression methods are not included in this table because these methods only compress the regression module, therefore the number of FLOPS does not change compared to the uncompressed models.

Table 1 shows that our proposed compressed encoding models have a remarkably lower computational cost compared to the uncompressed models. The structurally compressed model from the CNN layer 4 features has the highest compression ratio compared to other layers (compression ratio of 2), and layer 5 has the lowest compression ratio (ratio of 1.2). Note that the compression ratio based on the number of FLOPS is equal to the compression ratio based on the number weights for the structural compression because the entire filters are being removed during this form of compression. The structural compression is also not defined for the fully connected layers; therefore no number is reported for these layers. For the Deep compression, the second fully connected layer has the highest compression ratio in terms of the number of FLOPS (ratio of 17.12), while both first and second fully connected layers have the highest compression ratio in terms of the number of weights (ratio of 33.33). The last row in Table 1 shows the FLOPS, number of trainable weights and compression ratios for the model that uses features from all layers of the AlexNet. For this model, structural compression has a compression ratio of 1.8 while deep compression has a ratio of 2.4 (FLOPS) and 27 (weights). Overall, our findings suggest that both structural and deep compression methods offer a reduced model size and computational cost.

2.3. Structurally compressed models reveal more stable interpretations compared to the uncompressed models

Having established how compressed models can retain a high predictive accuracy similar to their uncompressed counterparts, we now examine the advantages of constructing compressed models in model interpretation. Here, interpretability is characterized by the visual patterns that elicit the largest response in each voxel (i.e., optimal visual patterns). This form of interpretability is hindered for the uncompressed CNN-based models due to their large number of parameters [25]. Additionally, the model-based optimal visual patterns are often unstable for the large uncompressed models [30]. Intuitively, we define stability as the similarity of visual patterns among the images with the highest model responses. Later in this subsection, we provide a formal definition for this notion of stability. Considering the lower number of parameters in structurally compressed models, we hypothesize that these models allow for a more stable interpretation of pattern selectivity for each voxel. More specifically, we ask whether the set of visual patterns that elicit the largest response in each voxel are more stable for the compressed model compared to the uncompressed model.

To perform this experiment, we first identified the visual patterns that elicit the largest response for both uncompressed models and the structurally compressed model. We chose 10,000 natural images from the validation set in the ILSVRC 2012 dataset [31] and computed the voxel-wise response of both structurally compressed and uncompressed encoding models to each image. The images with the highest response were selected for each voxel. Note that these images were neither used to train CNN nor included in the experimental stimulus set. Figure 4.A presents the top 5 images with the highest model response for the most accurately predicted voxel in each visual area. We have visualized these preferred images for the top 10 most accurately predicted voxels in each visual area in the Supplementary Materials (Figures A8 to A12). Overall, these images provide a qualitative representation of the patterns selected by each voxel. The patterns for the voxels in V1 and V2 do not contain category specific patterns. This is consistent with previous observations that V1 and V2 areas lack category selectivity and are more selective to lower-level image features such as edges and borders [3, 13, 32]. Top images for the V4 area contain semi-complex shapes such as circles, crosses and dense textures. Previous single cell studies have confirmed the selectivity of V4 area to these complex shapes and textures [30]. Top images for PPA and RSC contain environmental scenes. These findings are

Table 1: Comparison of the computational cost and the number of weights for the uncompressed, structurally compressed, and deep-compressed models. The computational cost is quantified by FLOPS which equals to the number of floating-point operations required in each layer to classify one image.

Layer	Compression method	#FLOPS	Compression ratio (FLOPS)	#Weights	Compression ratio (Weights)
conv1	Uncompressed	105.41M	-	35K	-
	Structural	58.56M	1.8	19.4K	1.8
	Deep	88.89M	1.18	11.43K	3.06
conv2	Uncompressed	223.95M	-	307K	-
	Structural	124.41M	1.8	170.55K	1.8
	Deep	86.38M	2.59	44.55K	6.89
conv3	Uncompressed	149.52M	-	885K	-
	Structural	93.45M	1.6	553.12K	1.6
	Deep	52.24	2.86	115.98K	7.63
conv4	Uncompressed	112.14M	-	663K	
	Structural	56.07M	2	331.5K	2
	Deep	41.89M	2.67	93.51K	7.09
conv5	Uncompressed	74.76M	-	442K	-
	Structural	62.3M	1.2	368.33K	1.2
	Deep	27.69M	2.69	61.09	7.14
fc6	Uncompressed	37748736	-	38M	-
	Deep	2332737	16.18	1.14M	33.33
fc7	Uncompressed	16777216	-	17M	
	Deep	979690	17.12	0.51M	33.33
fc8	Uncompressed	4096000	-	4M	
	Deep	530823	7.71	0.29M	13.69
All layers together	Uncompressed	724.37M	-	61M	-
	Structural	394.79M	1.8	33.8	1.8
	Deep	300.79M	2.4	2.25M	27

aligned with the current understanding about functional landscape of these visual areas, suggesting both uncompressed and compressed models are successful in determining the preferred images for the accurately predicted voxels.

We now examine the stability of the preferred visual patterns for each voxel and compare the stability between the compressed and uncompressed models. We used the similarity between CNN-extracted features from images to quantify the stability. This feature-based similarity measure is chosen over the pixel-based measure because CNN-extracted features encompass the overall content of each image and reflect the image patterns more reliably compared to the pixel values. Here, we constructed the feature space by concatenating features from all layers of the AlexNet [31]. Formally, the stability score is defined as the average pairwise Pearson correlation coefficient between CNN-extracted features across images. We considered the stability score computed over top 10 images with the highest model response. We computed this score for the top 100 most accurately predicted voxels in each area for both structurally compressed and uncompressed models. Scatterplots comparing these stability scores between the compressed and the uncompressed models are shown in Figure 4.B. These plots indicate that the stability is considerably higher for the compressed models in V1, V2, V3, V4, LO, MT, FFA, and RSC (with the mean pairwise correlation of 0.48 ± 0.02, 0.47 ± 0.04, 0.46 ± 0.06, 0.44 ± 0.06, 0.48 ± 0.04, 0.48 ± 0.04, 0.47 ± 0.04, and 0.47 ± 0.04 for the structurally compressed model, outperforming the uncompressed model by 0.07, 0.06, 0.04, 0.05, 0.06, 0.04, 0.05, and 0.06 respectively). The stability scores in PPA are only marginally higher for the compressed model (0.43 ± 0.06 for the structurally compressed model outperforming the uncompressed model by only 0.008). Considering only

the top 50 most-accurately predicted voxels in PPA, the compressed model dominates the the uncompressed model. The compressed model has a higher correlation coefficient for more than 81% of the top voxels across all visual areas indicating that the compression provides more stable features than the uncompressed model for the majority of accurately predicted voxels. Overall, our findings indicate that the structurally compressed models allow for a more stable interpretation of pattern selectivity for each voxel. In addition to the assessment of model interpretability, or findings indicate that the downstream areas in the ventral visual pathway are more category-selective compared to early areas. This is consistent with prior studies [3, 19], where category-selectivity is quantified for the areas in this visual pathway.

2.4. The receptive-field-compressed models reveal increased size and centralization of the receptive fields along the ventral visual pathway

The organization of the receptive field maps in different areas in visual cortex have been extensively studies in the past. Majority of these studies have provided evidence for larger and more centralized receptive fields in higher visual areas along the ventral visual pathway [33], [16]. Our receptive-field-compressed encoding models allow for a systematic and quantitative analysis of this observation in different visual areas along the ventral visual pathway. In this section, we leverage our compressed models to quantitatively and systematically characterize the optimal size and location of the model receptive fields.

Figure 5.A illustrates the selected receptive fields for the top 100 most accurate voxels in visual areas V1, V2, V3, V4, LO, MT, FFA, PPA, EBA, and RSC. Visually, the areas in the early visual pathway have smaller and more scattered receptive fields while higher visual areas have larger and more centralized receptive fields. To quantitatively compare the receptive field locations and sizes, we computed the mean absolute distance between the center of each receptive field and the mean receptive field center across all voxels in each visual area (Figure 5.B). We found that V1 has the smallest average receptive field size (1.62 ± 0.02 degrees) with the most scattered centers (7.13 ± 0.04 degrees). V2, V3, and V4 exhibit slightly larger mean receptive field size (1.75 ± 0.02 and 1.77 ± 0.02, and 1.86 ± 0.03 degrees, respectively) with gradually increasing centralization (the mean eccentricity of 7.08 ± 0.04, 6.64 ± 0.04, and 6.10 ± 0.04 respectively). On the other hand, voxels in FFA and PPA had the largest receptive fields (mean radius of 3.33 ± 0.08 and 3.40 ± 0.07 degrees, respectively) followed by LO, MT and EBA (mean radius of 2.22 ± 0.04 and 2.66 ± 0.07, and 2.55 ± 0.09 degrees, respectively). The receptive fields in MT, FFA, PPA, and EBA were the most centralized (with the mean eccentricity of 5.27 ± 0.09, 4.58 ± 0.08, 4.68 ± 0.07 and 4.78 ± 0.09 respectively). Overall, our receptive field compressed encoding model provides a quantitative and systematic framework to compare the receptive field sizes and centers along the visual hierarchy. Our models systematically confirm findings by previous studies that the receptive fields become larger and more concentrated as we move downstream in ventral pathways.

3. Discussion

Encoding and decoding models are powerful tools to investigate human vision. We have shown that compressing CNN-based encoding models significantly decreases the number of parameters involved while preserving accuracy. We used structural compression to remove less important filters, deep compression to remove less important connections, and the receptive field compression to pool the features. Our findings suggest that compressed encoding models provide an interpretable and quantitative framework to investigate the relationship between natural visual stimuli and the fMRI BOLD signal.

In this paper, we constructed a two-step encoding model that consists of a image feature extraction module and a linear mapping between features and BOLD signal. Our image feature extraction module was a deep convolutional neural network trained on a natural image classification task. An alternative modeling approach is to directly train a deep neural network that takes image stimuli as the input and predicts the voxel-wise responses. Such a design requires an order of magnitude larger fMRI dataset in order to allow for

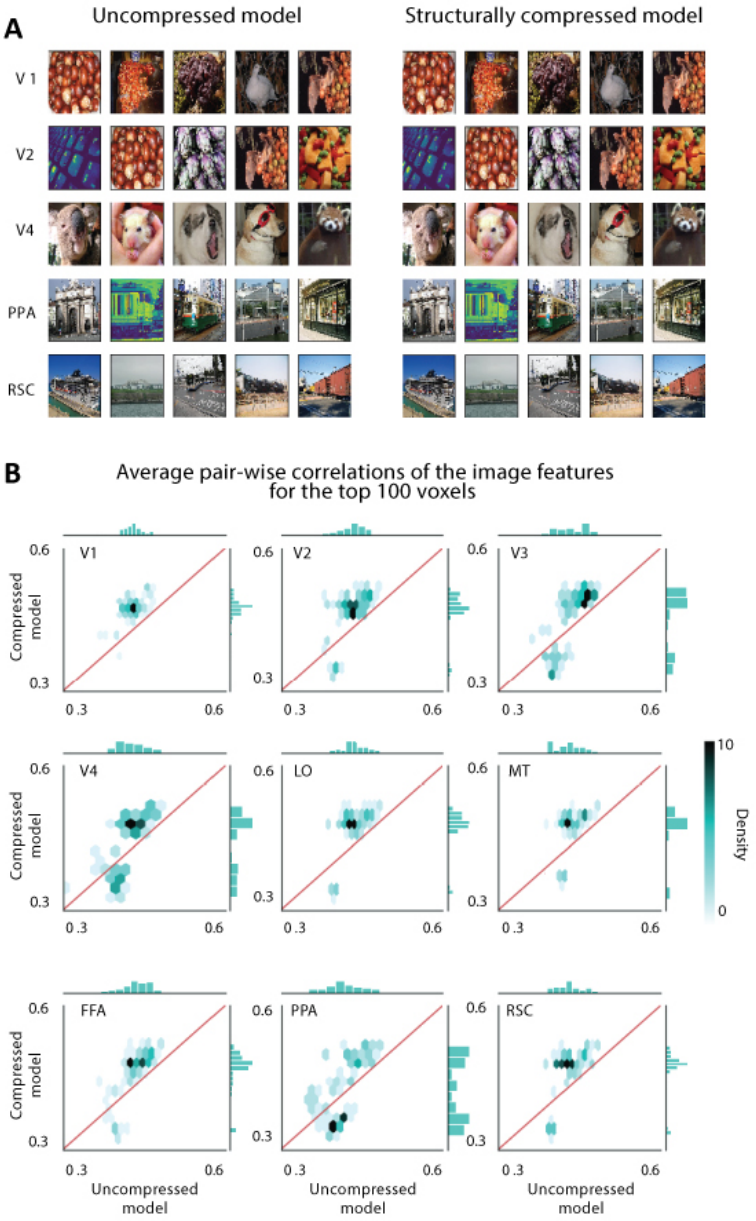

Figure 4: Structurally compressed models reveal more stable category-selectivity compared to the uncompressed models. A. Top 5 images with the highest model response for each visual area and for both structurally compressed and uncompressed models. Top images for the V1 and V2 areas are rich in lower-level image features such as edges and borders. Top images for the V4 area contain semi-complex shapes such as circles, crosses, and dense textures. Top images for PPA and RSC contain environmental scenes. **B.** Density plots comparing the stability of the selected images between compressed and uncompressed models. The stability is quantified by the average pairwise correlation coefficient between CNN-extracted features of the top 10 selected images for each voxel. The average correlation coefficients is reported across top 100 voxels for 9 visual areas. The color of each hexagonal bin indicates the density of voxels in that bin. The histograms at each plot represent the distribution of the relative stability score. Overall, top images are more stable for the structurally compressed encoding model compared to the uncompressed model.

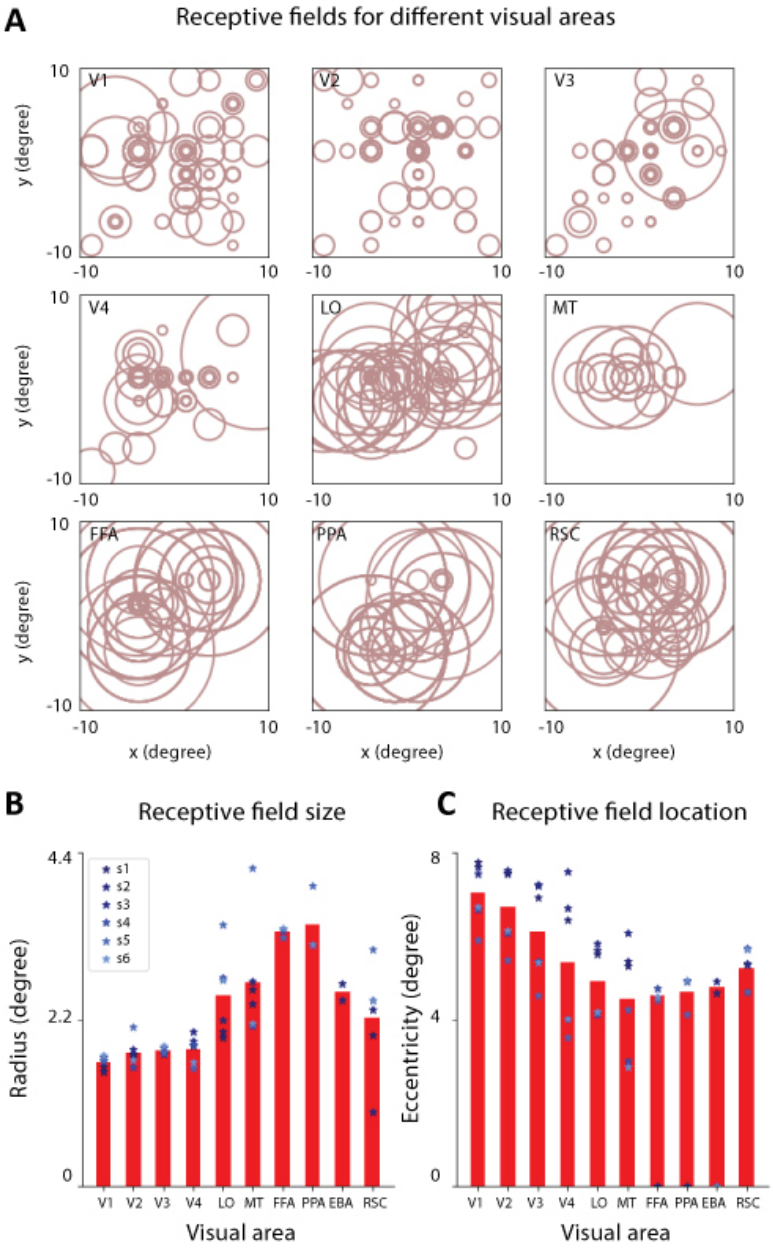

Figure 5: The receptive-field-compressed models reveal increased size and centralization of the receptive fields along the ventral visual pathway. A. Each circle illustrates the receptive field for an individual voxel from the top 100 most accurately predicted voxels. The radius and the center of each circle is the determined by the radius and the center of the Gaussian pooling field selected through the receptive field compression. **B.** The mean radius and **C.** the mean absolute distance between the center of each receptive field and the mean receptive field center across all voxels in each visual area. Receptive fields become larger and more concentrated as we move towards the downstream regions in the visual area.

effective training of the end-to-end deep learning model. In principle, a more aggressive compression allows for training an end-to-end model with limited fMRI data. However, the validation of this model requires further validation in future follow-up studies.

The deep neural network used in this study was trained on a single task (image classification), however, the representations across the human visual cortex emerge in response to various tasks such as classification, detection, and recognition. A more accurate encoding model would aggregate features relevant to various tasks, but the huge size of the feature space in these aggregated models may introduce feasibility issues. Our findings suggest that future studies aimed at the construction of such multi-modal systems should consider compression techniques as an essential part of their design to allow for more effective model interpretation.

4. Methods

4.1. Datasets

To build and investigate our compressed voxel-wise models, we used two separate fMRI BOLD signal datasets. Each dataset contains BOLD fMRI recordings from three healthy participants watching hours of natural movie clips. For each Subject, the dataset is divided into non-overlapping training and test sets. Further information about each dataset is provided below.

4.1.1. PURR Dataset

fMRI data were obtained from three healthy volunteers in a 3T MRI system with a temporal resolution of 2s. The training set contains 374 movie clips (continuous with frame rate of 30 fps) in a 2.4-h movie, divided randomly into 18 8-minute sections; the test set contains 598 movie clips in a 40-min movie, divided randomly into 5 8=minute sections. The training set was repeated two times whereas the test set was repeated ten times. Each examination contains several sections of clips that are 8 minutes and 24 seconds long. During each section, an 8 minute single video segment was shown; the first and the last movie frames were shown as a static picture for 12s. Stimuli were chosen from video blocks and YouTube. The fMRI data were preprocessed and co-registered onto a standard cortical surface template using the processing pipeline for the Human Connectome Project [34]. The visual areas were defined with multi-modal cortical parcellation [35]. Additional details on this dataset can be found in [8].

4.1.2. Vim-2 Dataset

fMRI data were obtained in a 4T MRI system with a temporal resolution of 1s [36] from three healthy volunteers. The training set includes a 2 hour movie; the test set includes a 9 minute movie. The training set was shown only once whereas the test set was repeated ten times. Stimuli were chosen from Apple Quick-Time HD gallery and YouTube. Retinotopic mapping data collected from the same Subjects in separate scan sessions was used to assign voxels to visual areas [37]. This dataset is further described in [38].

4.2. Feature extraction *via* Convolutional Neural Networks

Our voxel-wise encoding models consist of two modules. First, the CNN-based feature extraction module that constructs a feature set for each frame in our visual stimulus. Second, the response prediction module that predicts the BOLD signal from the CNN-based features.

To extract features from each frame in the visual stimulus, we used AlexNet [31], a well-known CNN model pre-trained on the ImageNet dataset [31]. AlexNet consists of eight layers - the first five layers are convolutional and the last three are fully connected. The five convolutional layers use rectified linear activation

functions, with the first layer receiving a 227×227 input image. Max-pooling is applied between layer 1 and layer 2, between layer 2 and layer 3, and between layer 5 and layer 6. The last layer uses a softmax function to generate a probability vector, by which an input stimulus frame is classified into 1000 classes. Layer 1 through layer 5 contain 96 kernels of 11×11, 256 kernels of 27×227, 384 kernels of 13×13, 384 kernels of 13×13, and 256 kernels of 13×13, respectively. Layers 6 to 8 contain 4096, 4096, and 1000 units, respectively.

For the response prediction module, we used a linear regression with L_2-norm regularization (Ridge regression [39]). We used each CNN layer output to predict the voxel-wise responses and then selected the layer with the highest accuracy for each visual area. We also built an encoding model with features from all CNN layers concatenated together. To reduce the dimension of the predictors in this case, we used PCA to keep 99% of the variance across all layers.

Formally, we model voxel v's response, y_v, as a linear weighted linear combination of the features ϕ^l from the l_{th} layer of CNN:

$$y_v = \phi^l W_v^l + b_v^l + \epsilon \tag{1}$$

where W_v^l is the regression coefficient vector, b_v^l is the bias factor and ϵ is the model error. We then used Ridge regression with the following cost function to approximate regression coefficients from the training data:

$$f(W_v^l) = \|y_v - \phi^l W_v^l\|_2^2 + \lambda \|W_v^l\|_2^2 \tag{2}$$

The regularization parameter λ is optimized through 10-fold cross-validation. After determining λ, the training data is used to estimate the final regression coefficients. Then, the prediction accuracy is obtained from the test set by calculating the average Pearson correlation coefficient between the predicted response and the measured response across test set segments.

4.3. Layer Contribution

For our encoding models from the entire CNN feature, we investigated the contribution of each CNN layer in predicting voxel-wise responses. While it is possible to directly use the regression weights as contribution, the value of the weights highly depend on the raw values of the feature sets [16]. Therefore, we calculated CNN layer contribution for each visual area as follows:

$$C = cov(\phi^l W_v, y_v) / \sqrt{var(\phi W_v), var(y_v)} \tag{3}$$

Where ϕ^l is the feature map extracted from CNN layer l, W_v is the regression coefficient vector, and y_v is the measured response.

4.4. Structural Compression

The process of structural compression involves removing redundant filters from the model to increase model interpretability. Here, we used a recently established structural compression technique called classification accuracy reduction (CAR) compression [25]. CAR compression quantifies the contribution of each filter to the model's prediction accuracy and then removes the filters with the least contribution. We iteratively used CAR compression to continuously score and prune convolutional filters in each layer of AlexNet. Model accuracy was used as a stopping criterion to constrain the iterative structural compression. We compressed each model while the hold-out test set accuracy is still in the 2% range of the uncompressed model accuracy. Note that the hold-out test used for compression is different from the the hold-out test set used for assessing the final accuracy. The compressed CNN-based features are then further compressed using PCA to retain 99% of their variance. These dimensionality-reduced features are then convolved with a canonical hemodynamic response function (HRF) [40] with the maximum at 5s and the outputs downsampled to match the fMRI sampling [8]. Finally, estimated fMRI responses are calculated with a ridge regression on the PCA-reduced,

down-sampled features. At this step, the pruned model's accuracy was determined by the Pearson correlation between measured and predicted responses on the hold-out test set.

4.5. Deep Compression

We performed deep compression (DC) [27] to reduce the number of connections in our CNN-based encoding models. Following the process proposed in [27], we pruned the small-weight connections from the CNN. More specifically, all connections with weights below a threshold were removed while there was no drop in the classification accuracy. We further reduced the number of weights by having multiple connections share the same weight. We used k-means clustering to identify the shared weights in each layer of the network and used the same weights for all the connections that fell into the same cluster. We then fine-tuned the shared weights through retraining of the network.

4.6. Receptive Field Compression

Receptive field compression takes inspiration from the biological visual pathways, much like how CNN architectures share similarities with the hierarchical organization of visual cortex. This form of compression consists of identifying the most important regions of visual stimulus for the prediction task and then removing features from any location outside this region. Here, we modeled the receptive fields with a 2D isotropic Gaussian function [16]. Thus, the receptive field , g, can be described as:

$$g(x, y; \mu_x, \mu_y, \sigma) = 1/(\sqrt{}(2\pi)\sigma)exp((x - \mu_x)^2 + (y - \mu_y)^2)/(2\sigma^2) \tag{4}$$

Where (μ_x, μ_y) is the receptive field center and σ is the receptive field radius. To simulate the effect of biological receptive fields, the 2D Gaussian function was convolved with the CNN-based features extracted from the stimulus. Formally:

$$\phi^l_{RF} = \iint g(x, y; \mu_x, \mu_y, \sigma)(\phi^l(x, y)dxdy) \tag{5}$$

where ϕ^l_{RF} is the receptive field feature map for layer l and ϕ^l is the feature extracted from layer l of the CNN network.

We used grid search to approximate the optimal receptive field configuration for the CNN model. Since we're using a 2D Gaussian function, we only varied the center and radius for the candidate receptive fields. For each voxel, we built a grid of candidate Gaussian pooling fields with varying sizes and locations. The size of the grid was 8 by 8 on the visual field (with Gaussian centers spaced 2.5 degrees apart). In each location on this grid, 8 log-spaced receptive fields were constructed with the sizes between $\sigma = 0.5$ and $\sigma = 8$. This provided a total of 512 Gaussian pooling fields on the visual field.

Once again, the Pearson correlation coefficient between measured and predicted responses was used to quantitatively pinpoint the best receptive field configuration. More specifically, we first applied each candidate receptive field to the features extracted from the CNN model, using Eq. 4. This was followed by a convolution with the HRF filter and downsampling to match the temporal frequency of the BOLD signal (0.5Hz for the PURR dataset, and 1Hz for the Vim-2 dataset). We used 80% of the dataset to train a Ridge regression model that predicts the fMRI BOLD signal from the compressed feature set. The prediction accuracy of this model was then assessed on the remaining 20% of the data. We used this process to identify the most accurate receptive field for each voxel. To determine the final accuracy of the compressed model, we retrained the Ridge regression using 100% of the training dataset and reported the accuracy on the hold-out test set for each voxel.

Acknowledgments

The authors would like to thank Gavin Cui and Gaurav Ghosal for their valuable feedback on the manuscript. R.A. would like to acknowledge support from the Weill Neurohub and the Sandler Program for Breakthrough Biomedical Research program.

Conflict of interest/Competing interests

The authors declare no competing interests.

Availability of data and materials

All the data used in this manuscript are publicly available at https://crcns.org/data-sets/vc/vim-2 and https://engineering.purdue.edu/libi/lab/Resource.html. The intermediate data files are available upon request.

Authors' contributions

R.A., A.S., and M.M. conceptualized and conceived the experiments. F.K. conducted the experiments. All authors analyzed the data. F.K. and R.A. wrote the manuscript with contributions from A.S. and M.M..

References

[1] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[2] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[3] Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.

[4] Kai Qiao, Chi Zhang, Jian Chen, Linyuan Wang, Li Tong, and Bin Yan. Neural encoding and interpretation for high-level visual cortices based on fmri using image caption features. *arXiv preprint arXiv:2003.11797*, 2020.

[5] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

[6] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[7] Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019.

[8] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, 2018.

[9] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[10] Umut Güçlü and Marcel AJ van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

[11] Yibo Cui, Chi Zhang, Linyuan Wang, Bin Yan, and Li Tong. Dense-gwp: An improved primary visual encoding model based on dense gabor features. *Journal of Mechanics in Medicine and Biology*, page 2140017, 2021.

[12] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[13] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[14] Umut Güçlü and Marcel AJ van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017.

[15] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[16] Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 2018.

[17] Katja Seeliger, Matthias Fritsche, Umut Güçlü, Sanne Schoenmakers, J-M Schoffelen, SE Bosch, and MAJ Van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018.

[18] Ziya Yu, Chi Zhang, Linyuan Wang, Li Tong, and Bin Yan. A comparative analysis of visual encoding models based on classification and segmentation task-driven cnns. *Computational and Mathematical Methods in Medicine*, 2020, 2020.

[19] Junxing Shi, Haiguang Wen, Yizhen Zhang, Kuan Han, and Zhongming Liu. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human brain mapping*, 39(5):2269–2282, 2018.

[20] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific reports*, 8(1):1–17, 2018.

[21] Kai Qiao, Chi Zhang, Linyuan Wang, Bin Yan, Jian Chen, Lei Zeng, and Li Tong. Accurate reconstruction of image stimuli from human fmri based on the decoding model with capsule network architecture. *arXiv preprint arXiv:1801.00602*, 2018.

[22] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[23] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[24] Gaurav R Ghosal and Reza Abbasi-Asl. Multi-modal prototype learning for interpretable multivariable time series classification. *arXiv preprint arXiv:2106.09636*, 2021.

[25] Reza Abbasi-Asl and Bin Yu. Structural compression of convolutional neural networks with applications in interpretability. *Frontiers in Big Data*, 4:73, 2021.

[26] Reza Abbasi-Asl and Bin Yu. Interpreting convolutional neural networks through compression. *NIPS 2017 Symposium on Interpretable Machine Learning*, 2017.

[27] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[28] Haitao Zhao, Pong Chi Yuen, and James T Kwok. A novel incremental principal component analysis and its application for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):873–886, 2006.

[29] Merim Bilalić, Tobias Lindig, and Luca Turella. Parsing rooms: the role of the ppa and rsc in perceiving object relations and spatial layout. *Brain Structure and Function*, 224(7):2505–2524, 2019.

[30] Reza Abbasi-Asl, Yuansi Chen, Adam Bloniarz, Michael Oliver, Ben DB Willmore, Jack L Gallant, and Bin Yu. The deeptune framework for modeling and characterizing neurons in visual cortex area v4. *bioRxiv*, page 465534, 2018.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[32] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[33] Serge O Dumoulin and Brian A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660, 2008.

[34] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[35] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[36] Shinji Nishimoto, AT Vu, T Naselaris, Y Benjamini, and B Yu. Gallant lab natural movie 4t fmri data. *CRCNS. org. Available online at: http://dx. doi. org/10.6080/K00Z715X*, 2014.

[37] Kathleen A Hansen, Stephen V David, and Jack L Gallant. Parametric reverse correlation reveals spatial linearity of retinotopic human v1 bold response. *Neuroimage*, 23(1):233–241, 2004.

[38] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[39] Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

[40] Rik Henson and Karl Friston. Convolution models for fmri. *Statistical parametric mapping: The analysis of functional brain images*, pages 178–192, 2007.

A. Supplementary Materials

A.1. Prediction accuracy of compressed encoding models on cortical maps for other Subjects in the PURR dataset.

A1 illustrates the voxel-wise prediction accuracy on cortical maps for Subjects 2 and 3 for the PURR dataset. The findings outlined in the main text for Subject 1 are consistent with those for Subjects 2 and 3.

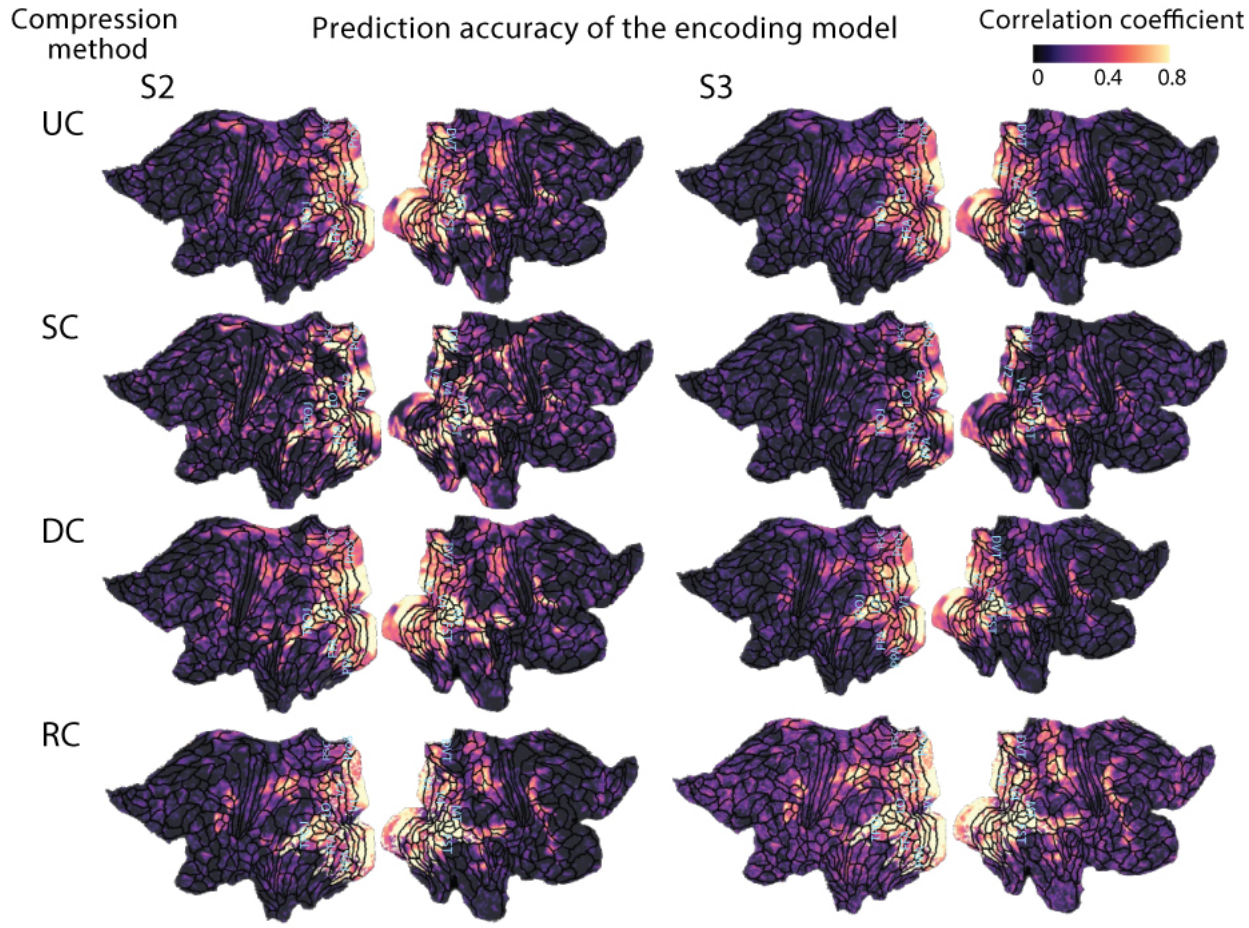

Figure A1: Prediction accuracy of compressed encoding models on cortical maps for Subjects 2 and 3 in the PURR dataset. Compared to the uncompressed (UC) model, the structurally compressed (SC) model better estimates fMRI responses in ProS, DVT and the lateral part of V1. The accuracy in the deep-compressed (DC) model is better in central part of V1, V2, V3 and the lateral part of V4. the receptive field compressed (RC) model better estimate lateral part of V1, V2, V3, V4 and also TPOJ and FST areas.

A.2. Voxel-wise comparison of correlation coefficients between compressed and uncompressed models

Figures A2, A3, A4, A5, A6, and A7 illustrate the voxel-wise comparisons of correlation coefficients between the compressed and uncompressed models.

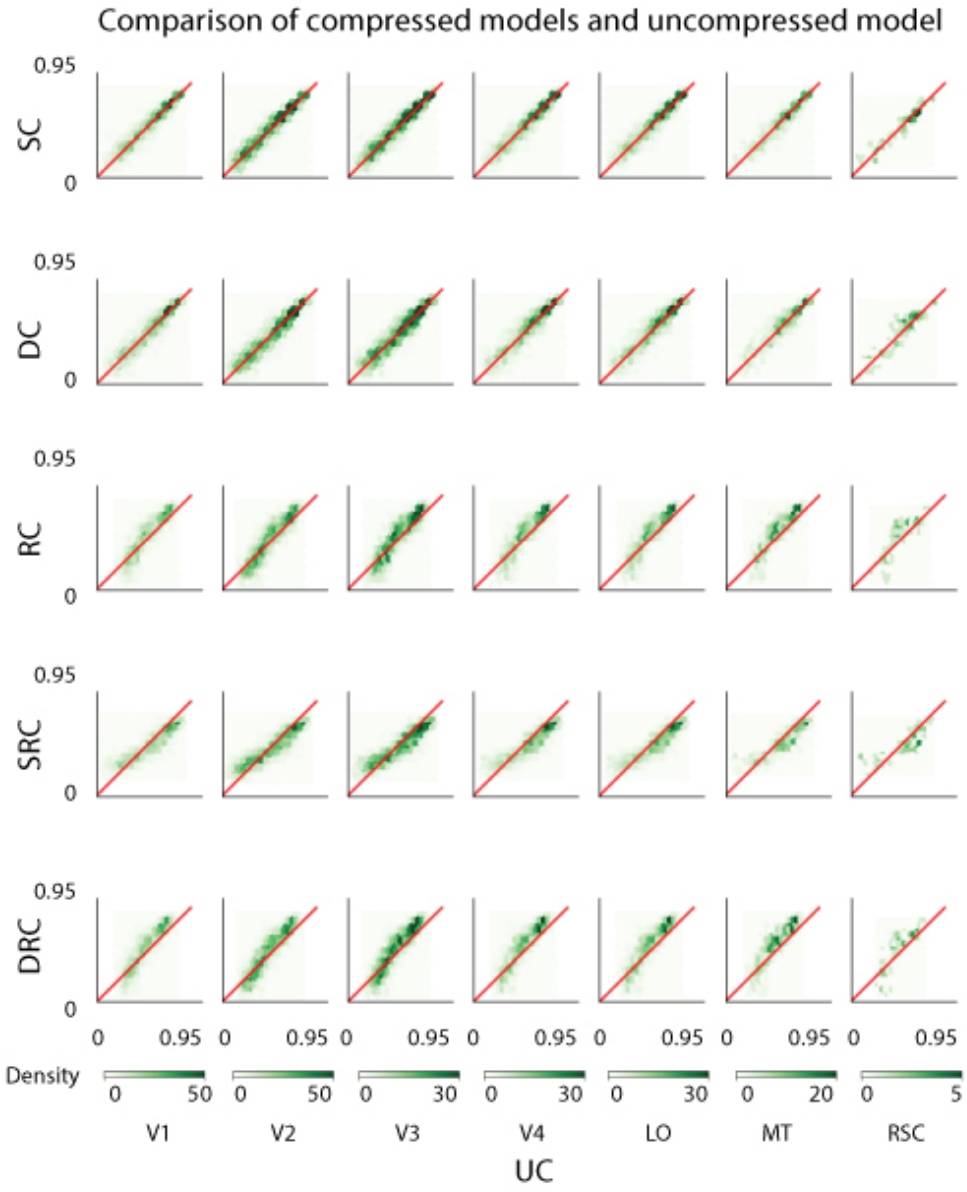

Figure A2: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 1 in the Vim-2 dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

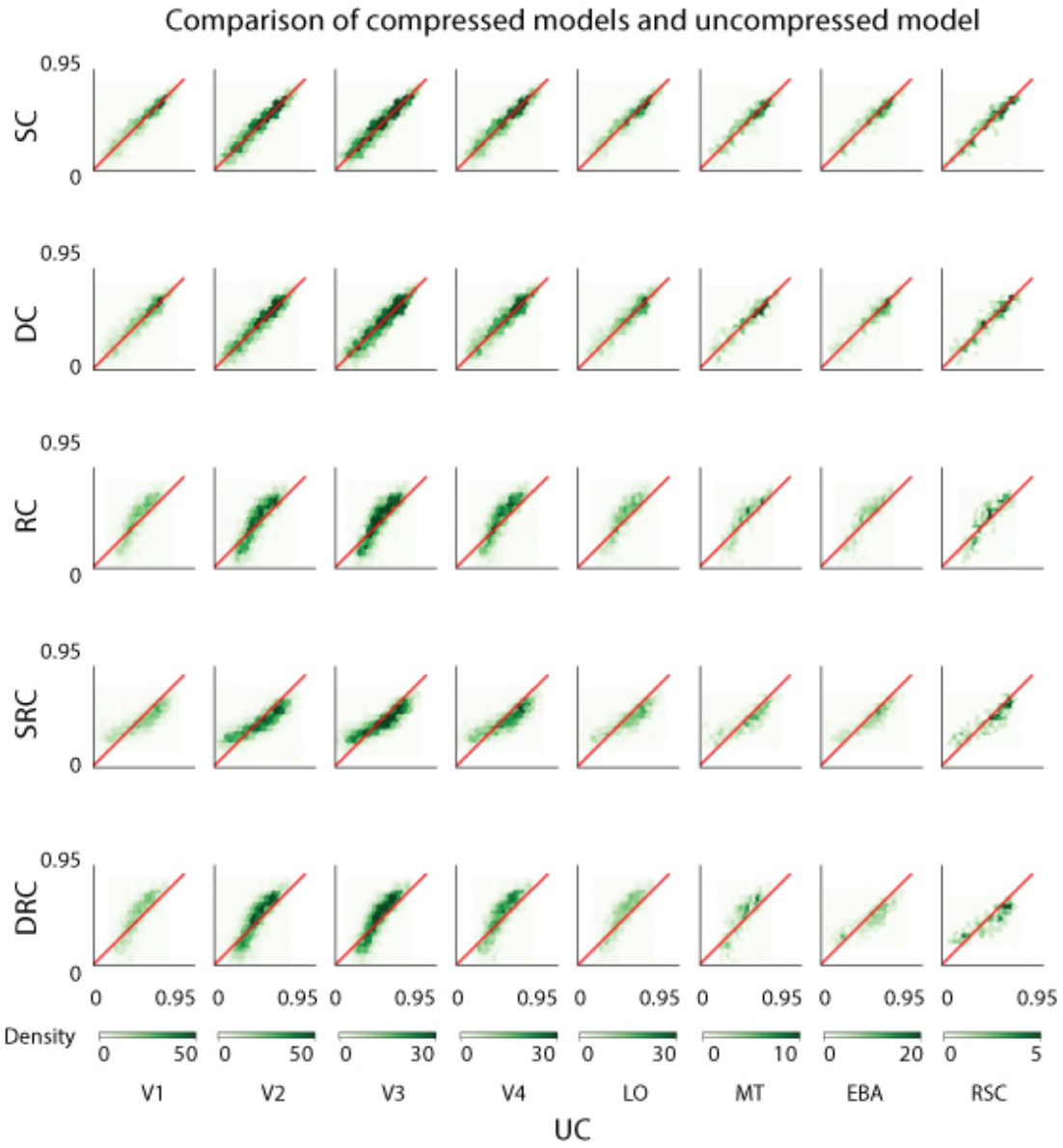

Figure A3: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 2 in the Vim-2 dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

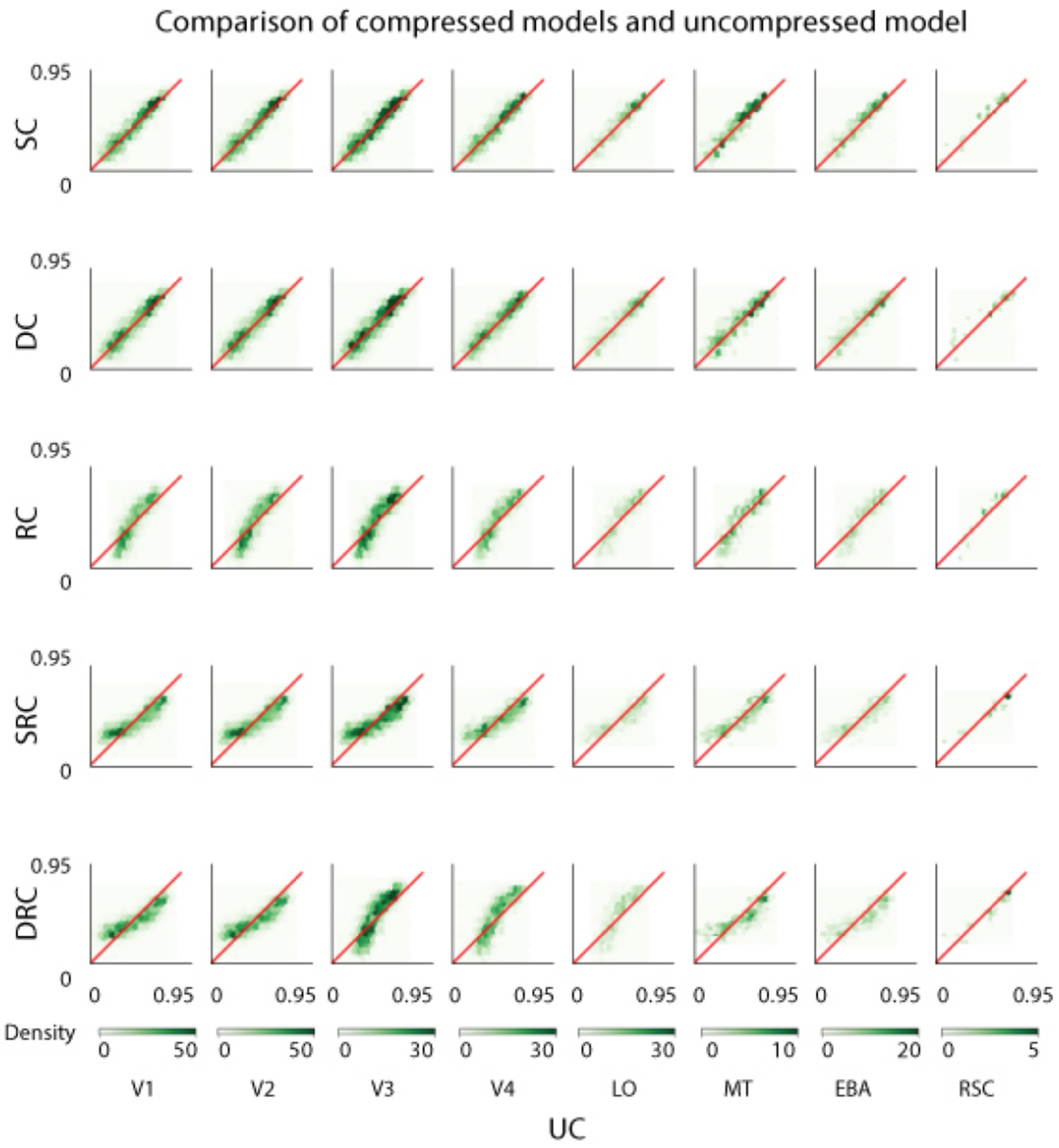

Figure A4: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 3 in the Vim-2 dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

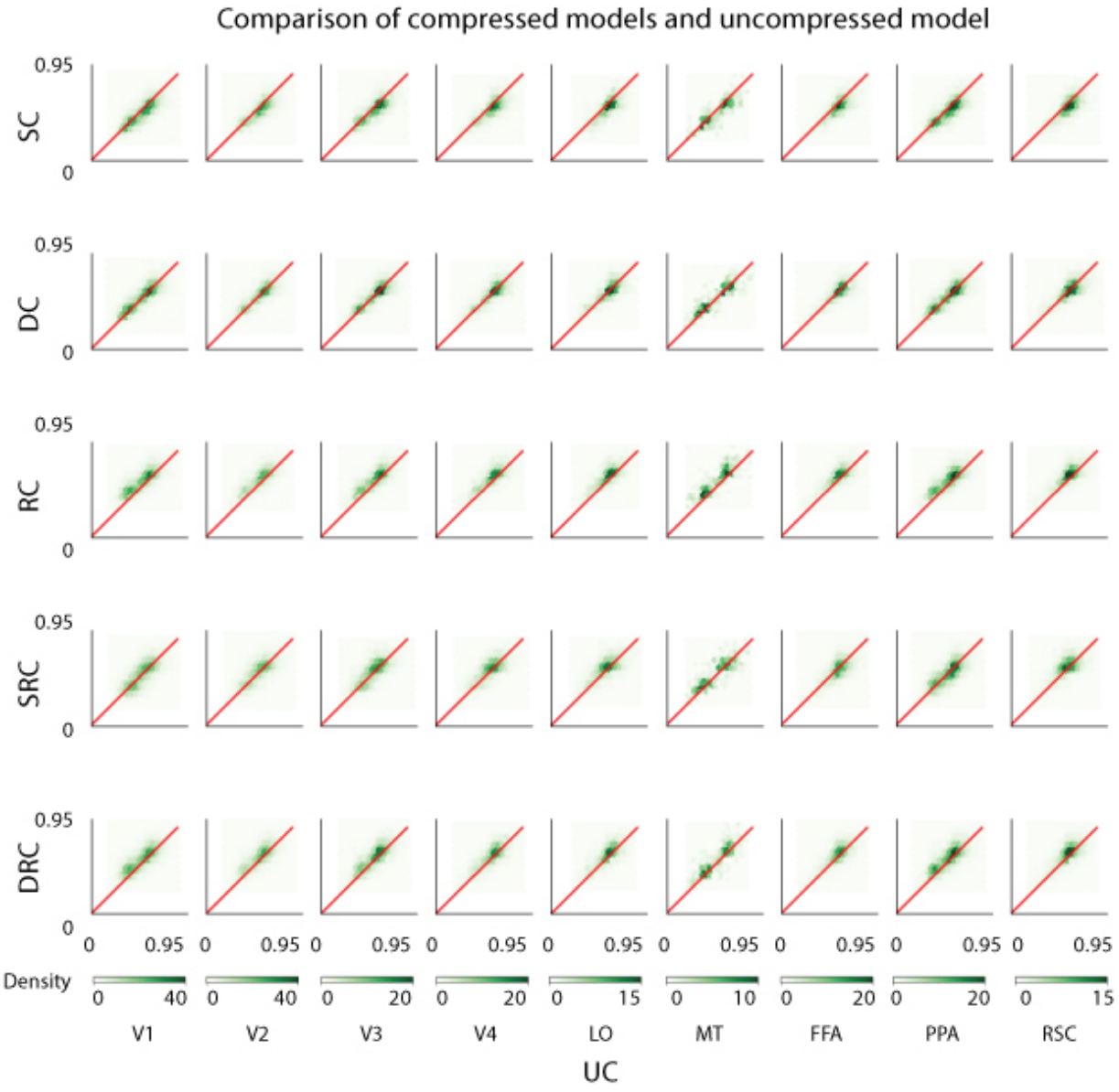

Figure A5: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 1 in the PURR dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

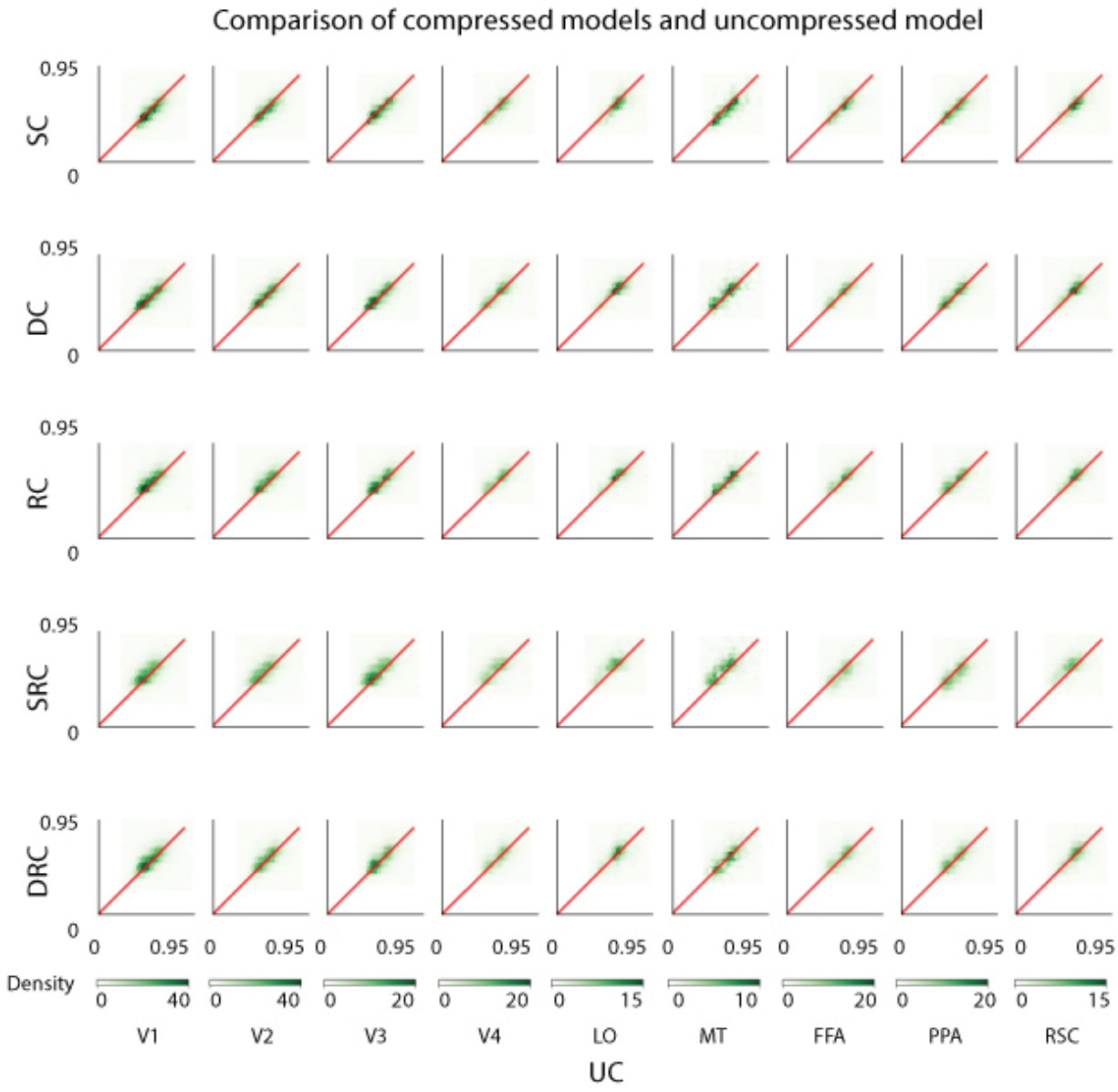

Figure A6: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 2 in the PURR dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

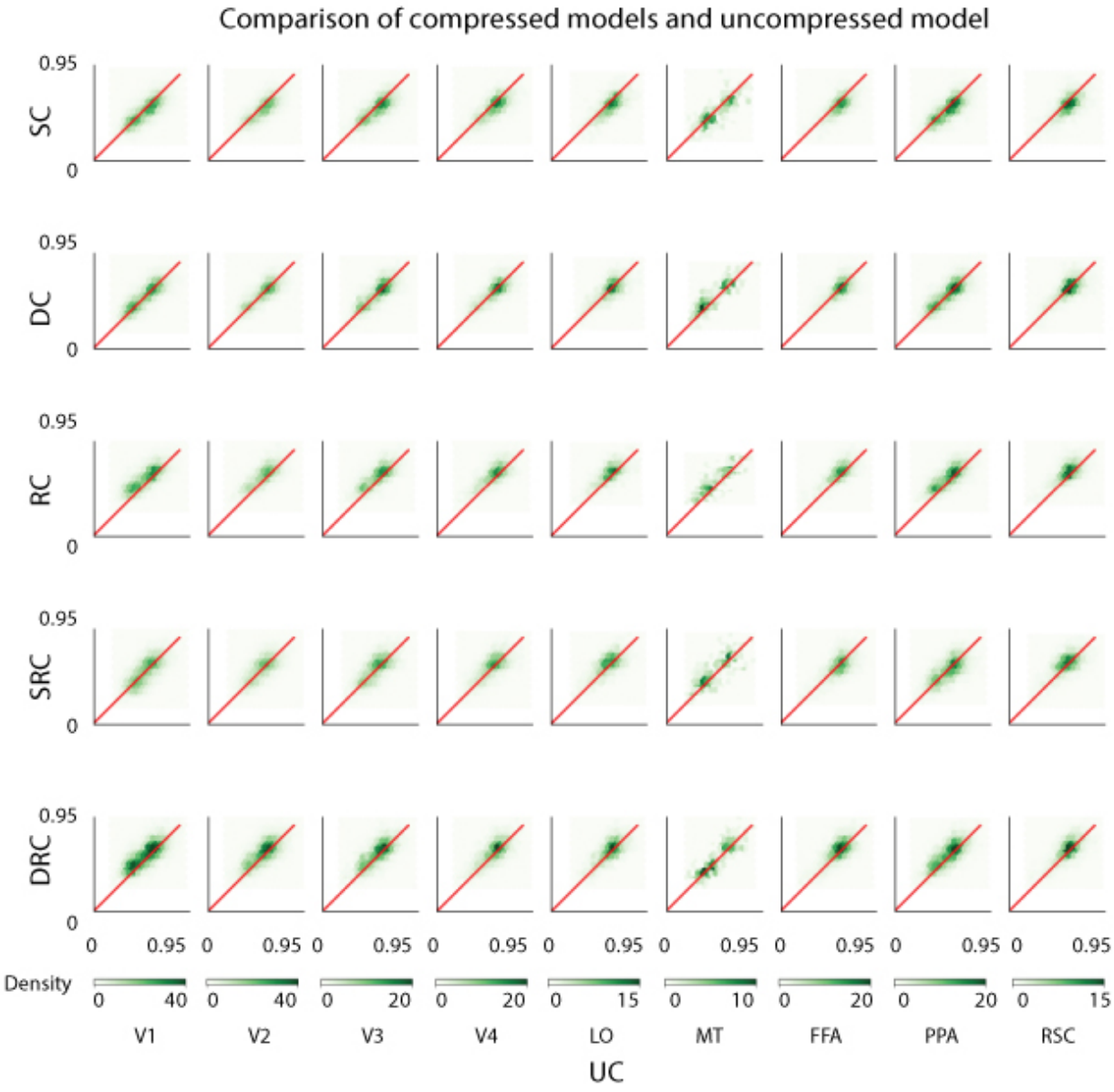

Figure A7: Scatterplots comparing voxel-wise correlation coefficients between compressed and uncompressed models for Subject 3 in the PURR dataset. Each dot corresponds to one voxel. Columns represent different visual areas. Rows represent different compression technique.

A.3. Top images with the highest model response for the structurally compressed and uncompressed models

Figures A8, A9, A10, A11, and A12 present the 5 top images with the highest response for the top 10 most accurately voxels in visual areas V1, V2, V4, PPA, and RSC, respectively. Images are shown for both uncompressed and structurally compressed models.

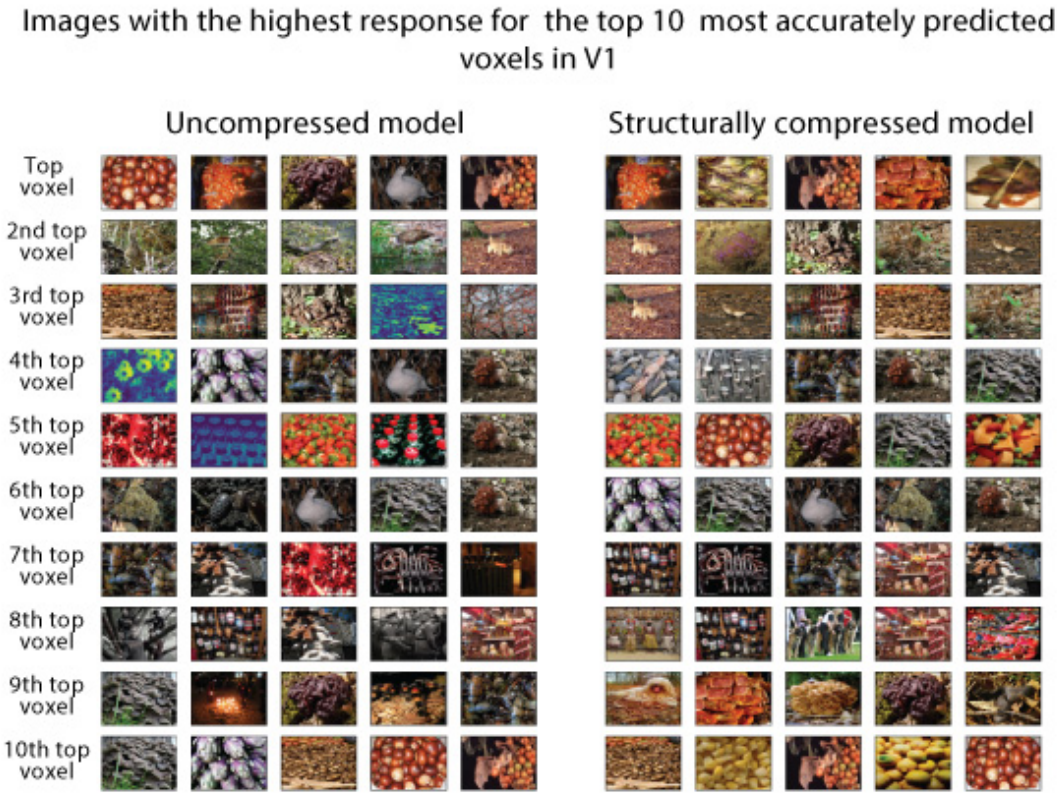

Figure A8: Comparing the performance of structurally compressed and uncompressed model: A.The five columns at left show the five images that the uncompressed model predicts will most increase activity for voxels in V1 area, and the five columns at right show the five images that the structurally compressed model predicts will most increase activity for voxels in V1 area.

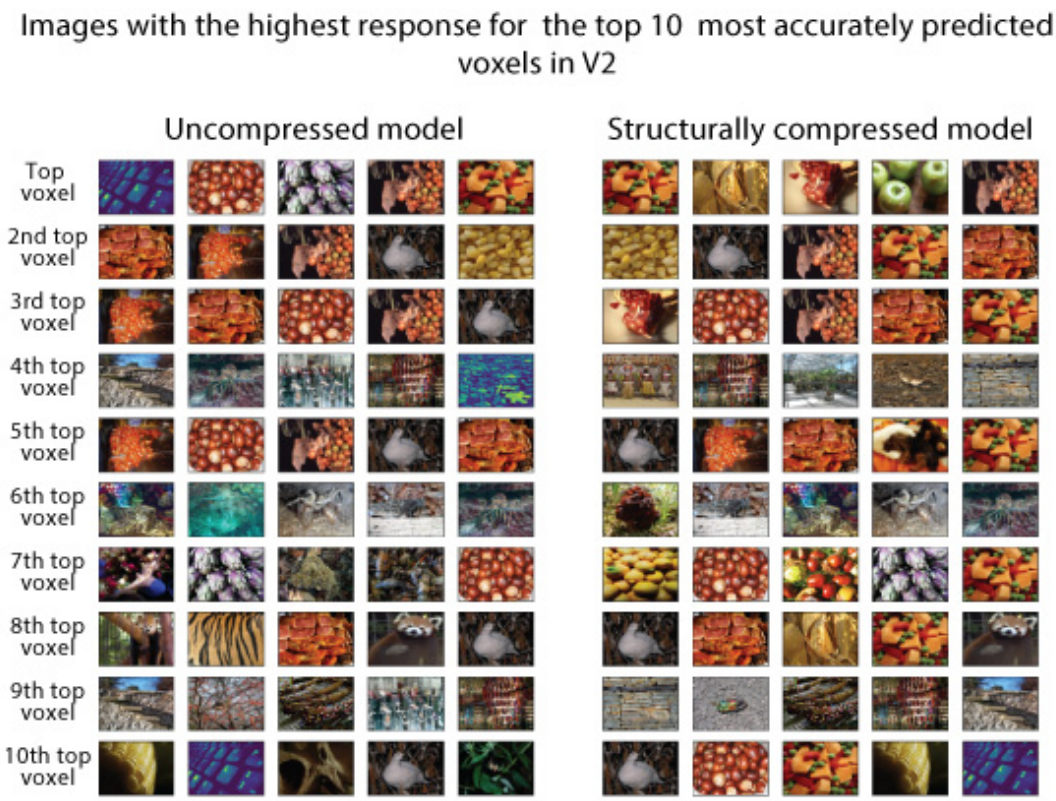

Figure A9: Comparing the performance of structurally compressed and uncompressed model: A.The five columns at left show the five images that the uncompressed model predicts will most increase activity for voxels in V2 area, and the five columns at right show the five images that the structurally compressed model predicts will most increase activity for voxels in V2 area.

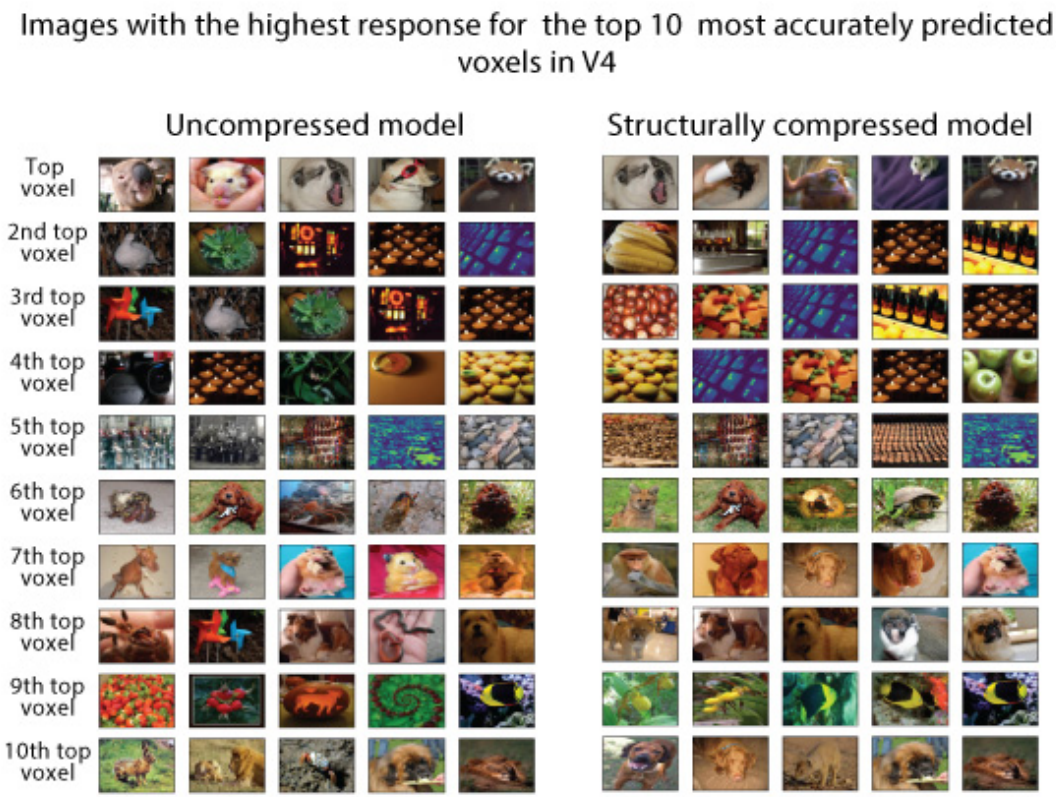

Figure A10: Comparing the performance of structurally compressed and uncompressed model: A.The five columns at left show the five images that the uncompressed model predicts will most increase activity for voxels in V4 area, and the five columns at right show the five images that the structurally compressed model predicts will most increase activity for voxels in V4 area.

Figure A11: Comparing the performance of structurally compressed and uncompressed model: A.The five columns at left show the five images that the uncompressed model predicts will most increase activity for voxels in PPA area, and the five columns at right show the five images that the structurally compressed model predicts will most increase activity for voxels in PPA area.

Figure A12: Comparing the performance of structurally compressed and uncompressed model: A.The five columns at left show the five images that the uncompressed model predicts will most increase activity for voxels in RSC area, and the five columns at right show the five images that the structurally compressed model predicts will most increase activity for voxels in RSC area.