PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Compression-enabled interpretability of voxelwise encoding models

Fatemeh Kamali[1], Amir Abolfazl Suratgar[1]*, Mohammadbagher Menhaj[1], 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, California, United States of America, **3** UCSF Weill Institute for Neurosciences, San Francisco, California, United States of America

* a-suratgar@aut.ac.ir (AAS); Reza.AbbasiAsl@ucsf.edu (RA-A)

Citation: Kamali F, Suratgar AA, Menhaj M, Abbasi-Asl R (2025) Compression-enabled interpretability of voxelwise encoding models. PLoS Comput Biol 21(2): e1012822. https://doi.org/10.1371/journal.pcbi.1012822

Editor: Daniele Marinazzo, Ghent University, BELGIUM

Received: October 11, 2023

Accepted: January 23, 2025

Published: February 19, 2025

Data availability statement: All the data used in this manuscript are publicly available at https://crcns.org/data-sets/vc/vim-2 and https://purr.purdue.edu/publications/2809. The software package developed under this study is available at https://github.com/abbasilab/compressed-bold-models.

Funding: The author(s) received no specific funding for this work.

Abstract

Voxelwise encoding models based on convolutional neural networks (CNNs) are widely used as predictive models of brain activity evoked by natural movies. Despite their superior predictive performance, the huge number of parameters in CNN-based models have made them difficult to interpret. Here, we investigate whether model compression can build more interpretable and more stable CNN-based voxelwise models while maintaining accuracy. We used multiple compression techniques to prune less important CNN filters and connections, a receptive field compression method to select receptive fields with optimal center and size, and principal component analysis to reduce dimensionality. We demonstrate that the model compression improves the accuracy of identifying visual stimuli in a hold-out test set. Additionally, compressed models offer a more stable interpretation of voxelwise pattern selectivity than uncompressed models. Finally, the receptive field-compressed models reveal that the optimal model-based population receptive fields become larger and more centralized along the ventral visual pathway. Overall, our findings support using model compression to build more interpretable voxelwise models.

Author summary

In this study, we explored the process of simplifying complex brain models and investigated the advantages of this simplification for improved interpretability of the models without losing accuracy. We focused on models that predict brain activity when people watch movies, which are usually based on advanced neural networks. These models are powerful, but they are often too complicated to interpret. By using compression techniques to reduce the size and complexity of these models, we found that they not only remained accurate but also became more stable and easier to understand. Our approach involved trimming unnecessary parts of the model and focusing on the most important areas that respond to visual stimuli. This suggests that simplifying models can help us better understand how the brain processes visual information. Our work highlights the potential of model compression as a tool for making complex scientific findings more accessible and easier to understand for both researchers and the general public.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

A prominent question in computational neuroscience is how sensory information is represented and processed in the visual cortex [1,2] and various computational models have been developed to predict brain activity during sensory tasks. Constructing accurate and data-driven models requires large-scale data collection of the brain's high-dimensional and complex activity. In the past decade, functional magnetic resonance imaging (fMRI) has emerged as a standard technique to record brain activity during natural visual tasks [1–7], and computational models of fMRI blood oxygen level-dependent (BOLD) signals have been developed. Voxelwise encoding models of BOLD signals use sensory stimuli to predict brain activity during various visual tasks and are the focus of this study. On the other hand, decoding models use brain activity to reconstruct and categorize visual stimuli [2,6,8]. Together, these encoding and decoding models provide functional descriptions of cortical areas.

Voxelwise 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 signals from stimulus features (Fig 1A). The feature extraction module has historically been 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]. Recently, however, deep networks and, specifically, convolutional neural networks (CNNs), have emerged as the state-of-the-art encoding models of BOLD signals [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 scale-invariant feature transform or 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 human visual cortex responses. For the response prediction module, 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].

Despite their promising performance, models based on deep CNNs are often extremely hard to interpret. Here, interpretation is defined as the process of extracting meaningful information about domain relationships contained in models output [22–24]. Specifically, millions of parameters and the highly non-linear transformations in these models make them impossible for human observers and domain experts to understand. 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 for interpreting 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 for a domain expert to understand. Additionally, a compressed model requires considerably fewer computational operations, and thus, is faster than the uncompressed model. This reduced computational cost could facilitate the application 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 responses. We use multiple compression techniques applied to both the feature extraction module and the response prediction module. These compression techniques include (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,

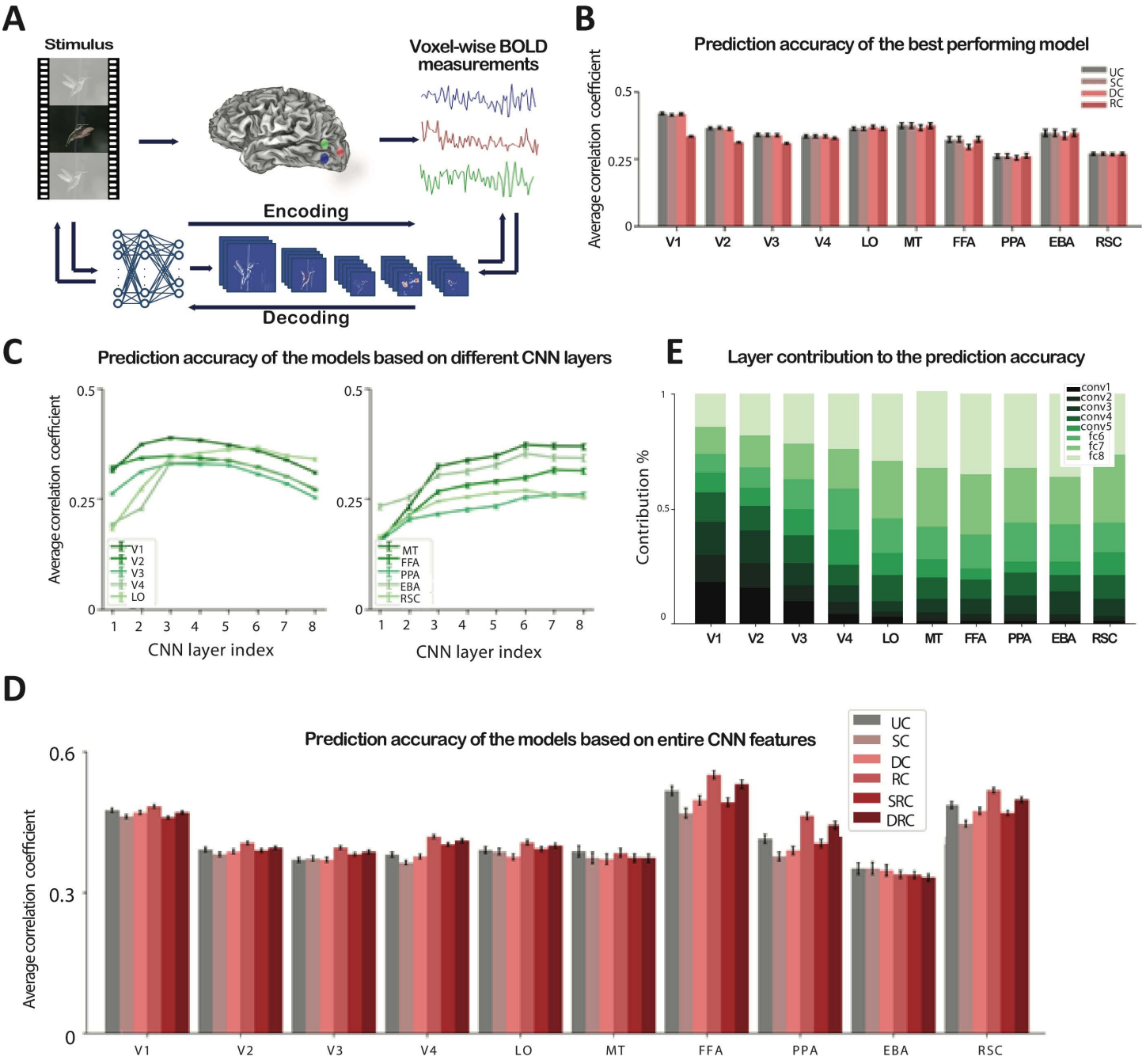

Fig 1. Compressed voxelwise encoding models accurately predict BOLD responses. A. An encoding model predicts the fMRI BOLD responses from the visual stimulus features. The decoding model predicts the optimal stimulus from the BOLD responses. **B.** The Pearson correlation coefficient between estimated response using the best CNN layer and measured fMRI response. The values are averaged over all voxels in each visual area across all subjects in both PURR and Vim-2 datasets and visualized for uncompressed (UC), structurally-compressed (SC), deep-compressed (DC), and receptive field-compressed (RC) models. Error bars show the 95% confidence interval. Compression does not significantly affect the accuracy. **C.** The Pearson correlation between estimated and measured fMRI responses for different CNN layers. Each data point shows the averaged correlation coefficient over all voxels in each visual area for the compressed model. Error bars show the 95% confidence interval. The contribution of the higher CNN layers is attenuated for early visual areas, while the reverse trend is visible for higher visual areas. **D.** Each bar indicates the Pearson correlation coefficient between estimated and measured fMRI response averaged over all voxels in a visual area and across all subjects in both PURR and Vim-2 datasets. The average correlation coefficients are shown for uncompressed (UC), structurally-compressed (SC), deep-compressed (DC) and receptive field-compressed (RC), structurally receptive field-compressed (SRC), and deep receptive field-compressed (DRC) models. Error bars show the 95% confidence interval. Compression does not significantly affect the accuracy. **E.** Each column indicates the contribution of each CNN layer to prediction accuracy. The contributions are normalized to 1. Feature maps extracted from lower CNN layers have a higher contribution to lower visual areas and feature maps extracted from higher CNN layers have a higher contribution to higher visual areas.

https://doi.org/10.1371/journal.pcbi.1012822.g001

and (4) principal component analysis (PCA) [28] to reduce the dimensionality of the model. Using two separate fMRI datasets collected during natural vision from 6 participants, we first demonstrate that the compression of CNN-based voxelwise models reduces their size and computational cost while maintaining their high predictive accuracy. We then show compression improves the accuracy of identifying visual stimuli from the BOLD signal in the hold-out test set. We 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 the compressed models to quantitatively compare the model-based population receptive field sizes and locations along different visual pathways. Our primary contributions are as follows: (1) New insights from the compressed models: Our study specific areas of the visual cortex may respond to a more refined set of features than previously understood. The higher correlation in the selected feature for the SC model indicates that these areas are tuned to a more consistent set of stimuli, which could refine our understanding of feature selectivity in cortical processing. (2) Implications for end-to-end models: One significant barrier to using neural networks directly for encoding models is the limited volume of clean fMRI data available (compared to ImageNet, which is used to train image classification networks.), which is insufficient to train complex networks end-to-end. By reducing the number of parameters through structural compression, we open up possibilities for developing end-to-end models with existing data volumes. This approach could lead to more efficient and potentially real-time applications in brain-computer interfaces and neuroimaging analysis. (3) Future predictions and experiments: The compressed model's ability to highlight a smaller set of influential features provides a pathway for designing future experiments that focus on these specific features, potentially leading to more precise hypotheses about neural coding in different cortical areas. This refined focus can help in removing noise from input data and improving model accuracy.

2. Methods

2.1. Datasets

To build and investigate the compressed voxelwise 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.

2.1.1. 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 a 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 sections of 8-minutes and 24s each. The training set was repeated two times whereas the test set was repeated ten times. 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 [29]. The visual areas were defined with multi-modal cortical parcellation [30]. Additional details on this dataset can be found in [8].

2.1.2. Vim-2 dataset. fMRI data were obtained from three healthy volunteers in a 4T MRI system with a temporal resolution of 1s [31]. 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 [32]. This dataset is further described in [33].

2.2. Feature extraction *via* convolutional neural networks

Our voxelwise encoding models consist of two modules: 1) The CNN-based feature extraction module which constructs a feature set for each frame in the visual stimulus. 2) the response prediction module which predicts the BOLD signal from the CNN-based features.

To extract features from each frame in the visual stimulus, we used AlexNet [34], a well-known CNN model pre-trained on the ImageNet dataset [34]. 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, from which an input stimulus frame is classified into 1000 classes. Layer 1 through layer 5 contain 96 kernels of 11×11, 256 kernels of 27×27, 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. It is important to notice that although this CNN is primarily designed for static images, it can be adapted for video data by treating each frame as a separate input.

For the response prediction module, we used a linear regression model with L_2-norm regularization (Ridge regression [35]). We used each CNN layer output to predict the voxelwise responses and then selected the layer with the highest accuracy for each visual area. Principal component analysis (PCA) was used to reduce the dimension of the features while keeping 99% of the variance in each layer. We also built an encoding model with features from all CNN layers concatenated together. Again, PCA was used to keep 99% of the variance across all layers.

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

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

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

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

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.

2.3. Layer contribution

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

$$C = \frac{cov\left(\phi^l W_v^l, y_v\right)}{var\left(\phi W_v^l\right), var\left(y_v\right)} \tag{3}$$

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.

2.4. 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, i.e., the compression stopped when the hold-out validation set accuracy dropped 2% from the uncompressed model accuracy. Note that the hold-out validation set used for compression is different from 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) [36] 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. Algorithm 1 summarizes the pseudo-code of the structural compression encoding model.

Algorithm 1 Structurally-compressed encoding model

Input: UC encoding model accuracy, sorted CNN filters based on CAR algorithm, target layer L with nl filters, fMRI dataset
Set SC encoding model accuracy = UC encoding model accuracy
Split the fMRI dataset into 3 subsets: training set, validation set, and testing set
While (UC encoding model accuracy - SC encoding model validation accuracy < 0.02):
    Remove least important filter in layer L
    Update the number of filters nl = nl-1
    Build SC model using updated nl filters and training set
    Compute SC model validation accuracy using validation set
    Update UC encoding model accuracy - SC encoding model validation accuracy
End while
Compute SC encoding model accuracy using test set

2.5. Deep compression

We performed deep compression (DC) [27] to reduce the number of connections in the 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 by retraining the network. Similar to structural compression, we further compressed the features using PCA followed by a convolution with the HRF. Finally, the BOLD responses were calculated using a ridge regression model. The model performance was assessed by the Pearson correlation coefficient between measured and predicted responses. Algorithm 2 presents the pseudo-code of the deep compression encoding model.

Algorithm 2 Deep-compressed encoding model

Input: AlexNet weights and classification accuracy for ImageNet, target layer L with nc connections, ImageNet dataset, fMRI dataset
Split the fMRI dataset into 2 subsets: training set and testing set
Set threshold = 1e-10
While (deep compression image recognition accuracy = AlexNet Image recognition accuracy for ImageNet dataset)
    Update threshold = threshold *10
    Remove weights below threshold
    Compute Image classification accuracy for ImageNet dataset for pruned network
    Update deep compression image classification accuracy = Image classification accuracy for ImageNet dataset for pruned network
End while
Use k-means clustering to cluster weights in each layer of the network
Use the same weights for all the connections that fell into the same cluster
Fine-tune the shared weights using ImageNet
Build DC encoding model using the training set
Compute DC encoding model accuracy using test set

2.6. 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 the 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 population receptive fields with a 2D isotropic Gaussian function [16]. Thus, the population receptive field, g, can be described as:

$$g(x,y;x,y,) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu_x)^2(y-\mu_y)^2}{2\sigma^2}\right] \quad (4)$$

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, a dot product is computed between a CNN-bases feature map (dimensions p×p×k) extracted from the stimulus and the 2D Gaussian function, resulting in a k-dimensional vector.Formally

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

where ϕ_{RF}^{l} 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). Similar to the structural and deep compression,

PCA was used to reduce the feature dimension while keeping 99% of the variance. 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.

2.7. Decoding

Our objective in decoding is to determine which stimulus are responsible for voxel activity pattern. To assess the model's decoding ability, we used a test set of various clips that were not seen by the encoding model during training. We then estimated the voxel activity pattern for this test set using the encoding model. Subsequently, for each clip in the test set, we calculated the Pearson correlation coefficient between the estimated response and the measured response and identified the clip that yields the highest correlation coefficient. If the clip was the same as the actual stimulus, we concluded that the model has successfully performed decoding; otherwise, the model was unable to correctly decode.

3. Results

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

While our goal is to improve the interpretability of the CNN-based voxelwise models, we must also maintain the prediction accuracy of compressed models. To quantify prediction accuracy, we computed the voxelwise 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. Principal component analysis (PCA) was integrated with all these compression techniques to reduce the dimensionality of CNN features (see Methods). notably, our compressed models achieved significant reductions in size, up to 27x compared to uncompressed model (details in Section 3.2). We have systematically compared the prediction accuracy of these compressed models with the uncompressed CNN-based model for each visual area. 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).

On average, the structurally- and deep-compressed models performed as well as the uncompressed model for most visual areas. The accuracies for the best-performing models among 8 distinct models that are built based on each layer of the CNN are presented in Fig 1B and 1C, while the accuracies using the entire set of CNN features from all layers are presented in Fig 1D and 1E. The average correlation coefficient is reported across all the voxels within each visual area. For FFA, the deep-compressed model underperforms other models by 3%, most likely due to the higher sensitivity of face features to the pruning of 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 similar accuracy to the uncompressed model for the higher visual areas. Fig 1C illustrates the predictive accuracy of the models based on each individual CNN layer

for structurally-compressed models (as opposed to Fig 1B where the best-performing model is selected). 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 are the most accurate.

We further used the features extracted from all 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, receptive-field compression is used after the structural or deep compression. The prediction accuracies reported in Fig 1D suggest that using the entire feature map to predict the BOLD signal results in higher accuracies (92% on average) compared to regressing single layers of CNNs. This is not surprising because a larger number of features with complementary image statistics (from different layers of CNN) are used in these models compared to the single-layer models.

We also repeated the structural compression and deep compression analyses for randomly pruned filters with the same amount of pruning. Over 10 repetitions of this procedure, the average correlation coefficient score dropped by 6%, 4% for V1, 7%, 4% for V2, 8%, 5% for V3, 8%, 5% for V4, 4%, 6% for LO, 3%, 5% for MT, 2%, &% for FFA, 1%, 5% for PPA, 0.5%, 6% for EBA, and 1%, 7% for RSC, for structural and deep compression, respectively.

For the best-performing models in Fig 1D, we also quantified the contribution of each layer to the prediction accuracy for each visual area (Fig 1E). Overall, our results suggest that the average contribution of the extracted feature maps to predicting the 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 a 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 voxelwise accuracies based on the entire CNN layers on the cortical map in one of the subjects (see S1-S8 Figs 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 Fig 2. The structurally-compressed model outperforms the uncompressed model in ProS, DVT, and the lateral part of the V1, suggesting that these regions are better modeled using fewer 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 V1, V2, V3, V4, and also TPOJ and FST areas are modeled more accurately compared to the uncompressed model. Other subjects display similar results (S1-S8 Figs).

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

So far, we have demonstrated that the compressed encoding models (see Methods) retain a satisfactory prediction accuracy. Here, we examine the model size, the computational cost, and the compression ratio 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

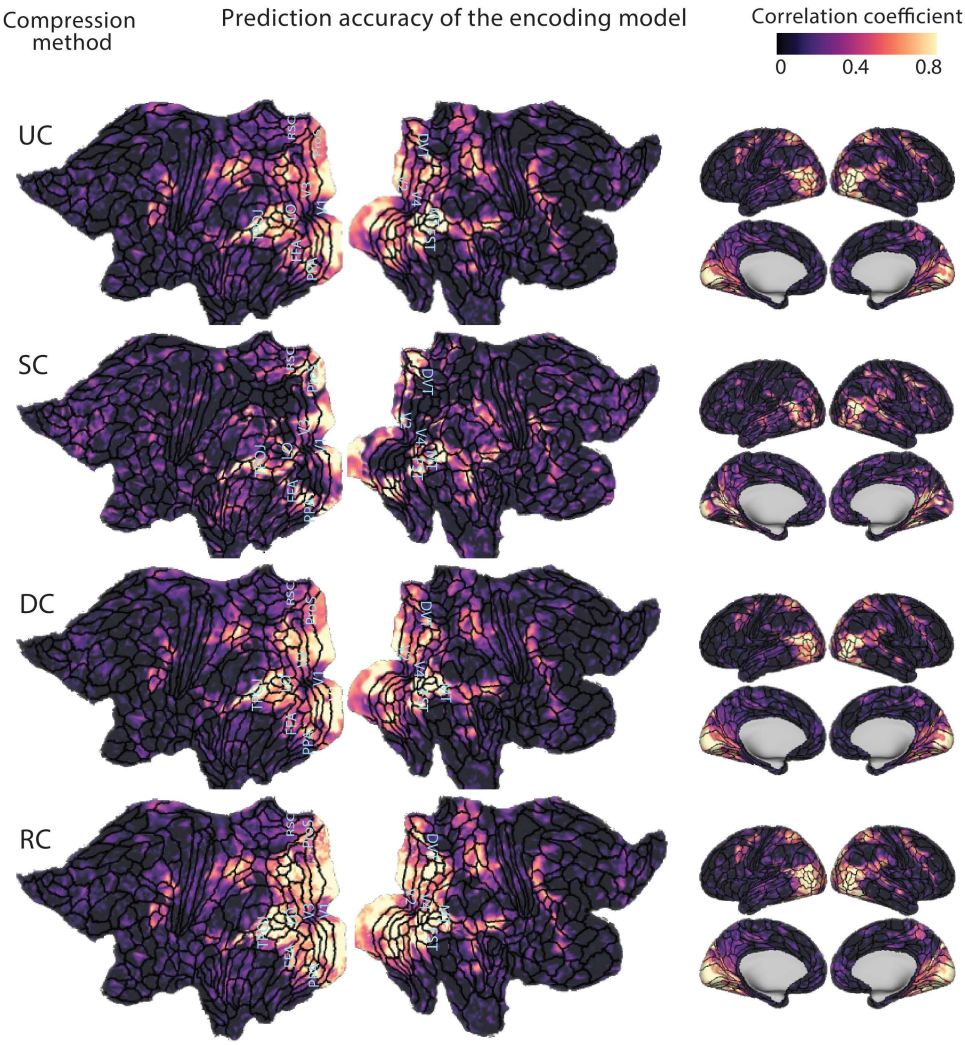

Fig 2. Prediction accuracy of the compressed encoding models on cortical maps. The maps show the correlations between estimated and measured responses for the uncompressed model (UC), structurally-compressed model (SC), deep-compressed model (DC), and receptive field-compressed model (RC) for subject 1 in the PURR dataset. Compared to uncompressed models, structurally-compressed models better estimate fMRI responses in ProS, DVT, and the lateral part of V1. The deep-compressed model is better in the central part of V1, V2, and V3, and the lateral part of V4. The receptive field compressed model better estimates the lateral part of V1, V2, V3, and V4, as well as TPOJ and FST areas.

https://doi.org/10.1371/journal.pcbi.1012822.g002

the number of floating point operations in that layer to extract all of the feature maps, without accounting for the regression overhead. The compression ratio was computed by dividing the number of FLOPS (or weights) required for the uncompressed model by that of the compressed model.

Our compressed encoding models have a reduced model size and a remarkably lower computational cost compared to the uncompressed models (Table 1). For structural compression, the model based on the CNN layer 4 has the highest compression ratio (compression ratio of 2) and the model based on layer 5 has the lowest compression ratio (ratio of 1.2), suggesting a high redundancy across layer 4 filters. Note that the compression ratio for structural compression is reported for a compressed model with a validation set accuracy that is less than 2%

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, i.e., the number of floating-point operations required in each layer to classify one image. The compression ratio is defined as the number of FLOPS (or weights) required for the uncompressed model divided by those required for the compressed model. Note that for the structurally-compression model, the compression ratio based on the number of FLOPS is equal to the compression ratio based on the number weights because all filters are removed during this form of compression. Structural compression is also not defined for the fully connected layers; therefore no number is reported for these layers.

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	52.24M	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	61.09K	7.14
fc6	Uncompressed	37.75M	–	38M	–
	Deep	2.33M	16.18	1.14M	33.33
fc7	Uncompressed	16.78M	–	17M	
	Deep	0.98M	17.12	0.51M	33.33
fc8	Uncompressed	4.10M	–	4M	
	Deep	0.53M	7.71	0.29M	13.69
All layers together	Uncompressed	724.37M	–	61M	–
	Structural	394.79M	1.8	33.8M	1.8
	Deep	300.79M	2.4	2.25M	27

https://doi.org/10.1371/journal.pcbi.1012822.t001

of the uncompressed model (see Methods). For the deep compression, the model based on the second fully connected layer has the highest compression ratio in terms of the number of FLOPS (ratio of 17.12), while models based on both the first and second fully connected layers have the highest compression ratio in terms of the number of weights (ratio of 33.33). These extraordinarily high compression ratios suggest the considerable amount of weight redundancy in fully connected layers of CNN. 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). For this model, the high compression ratio for the weights is due to the inclusion of fully connected layers in the model design. 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. However, it is worth noting that across different compression techniques, PCA explains 99% of the variance using 6664 to 10917 PCs for the convolutional layers, 2969 to 3463 PCs for the fully connected layers, and 241 components for the final softmax layer. For the models that are built based on all CNN layers,

PCA explains 99% of variance using 1320-2716 PCs across different compression techniques. Overall, our findings suggest that compression methods offer a reduced model size and computational cost.

3.3. Compressed models identify natural movie frames from BOLD responses more accurately than uncompressed models

So far, we have established that compressed models reduce computational cost while retaining a high predictive accuracy compared to their uncompressed counterparts. We now investigate whether compression improves the model's ability to decode visual stimuli from brain activity (BOLD signal). To test this, we compared the ability of the compressed and uncompressed models to identify which video frame was shown to the subject during a hold-out test set with 598 clips. We first used structurally-compressed (SC), deep-compressed (DC), receptive field-compressed (RC), and the uncompressed (UC) model to predict voxelwise activity patterns for voxels in V1, V2, V3, V4, LO, MT, FFA, PPA, EBA and RSC areas evoked by each clip in the hold-out test set. We then selected the video clip with the highest Pearson correlation coefficient between predicted and the measured BOLD signal. The identification performance was obtained as the ratio of the number of correctly selected clips to the total number of clips. The average identification performance across all subjects in the PURR dataset is 91%, 90.6%, 92%, and 93% for UC, SC, DC, and RC models, respectively. For the vim-2 dataset, the identification performance was 81.6%, 83.3%, 85%, and 87% for UC, SC, DC, and RC models, respectively. Overall, the RC model performed better in identification accuracy than the uncompressed model and the other compressed models. Note that the chance-level performance is 0.1% for subjects in the PURR dataset and 1% for the subjects in the vim-2 dataset.

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

An established method for the interpretability of a voxelwise model is visualizing the images that elicit the largest response in each voxel (i.e., optimal visual patterns) [3]. This form of interpretability may be hindered for CNN-based models due to their large number of parameters [25]. More specifically, the optimal visual patterns for each voxel could be *unstable* when a large model is used to identify or generate that pattern [37]. Here, stability is defined as the similarity of visual patterns among the images with the highest model responses. This form of stability is a requirement for the reliable interpretation of predictive models for scientific discovery [22]. Considering the lower number of parameters in the structurally-compressed model, we hypothesize that the images that elicit the largest response in each voxel are more stable for the structurally-compressed model compared to the uncompressed model.

To determine the optimal visual pattern for each voxel, we searched for the natural images that elicit the largest response for both the uncompressed and structurally-compressed models. We chose 10,000 natural images from the validation set in the ILSVRC 2012 dataset [34] and computed the voxelwise response of both 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 the CNN nor included in the experimental stimulus set. Overall, these images provided a qualitative representation of the patterns selected by each voxel. The patterns for the voxels in V1 and V2 did 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 [3,13,38]. The top images for the V4 area contained semi-complex shapes such as curvatures, circles, crosses, and dense textures. For FFA, the top 5 images

included face features while the top images for PPA and RSC contained environmental scenes. Notably, we found a significant overlap in preferred images between uncompressed (uc) and compressed (ss) models across various brain areas: V1: 55%, V2: 50.3%, V4: 41.4%, FFA: 81%, PPA: 75.4%, RSC: 73.2% for the top 50 images. These findings are aligned with observations in previous studies [37], suggesting both uncompressed and compressed models are successful in determining the preferred images for the accurately predicted voxels.

We then examined the stability of the preferred stimulus images for each voxel and compared 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 [34]. Formally, the stability score is defined as the average pairwise Pearson correlation coefficient between CNN-extracted features across images. We empirically chose to compute the stability score over the top 10 images with the highest model response because this was sufficient to convey the overall pattern selectivity of the voxel. This score was computed for the most accurately predicted voxels (top 100) within each area, for both structurally-compressed and uncompressed models. We chose only 100 top voxels to ensure that our stability analysis remained unbiased due to potential predictive inaccuracies. The top 100 voxels selected for stability analysis show over 80% overlap between the uncompressed (UC) and compressed (SC) models. However, to account for the worst-case scenario—given that stability is generally higher for the SC model—we chose the top voxels based on the UC model.

Scatterplots comparing these stability scores indicate that the stability is considerably higher for the structurally-compressed model in V1, V2, V3, V4, LO, MT, FFA, and RSC outperforming the uncompressed model by 0.07, 0.06, 0.04, 0.05, 0.06, 0.04, 0.05, and 0.06, respectively (Fig 3). The stability scores in PPA are only marginally higher for the 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 uncompressed model – the compressed model has a higher correlation coefficient for more than 81% of the top voxels across all visual areas. These results indicate that the compressed model provides more stable features than the uncompressed model for the majority of accurately predicted voxels.

To further examine our approach we repeated the procedure for top 20 and 50 images. The result drop 0.02–0.05 for 20 top frames and 0.05–0.09 for top 50 images. we conducted additional analyses using ground truth simulations (input video frame shown to the subjects). For op 10 frames lead to highest output, average correlation are 0.31, 0.33, 0.32, 0.31, 0.31, 0.30, 0.32, 0.31, 0.31 for V1, V2, V3, V4, LO, MT, FFA,PPA and RSC respectively. The result drop 0.03 - 0.07 for 20 top frames and 0.05-0.1 for top 50 images.

We then asked whether the stability is a consequence of structural compression or any reduction in the number of parameters. To answer this question, we repeated the stability analysis for randomly-pruned filters with the same compression rate as structural compression. Averaged over 100 repeats of this random pruning procedure, the stability score dropped by 8% for V1, 8% for V2, 6% for V3, and 4% for V4, suggesting the significance of structural compression in improved stability for these areas. The stability score remained the same for RSC, LO, MT, FFA, and PPA, which is not surprising because these areas are better modeled by fully connected layers of CNN (layers 6, 7, and 8, see Fig 1B) which are not affected by structural compression.

Overall, our findings indicate that structurally-compressed models allow for a more stable interpretation of pattern selectivity for each voxel. Consistent with prior studies [3,19], our findings indicate that the downstream areas in the ventral visual pathway are more category-selective compared to early areas.

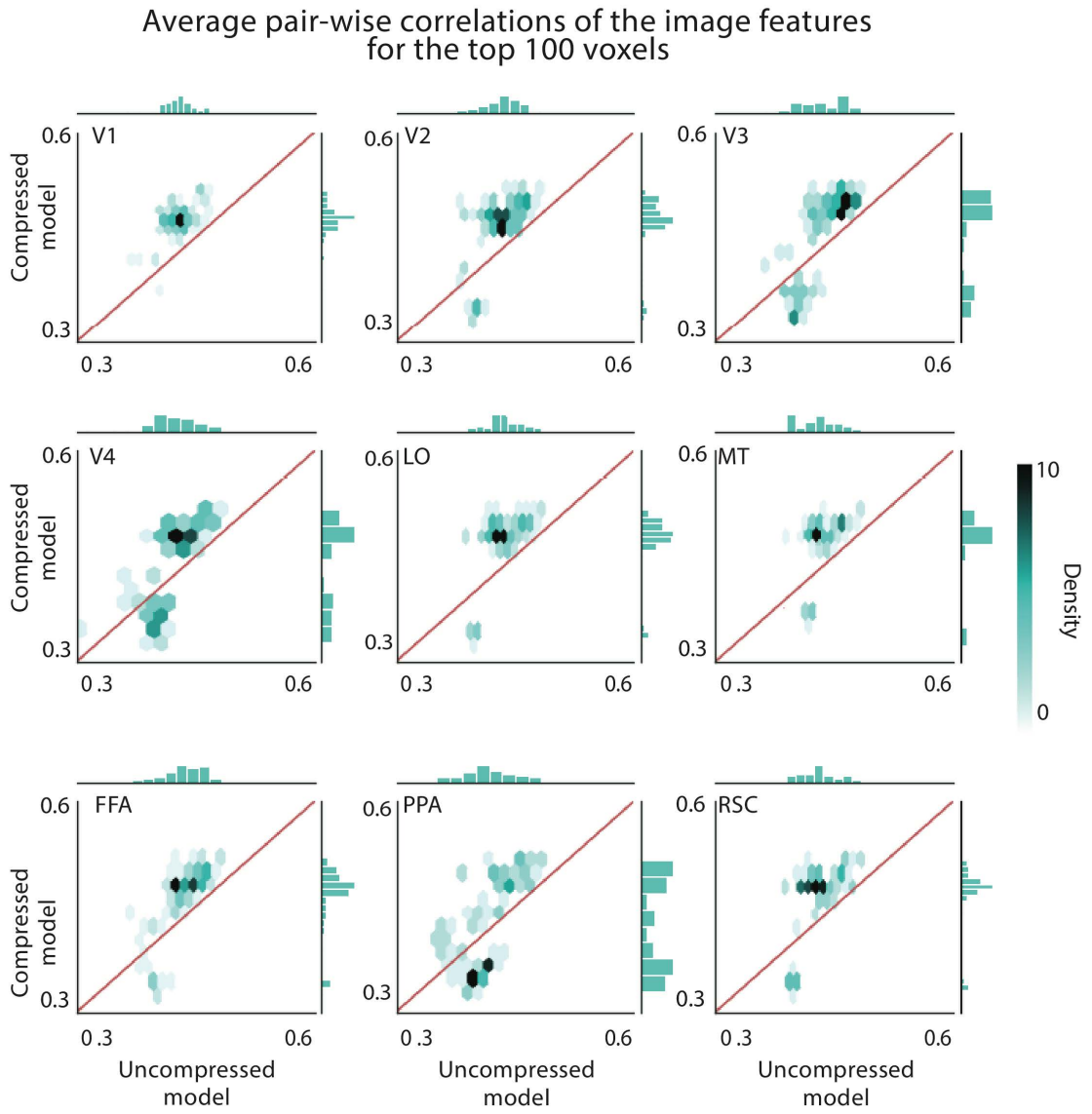

Fig 3. Structurally-compressed models reveal more stable category selectivity compared to the uncompressed models. Density plots comparing the stability of the top 5 images with the highest model response in each visual area between compressed and uncompressed models. The stability is quantified by the average pairwise correlation coefficient between CNN-extracted features of the top 10 selected images for each voxel. The average correlation coefficient is reported across the top 100 voxels for 9 visual areas. The color of each hexagonal bin indicates the density of voxels in that bin. The histograms below 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.

https://doi.org/10.1371/journal.pcbi.1012822.g003

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

The organization of the population receptive field maps (characterized by neural population-level measurements such as BOLD signal) in different areas of the visual cortex have been extensively studied in the past [16,39,40]. These studies have provided evidence for larger and more centralized receptive fields in higher visual areas along the ventral visual pathway.

We confirmed this observation using the receptive-field-compressed encoding models. These models allow for a systematic and quantitative analysis of the optimal size and location of the model-based population receptive fields in different visual areas. Fig 4A illustrates the selected population 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 population receptive fields while higher visual areas have larger and more centralized receptive fields. To quantitatively compare the population 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 (Fig 4B). We found that V1 has the smallest average population receptive field size (mean radius: 1.6 ± 0.02 degrees) with the most scattered centers (mean eccentricity: 7.13 ± 0.04 degrees). V2, V3, and V4 exhibit slightly larger mean receptive field size (mean radius: 1.75 ± 0.02 and 1.77 ± 0.02, and 1.86 ± 0.03 degrees, respectively) with gradually increasing centralization (mean eccentricity: 7.08 ± 0.04, 6.64 ± 0.04, and 6.10 ± 0.04 degrees, respectively). On the other hand, voxels in FFA and PPA had the largest population receptive fields (mean radius: 3.33 ± 0.08 and 3.40 ± 0.07 degrees, respectively) followed by LO, MT, and EBA (mean radius: 2.22 ± 0.04 and 2.66 ± 0.07, and 2.55 ± 0.09 degrees,

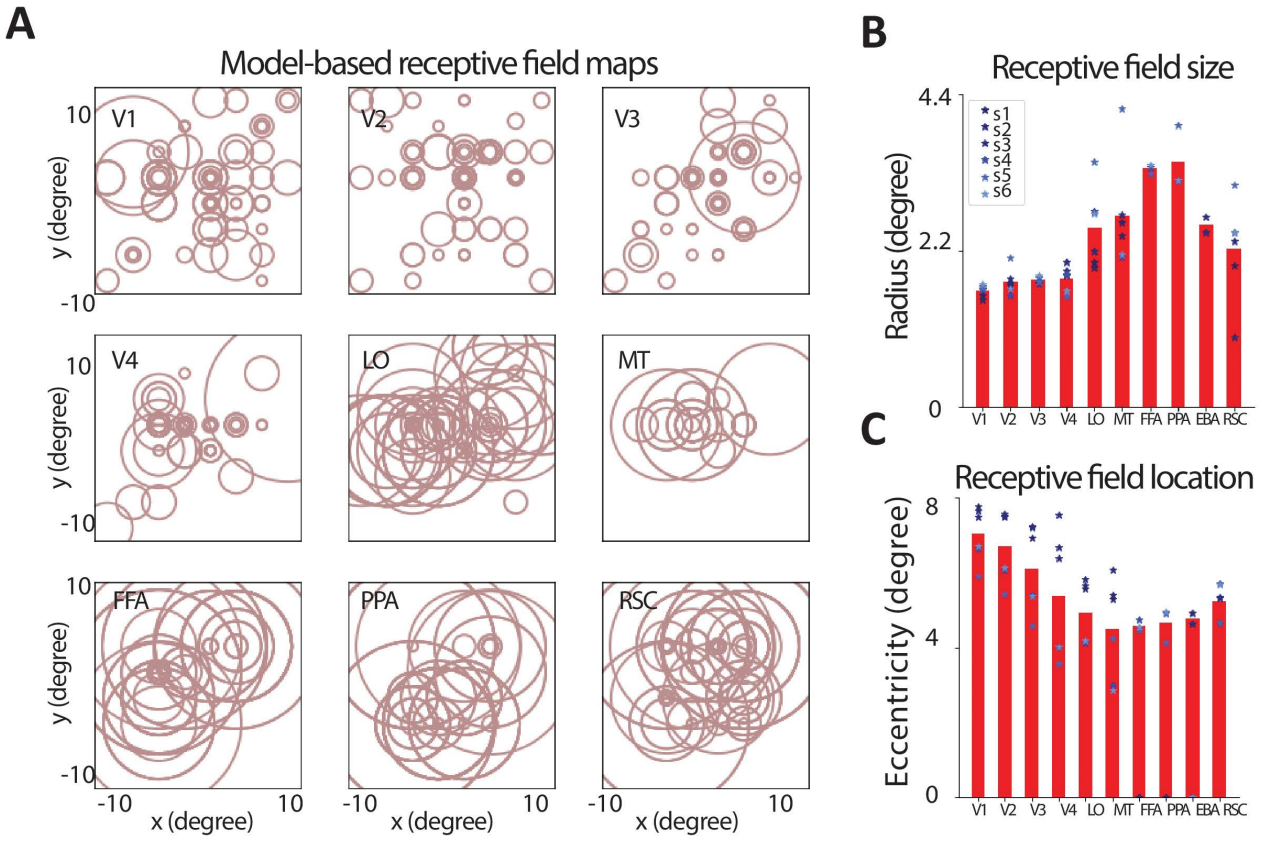

Fig 4. The receptive-field-compressed models reveal increased size and centralization of the population receptive fields along the ventral visual pathway. A. Each circle illustrates the receptive field for an individual voxel from the top 100 most accurately predicted voxels in subject 1 from the PURR dataset. The radius and the center of each circle are determined by the radius and the center of the Gaussian pooling field selected through receptive field compression. **B.** The mean radius of receptive fields in each visual area. **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 toward the downstream regions in the visual area.

https://doi.org/10.1371/journal.pcbi.1012822.g004

respectively). The population receptive fields in MT, FFA, PPA, and EBA were the most centralized (mean eccentricity: 5.27 ± 0.09, 4.58 ± 0.08, 4.68 ± 0.07 and 4.78 ± 0.09, respectively). Overall, the receptive field-compressed encoding model provides a quantitative and systematic framework to compare the population receptive field sizes and centers along the visual hierarchy. These models systematically confirm findings by previous studies that the population receptive fields become larger and more concentrated as we move downstream in ventral pathways.

4. 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 and reduces the computational cost while preserving accuracy. We used structural compression to remove less important filters, deep compression to remove less important connections, and 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. Furthermore, we show that structurally compressed models provide a more stable framework to investigate visual pathways. Here, stability is defined as the similarity between different images that activate each voxel in the visual pathway. We believe that It is crucial to establish stability as a prerequisite for reliable interpretations, as misinterpretation can easily occur when attempting to interpret or draw conclusions from unstable or inconsistent images. Therefore, it is essential to prioritize stability when analyzing visual patterns and their corresponding voxel activations. By doing so, we enhance the validity and accuracy of our interpretations while minimizing the risk of misinterpretation. Overall, compared to previous works [3,8] we provide simpler and more stable encoding models, while maintaining a similar predictive performance.

Overall, we demonstrated that the small set of visual stimulus features identified by compression could accurately predict the BOLD signal. However, modeling higher visual areas (e.g., FFA) that encode more complex visual patterns (e.g., faces) may still require a larger number of features. This is evident from our finding for FFA which showed high accuracy for a model that uses all layers of CNN (correlation coefficient of 0.52 ± 0.009) compared to the best-performing individual-layer model (correlation coefficient of 0.30 ± 0.009) (Fig 1D). This suggests that FFA models require a diverse set of features including both low-level and high-level image statistics, which is not surprising due to the complexity of the face images that are primarily encoded by FFA.

Our findings suggest that combining multiple compression strategies could increase the compression ratio while maintaining predictive accuracy (Fig 1D). Receptive field compression combined with structural compression, however, could introduce aggressive feature pruning in the design of the predictive model, leading to a partial reduction in predictive performance (Fig 1D). For the structural compression, our findings confirmed previous studies [25] that the convolutional layers in a CNN such as AlexNet contain redundant but diverse filters. Here we reduced the redundancy but maintained the diversity [25], which is key for the high accuracy of the structurally-compressed models, especially for early visual areas.

In light of the recent release of the Natural Scenes Dataset (NSD) [41], we anticipate that our results could be further validated with this higher-resolution dataset. The NSD employs a sampling strategy designed to maximize the number of distinct images, providing a richer set of stimuli for analysis. Additionally, the experiment conducted across 40 scan sessions for each subject may enhance the accuracy of the fMRI measurements, allowing for a more

detailed exploration of brain responses to visual stimuli. We speculate that our findings regarding the stability of UC and SC models would likely hold true with the NSD dataset. The increased resolution may allow for a more nuanced representation of visual features, potentially leading to improved model performance. Additionally, the fact that the experiment was split across 40 scan sessions for each subject suggests a greater accuracy in capturing individual differences in brain responses, which could enhance the reliability of our stability metrics. The larger and more diverse set of images in the NSD may also contribute to a richer exploration of feature representation and stability across varying stimuli. We anticipate that the SC model would continue to demonstrate higher stability in this context, as it is designed to filter out irrelevant features while preserving critical information.

In terms of the encoding model design, we constructed a two-step encoding model that consists of an image stimulus feature extraction module and a BOLD response prediction module. The 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 voxelwise responses. This end-to-end model architecture could present a simpler model design and training but will require an order of magnitude larger fMRI dataset than what was used here. We expect that the compression methods used in this study would allow for training such an end-to-end model with limited fMRI data. This was beyond the scope of this study but is a direction for future follow-up studies.

Furthermore, our voxel-wise compressed models provide a pathway for designing more enhanced experiments using fMRI data. The ability of compressed models to select a smaller set of influential features and understanding the pattern selectivity of these select features enables the experimentalists to explore interpretable relationships between stimulus and brain data. Additionally, the compressed models are more stable and have a smaller number of features in their architecture exhibiting an enhanced robustness to noise or perturbations in the input data. This improves model reliability in real-world applications where data may not be perfectly clean.

Finally, 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 a variety of tasks such as classification, detection, and recognition. A more accurate encoding model would aggregate features relevant to these 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 decrease the feature space and allow for more effective model interpretation.

Supporting information

S1 Fig. 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 deep-compressed (DC) model is more accurate in the central part of V1, V2, V3, and the lateral part of V4. The receptive field-compressed (RC) model better estimates the lateral part of V1, V2, V3, V4, as well as the TPOJ and FST areas.
(PDF)

S2 Fig. The difference between the prediction accuracy of compressed encoding models and uncompressed encoding models.
(PDF)

S3 Fig. 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 techniques.
(PDF)

S4 Fig. Scatterplots comparing voxelwise 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 techniques.
(PDF)

S5 Fig. Scatterplots comparing voxelwise 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 techniques.
(PDF)

S6 Fig. Scatterplots comparing voxelwise 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 techniques.
(PDF)

S7 Fig. Scatterplots comparing voxelwise 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 techniques.
(PDF)

S8 Fig. Scatterplots comparing voxelwise 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 techniques.
(PDF)

Acknowledgments

The authors would like to thank Gavin Cui, Gaurav Ghosal for their valuable feedback on the manuscript.

Author contributions

Conceptualization: Fatemeh Kamali, Amir Abolfazl Suratgar, Mohammadbagher Menhaj, Reza Abbasi-Asl.

Data curation: Fatemeh Kamali.

Formal analysis: Fatemeh Kamali, Reza Abbasi-Asl.

Methodology: Fatemeh Kamali.

Software: Fatemeh Kamali, Reza Abbasi-Asl.

Supervision: Amir Abolfazl Suratgar, Mohammadbagher Menhaj, Reza Abbasi-Asl.

Validation: Fatemeh Kamali.

Visualization: Fatemeh Kamali, Amir Abolfazl Suratgar, Reza Abbasi-Asl.

Writing – original draft: Fatemeh Kamali, Reza Abbasi-Asl.

Writing – review & editing: Amir Abolfazl Suratgar, Mohammadbagher Menhaj.

References

1. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;452(7185):352–5. https://doi.org/10.1038/nature06713 PMID: 18322462
2. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011;56(2):400–10. https://doi.org/10.1016/j.neuroimage.2010.07.073 PMID: 20691790
3. Agrawal P, Stansbury D, Malik J, Gallant JL. Pixels to voxels: modeling visual representation in the human brain. arXiv:1407.5104 [Preprint]. 2014. Available from: https://doi.org/arXiv:1407.5104
4. Qiao K, Zhang C, Chen J, Wang L, Tong L, Yan B. Neural encoding and interpretation for high-level visual cortices based on fmri using image caption features. arXiv:2003.11797 [Preprint]. 2020. Available from: https://doi.org/10.48550/arXiv.2003.11797
5. Han K, Wen H, Shi J, Lu K, Zhang Y, Fu D, et al. Variational autoencoder: an unsupervised model for encoding and decoding fMRI activity in visual cortex. Neuroimage. 2019;198:125–36.
6. Naselaris T, Prenger R, Kay K, Oliver M, Gallant J. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009;63(6):902–15.
7. Nunez-Elizalde AO, Huth AG, Gallant JL. Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage. 2019;197:482–92.
8. Wen H, Shi J, Zhang Y, Lu K, Cao J, Liu Z. Neural encoding and decoding with deep learning for dynamic natural vision. Cereb Cortex. 2018;28(12):4136–60.
9. Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell. 2006;28(12):2037–41. https://doi.org/10.1109/TPAMI.2006.210
10. Güçlü U, van Gerven MAJ. Modeling the dynamics of human brain activity with recurrent neural networks. Front Comput Neurosci. 2017;11:7. https://doi.org/10.3389/fncom.2017.00007 PMID: 28232797
11. Cui Y, Zhang C, Wang L, Yan B, Tong L. Dense-gwp: an improved primary visual encoding model based on dense gabor features. J Mech Med Biol. 2021;21(05):2140017.
12. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014;10(11):e1003915. https://doi.org/10.1371/journal.pcbi.1003915 PMID: 25375136
13. Güçlü U, Van Gerven M. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci. 2015;35(27):10005–14. https://doi.org/10.1523/JNEUROSCI.0450-15.2015
14. Güçlü U, van Gerven M. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. Neuroimage. 2017;145(1):329–36.
15. Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: convolutional network layers map the function of the human visual system. Neuroimage. 2017;152:184–94. https://doi.org/10.1016/j.neuroimage.2016.10.001 PMID: 27777172
16. St-Yves G, Naselaris T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. Neuroimage. 2018;180(Pt A):188–202. https://doi.org/10.1016/j.neuroimage.2017.06.035 PMID: 28645845
17. Seeliger K, Fritsche M, Güçlü U, Schoenmakers S, Schoffelen J, Bosch S, et al. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. NeuroImage. 2018;180:253–66.
18. Yu Z, Zhang C, Wang L, Tong L, Yan B. A comparative analysis of visual encoding models based on classification and segmentation task-driven CNNs. Comput Math Methods Med. 2020;2020(1):5408942. https://doi.org/10.1155/2020/5408942
19. Shi J, Wen H, Zhang Y, Han K, Liu Z. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. Hum Brain Mapp. 2018;39(5):2269–82.
20. Wen H, Shi J, Chen W, Liu Z. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Sci Rep. 2018;8(1):3752. https://doi.org/10.1038/s41598-018-22160-9 PMID: 29491405
21. Qiao K, Zhang C, Wang L, Yan B, Chen J, Zeng L, et al. Accurate reconstruction of image stimuli from human fMRI based on the decoding model with capsule network architecture. arXiv:1801.00602 [Preprint]. 2018. Available from: https://doi.org/10.48550/arXiv.1801.00602

22. Murdoch W, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116(44):22071–80. https://doi.org/10.1073/pnas.1900654116
23. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 [Preprint]. 2017. Available from: https://doi.org/10.48550/arXiv:1702.08608
24. Ghosal G, Abbasi-Asl R. Multi-modal prototype learning for interpretable multivariable time series classification. arXiv:2106.09636 [Preprint]. 2021. Available from: https://doi.org/arXiv:2106.09636
25. Abbasi-Asl R, Yu B. Structural compression of convolutional neural networks with applications in interpretability. Front Big Data. 2021;4704182. https://doi.org/10.3389/fdata.2021.704182 PMID: 34514381
26. Abbasi-Asl R, Yu B. Interpreting convolutional neural networks through compression. arXiv:1711.02329 [Preprint]. 2017.
27. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv:1510.00149 [Preprint]. 2015. Available from: https://doi.org/arXiv:1510.00149
28. Zhao H, Yuen P, Kwok J. A novel incremental principal component analysis and its application for face recognition. IEEE Tran Syst Man Cybern Part B (Cybern). 2006;36(4):873–86.
29. Glasser M, Sotiropoulos S, Wilson J, Coalson T, Fischl B, Andersson J, et al. The minimal preprocessing pipelines for the human connectome project. NeuroImage. 2013;80:105–24.
30. Glasser M, Coalson T, Robinson E, Hacker C, Harwell J, Yacoub E. A multi-modal parcellation of human cerebral cortex. Nature. 2016;536(7615):171–8. https://doi.org/10.1038/nature18933
31. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B. Gallant lab natural movie 4T fMRI data. CRCNS.org. 2014. Available from: http://dx.doi.org/10.6080/K00Z715X
32. Hansen K, David S, Gallant J. Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. Neuroimage. 2004;23(1):233–41.
33. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol. 2011;21(19):1641–6. https://doi.org/10.1016/j.cub.2011.08.031 PMID: 21945275
34. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25.
35. McDonald G. Ridge regression. Wiley Interdiscip Rev Comput Stat. 2009;1(1):93–100.
36. Henson R, Friston K. Convolution models for fMRI. In: Statistical parametric mapping: the analysis of functional brain images. Elsevier; 2007.
37. Abbasi-Asl R, Chen Y, Bloniarz A, Oliver M, Willmore BD, Gallant JL. The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. bioRxiv. 2018:465534.
38. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962;160(1):106–54. https://doi.org/10.1113/jphysiol.1962.sp006837 PMID: 14449617
39. Dumoulin SO, Wandell BA. Population receptive field estimates in human visual cortex. Neuroimage. 2008;39(2):647–60.
40. Klink P, Chen X, Vanduffel W, Roelfsema P. Population receptive fields in nonhuman primates from whole-brain fMRI and large-scale neurophysiology in visual cortex. Elife. 2021;10:e67304.
41. Allen E, St-Yves G, Wu Y, Breedlove J, Prince J, Dowdle L. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat Neurosci. 2021;25(1):116–26. https://doi.org/10.1038/s41593-021-00800-0