A finer mapping of convolutional neural network layers to the visual cortex

Anonymous Author(s)
Affiliation
Address
email

Abstract

There is increasing interest in understanding similarities and differences between convolutional neural networks (CNNs) and the visual cortex. A common approach is to use features extracted from intermediate CNN layers to fit brain encoding models. Each brain region is then typically associated with the best predicting layer. However, this winner-take-all mapping is non-robust, because consecutive CNN layers are strongly correlated and have similar prediction accuracies. Moreover, the winner-take-all approach ignores potential complementarities between layers to predict brain activity. To address this issue, we propose to fit a joint model on all layers simultaneously. The model is fit with banded ridge regression, grouping features by layer, and learning a separate regularization hyperparameter per feature space. By performing a selection over layers, this model effectively removes non-predictive or redundant layers and disentangles the contributions of each layer on each voxel. This model leads to increased prediction accuracy and to finer mappings of layer selectivity.

Introduction

Convolutional neural networks (CNNs) were inspired originally by the anatomy of the brain, and they have been remarkably successful in computer vision [1, 2]. However, these networks still fail in many tasks that humans can perform easily. Therefore, there is increasing interest in understanding the similarities and differences between CNNs and the brain. To investigate this issue, a common method is to use some specific layer of a pre-trained CNN as a source of features to fit a brain encoding model [3]. With this approach, many studies have shown that early CNN layers best predict brain activity in low-level visual areas, while late layers best predict brain activity in intermediate and higher-level visual areas, with gradual changes of layer selectivity over the cortical surface [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. A similar approach has also been applied to speech [14, 15] and language tasks [16, 17].

A notable difficulty encountered in this approach is the strong correlation of CNN activations from one layer to the next. This confound causes different layers to have similar predictive power in encoding models [8, 16, 17, 18], and it is hard to separate which part of the predictive power is specific to a layer and which part is shared with other layers. Most studies ignore this issue and select the best-predicting layer for each voxel [5, 6, 8, 9, 14, 19, 11], but this winner-take-all approach is not robust and ignores potential complementarities between layers. Some studies use variance partitioning [20] or canonical component analysis [21] to disentangle the different layers, but these approaches are limited in terms of number of layers that can be disentangled.

To address this issue, we use banded ridge regression, which has been shown to disentangle contributions from correlated feature spaces in encoding models [22]. Specifically, we fit a predictive model

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

using features from all layers at once, grouping the features by layers, and learning optimal regularization for each layer through cross-validation. We show empirically that this joint model performs a selection over layers, effectively removing non-predictive or redundant layers, and disentangling the contributions of each layer on each voxel. Using this joint model increases prediction accuracy, and leads to smoother cortical maps of layer selectivity. These methodological improvements pave the way for high-precision mappings between CNNs and human brains, which may help both designing and interpreting CNNs, and defining high-resolution information pathways over the cortical surface.

1 Conventional approach

The conventional approach for mapping CNN layers to brain regions [4, 5, 6, 7, 8, 9, 19, 11, 12, 13] follows the voxelwise encoding model framework [23, 24]. First, brain activity is recorded while subjects perceive a visual stimulus. Then, the same stimulus is presented to a pretrained CNN, activations are extracted from intermediate CNN layers and preprocessed into features (see below). Finally, a regression model is trained on each voxel to predict brain activity from the features.

1.1 Feature extraction

To extract features, each image of the stimulus is first presented to a pretrained CNN. Then, the activations of one convolutional or fully-connected layer are extracted (typically after ReLU and max-pooling layers). Next, a compressive nonlinearity $x \mapsto \log(1 + x)$ is applied, and features are centered individually along the train set. Then, to account for the delay between the stimulus and the hemodynamic response, features are either convolved with a hemodynamic response function, or duplicated with multiple temporal delays. This process is repeated for each CNN layer.

Limitations. This process has been used with both static images [5, 6, 7, 8, 11, 13] and video stimuli [8, 9]. Video stimuli are closer to the human vision naturalistic experience and should thus be preferred. With a video stimulus, features from each frame must be down-sampled to the fMRI sampling frequency (typically 0.5 Hz). This down-sampling is done with an anti-aliasing low-pass filter [8, 9]. However this low-pass filter likely removes valuable information contained in the CNN activations. Indeed, a video stimulus can induce brain activity linked to movement, and this activity has been shown to be poorly predicted by low-pass filtered static features [25].

1.2 Winner-take-all model

In the conventional approach, a separate ridge regression model [26] is fit on each layer independently. (Here we use the term "layer" to refer to the features extracted from one CNN layer.) Then, the best layer is selected for each voxel, based on cross-validated prediction accuracy. Finally, differences in "layer selectivity" in terms of the selected best layer are analyzed across the brain.

Limitations. When two layers lead to very similar prediction accuracies, the best-layer selection can result from arbitrarily small differences in prediction accuracy, leading to a non-robust layer selectivity. The winner-take-all approach also ignores the possibility that one low-performing layer might contain information that is not captured by the best predicting layer and that might be relevant for brain prediction in conjunction with it. In this case, both layers would be complementary, in the sense that using both of them jointly would lead to better performance than any single one individually.

2 Proposed approach

2.1 Temporal filtering

To address the limitations of the feature extraction from video stimuli, we propose to filter the layer activations with eight complex-valued band-pass filters (of frequency bands [0, 0.5], [0.5, 1.5], ..., [6.5, 7.5] Hz). These band-pass filters preserve information from different temporal frequency bands, similarly to spatio-temporal features from [25]. Then, the amplitude of each filtered complex-valued signal is computed, and the resulting signal is down-sampled (with an anti-aliasing low-pass filter). The features extracted with the eight filters were either used in separate models (one per filter) or in a joint model with all features concatenated.

2.2 Joint model

To address the limitations of the winner-take-all model, we propose to use a joint model over all layers simultaneously. In this model, features from all layers are concatenated and a ridge regression is fit on all concatenated features. Fitting a joint model maximizes the model prediction accuracy by taking into account possible complementarities between layers. Then, to derive a measure of layer selectivity from the joint model, it is necessary to estimate the relative contribution of each layer to the model. To do so, we adapt the product measure [27, 28], a variance decomposition measure that quantifies relative importance at the level of individual features [29] (see Appendix). The adapted product measure decomposes the explained variance into contributions of each layer.

Continuous layer selectivity. To derive a layer selectivity from the joint model, we use a weighted average of layer indices, weighted by variance decomposition ratios. Starting from a variance decomposition $\rho \in \mathbb{R}^m$, where m is the number of layers, negative values were clipped to zero, and ρ was normalized to sum to one. Then, layer selectivity is computed as a weighted average $s = \sum_{i=1}^{m} i\rho_i$. This formula leads to a continuous layer selectivity, which increases both its robustness and its ability to describe gradual changes over the cortical surface.

One issue arises from the proposed approach. Because CNN activations are strongly correlated from one layer to the next, a joint model tends to use almost all layers. Therefore, the continuous layer selectivity gives an estimate that is biased toward middle values. To solve this issue, we refine our approach using banded ridge regression instead of ridge regression.

2.3 Banded ridge regression

When fitting joint models over multiple feature spaces simultaneously, using ridge regression assumes that all feature spaces require the same amount of regularization. However, the optimal amount of regularization depends on many characteristics of a feature space, such as its covariance, its number of features, or its predictive power. To solve this issue, ridge regression can be extended to banded ridge regression [22], which uses a separate regularization hyperparameter per feature space (see Appendix). Hyperparameters are learned with cross-validation.

By varying regularization strength on each feature space, banded ridge regression is able to select feature spaces that have good predictive power, and ignore those that have little predictive power or that are redundant with other feature spaces. This process leads to a feature-space selection which can improve the generalization performance on the test set, and follows the principle of parsimony (Occam's razor). In the case of continuous layer selectivity, using fewer layers also mitigates the bias toward middle values.

3 Results

This section presents a series of experiments comparing the proposed approach with the conventional approach. The experiments used functional magnetic resonance imaging (fMRI) brain recordings of a subject watching short movie clips [25]. Feature extraction was based on a pretrained CNN "Alexnet" [30] implemented in PyTorch [31]. See the Appendix for more details about model fitting procedures.

3.1 Comparing prediction accuracy

To compare the proposed approach with the conventional approach, we fit multiple regression models and estimated their prediction accuracy on a test set using the R^2 score. To summarize the performance of each model into a single score, the R^2 score was averaged over all voxels significantly predicted by at least one of the compared models ($p < 0.01$, permutation test, see Appendix).

First, we investigated the benefit of using band-pass filters during feature extraction. A separate ridge regression was fit on each combination of layer and filter, including concatenated layers and concatenated filters. The results (Figure 1 (left)) show that for each layer (except layer 4), each band-pass filter outperformed the low-pass filter. Furthermore, for each layer, concatenating features from all frequency bands outperformed the low-pass filter features.

Then, we investigated the benefit of using a joint model compared to a winner-take-all model. Three models were compared: a winner-take-all model, a joint model fit with ridge regression, and a joint

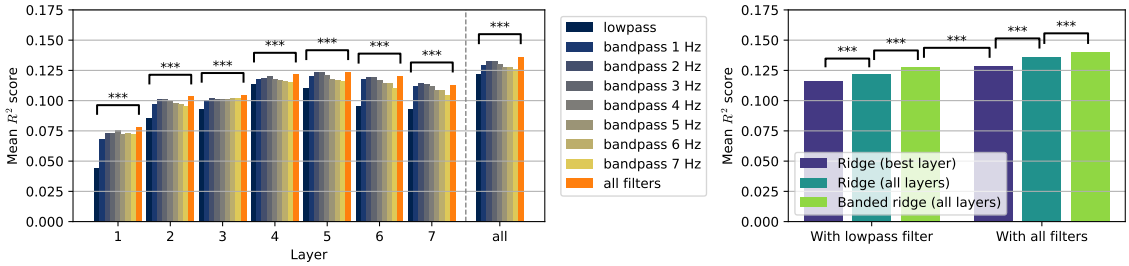

Figure 1: Model comparison in terms of prediction accuracy. Prediction accuracy was estimated with the R^2 score computed on a test set. The R^2 score was averaged over voxels predicted significantly by at least one of the compared models. **(Left)** To describe the benefit of band-pass filters, a ridge regression was fit on each combination of layer and frequency band. For every layer (except layer 4), the low-pass filter led to a lower score than every band-pass filter ($p < 10^{-4}$, paired t-test). For all layers, concatenated features from all filters outperformed the low-pass filter features ($p < 10^{-4}$, indicated by ***). **(Right)** To describe the benefit of the joint model, three models were compared: a winner-take-all model, a joint model fit with ridge regression, and a joint model fit with banded ridge regression. The three models were compared in two settings, either using a low-pass filter, or using a set of low-pass and band-pass filters. In both settings, the joint models were significantly better than the winner-take-all model ($p < 10^{-4}$), and fitting with banded ridge was significantly better than fitting with ridge ($p < 10^{-4}$). Between both settings, the set of filters led to significantly better predictions than with the low-pass filter ($p < 10^{-4}$).

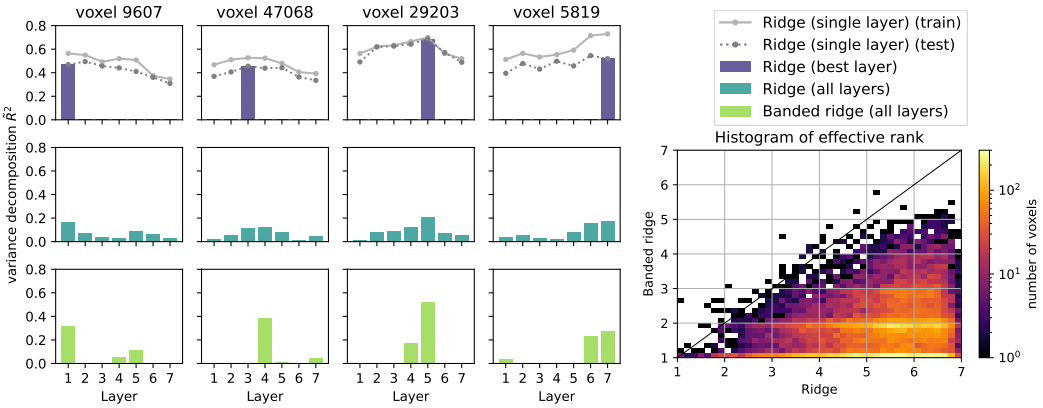

Figure 2: Model sparsity at the layer level. (Left) Variance decomposition of three different models over the seven layers, for four individual voxels. The winner-take-all model (*top*) uses only the best layer, ignoring potential complementarity between layers. The joint model fit with ridge (*middle*) uses almost all layers to make the predictions. The joint model fit with banded ridge (*bottom*) performs a feature-space selection leading to sparsity at the layer level, but allowing multiple layers to be used simultaneously. **(Right)** To quantify sparsity at the layer level, the effective rank is computed on the variance decomposition. The effective rank estimates the number of layers effectively used in each voxel. The histogram of effective rank over voxels shows that banded ridge regression (*vertical*) leads to a sparser model (smaller effective rank) than ridge regression (*horizontal*).

model fit with banded ridge regression. The results (Figure 1 (right)) show that both joint models outperform the winner-take-all model, and that fitting the joint model with banded ridge further improves the prediction accuracy gain. This result was observed with two different sets of features, either extracted with a low-pass filter or with a set of filters.

3.2 Comparing layer sparsity

In the previous subsection, banded ridge regression was shown to outperform ridge regression in terms of prediction accuracy. In this subsection, we show that banded ridge regression also uses fewer layers than ridge regression (*i.e.* it gives rise to a sparser model). Figure 2 (left) shows examples of variance decomposition over layers of different models. In each voxel shown, the joint model fit with banded ridge regression effectively used fewer layers than when fitting with ridge regression.

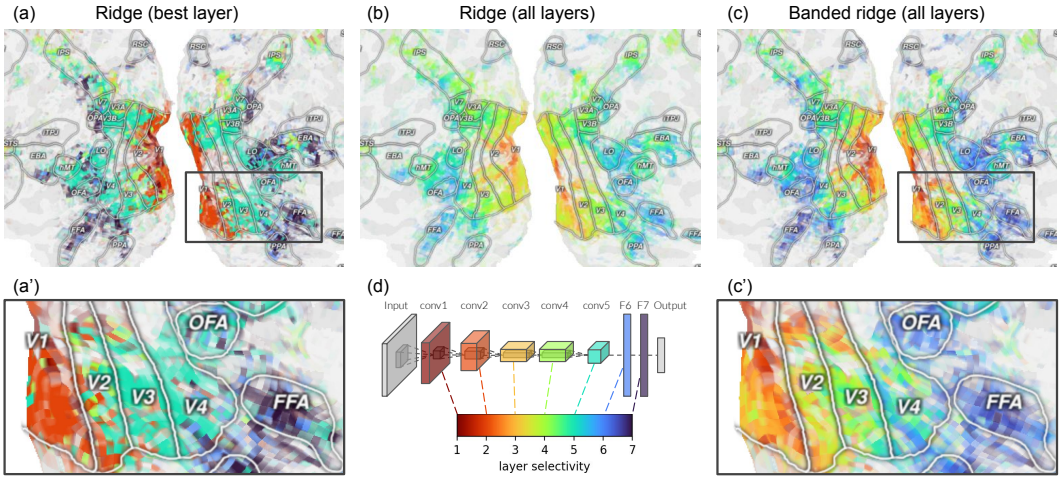

Figure 3: Layer selectivity of each voxel in the visual cortex, for three different models. The layer selectivity of each voxel is computed as a weighted average of the layer index, weighted by the variance decomposition. **(a,b,c)** Projection on a flattened cortical surface using Pycortex [33]. **(a)** The winner-take-all model gives a non-robust estimate of layer selectivity, because the argmax operator can flip from one layer to another due to small variations in prediction accuracy. **(b)** A joint model fit on all layers gives a continuous measure of layer selectivity, with a smooth gradient over the cortical surface without using any smoothing. However, fitting the joint model with ridge regression gives a biased estimate of layer selectivity toward middle values, because its variance decomposition tends to use all layers. **(c)** Fitting a joint model with banded ridge regression mitigates this bias thanks to its feature-space selection. **(a', c')** Zoomed views of (a, c). **(d)** Correspondence with the CNN architecture. The layer selectivity is derived from the layer indices of a pretrained Alexnet model.

To quantify layer sparsity over all significantly predicted voxels ($p < 0.01$), we computed the effective rank [32] based on variance decompositions over layers. Starting from a variance decomposition $\rho \in \mathbb{R}^m$, negative values were clipped to zero, and ρ was normalized to sum to one. Then, the effective rank was defined as $\tilde{m} = \exp(-\sum_{i=1}^{m} \rho_i \log(\rho_i))$. The effective rank is a continuous measure of the number of layers effectively contributing to the model. It is equal to k when the variation is equally split between k layers. Comparing the distribution of effective rank over voxels, Figure 2 (right) shows that banded ridge regression leads to a sparser joint model than ridge regression.

3.3 Comparing layer selectivity

In this subsection, we qualitatively compare the proposed approach and the conventional approach in terms of layer selectivity. In the conventional approach, each voxel is associated with the index of the best-predicting layer, leading to a discrete layer selectivity. In the proposed approach, a joint model is fit on all layers, and a continuous layer selectivity is computed based on the variance decomposition over layers. Figure 3 shows the layer selectivity projected on the cortical surface of the visual cortex. Compared to the discrete layer selectivity, the continuous layer selectivity leads to a smoother gradient over the cortical surface, without using any explicit smoothing or implicit smoothing prior. The reduced range of value of the continuous layer selectivity can be mitigated by using banded ridge regression instead of ridge regression.

Conclusion

Mapping CNNs to the visual cortex has drawn a lot of attention, both from the Machine Learning community looking for CNN interpretation, and from the Neuroscience community looking for brain mappings. As part of the encoding model framework, this methodology heavily relies on the comparison of different models in terms of their prediction accuracy, and usually leads to a winner-take-all approach where only the best model is considered. In this work, we propose to go beyond the winner-take-all approach, by fitting joint models and giving fine decompositions over feature spaces. Applied to layer selectivity, our approach both leads to higher prediction accuracy and to smoother cortical maps. This approach could be applied to most applications based on the encoding model framework, to improve both model performances and interpretations.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[2] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[3] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

[4] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[5] 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.

[6] 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.

[7] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio- temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

[8] 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, 2016.

[9] 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.

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

[11] Chi Zhang, Kai Qiao, Linyuan Wang, Li Tong, Guoen Hu, Ru-Yuan Zhang, and Bin Yan. A visual encoding model based on deep neural networks and transfer learning for brain activity measured by functional magnetic resonance imaging. *Journal of Neuroscience Methods*, 325:108318, 2019.

[12] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021.

[13] Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9):103013, 2021.

[14] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[15] Juliette Millet and Jean-Remi King. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032*, 2021.

[16] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fMRI. In *Advances in neural information processing systems*, pages 6628–6637, 2018.

[17] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language- processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938, 2019.

[18] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task- derived representations from brain activity. In *Advances in Neural Information Processing Systems*, pages 15475–15485, 2019.

[19] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.

[20] Iris IA Groen, Michelle R Greene, Christopher Baldassano, Li Fei-Fei, Diane M Beck, and Chris I Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, 7:e32962, 2018.

[21] Ying Yang, Michael J Tarr, Robert E Kass, and Elissa M Aminoff. Exploring spatiotemporal neural dynamics of the human visual cortex. *Human brain mapping*, 40(14):4213–4238, 2019.

[22] 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.

[23] Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29:477–505, 2006.

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

[25] 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.

[26] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[27] Paul J Hoffman. The paramorphic representation of clinical judgment. *Psychological bulletin*, 57(2):116, 1960.

[28] John W Pratt. Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international Tampere conference in statistics, 1987*, pages 245–260. Department of Mathematical Sciences, University of Tampere, 1987.

[29] Ulrike Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152, 2015.

[30] 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.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[32] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610. IEEE, 2007.

[33] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9:23, 2015.

A Appendix

A.1 Product measure

The product measure [27, 28] is a variance decomposition measure with many good properties [28, 36, 37, 29]. In particular, the product measure gives a proper decomposition of the variance explained in the training set $\sum_j \rho_j = R^2$. Yet, this property holds only for ordinary least squares and on the training set, because it uses the fact that the error vector $y - \hat{y}$ is orthogonal to the prediction vector $\hat{y}$. We thus adapt the product measure to decompose a definition of the R^2 score used for regularized regression method and/or on held-out data, $R^2 = 1 - ||y - \hat{y}||^2/||y||^2$, where y is zero-mean [39]. We then define the variance explained by layer j as

$$\rho_j = \frac{\sum_t \hat{y}_j(2y - \hat{y})}{\sum_t y^2}. \tag{1}$$

Here, $\hat{y}_j$ is the sub-prediction $\hat{y}_j = X_j b_j$, computed on layer X_j alone using the weights b_j of the joint model. Using the definition in (1), the variance explained R^2 by the joint model is the sum of the variance explained by all layers: $\sum_j \rho_j = R^2(\hat{y})$. This is an essential property to interpret the measure as a proper variance decomposition. Moreover, when the sub-predictions $\hat{y}_j$ are all orthogonal, the decomposition is identical to using the R^2 scores computed on each sub-prediction, $\rho_j = R^2(\hat{y}_j)$. Because in the case of ordinary least squares, the decomposition in (1) is equal to the product measure on the training set, this paper simply refers to it as the product measure.

A.2 Banded ridge regression

Definition. In banded ridge regression [22], the features are grouped into m feature spaces. A feature space i is formed by a matrix of features $X_i \in \mathbb{R}^{n \times p_i}$, with n samples and p_i features, and each feature space is associated with a different regularization hyperparameter $\lambda_i > 0$. To model brain activity $y \in \mathbb{R}^n$ on a particular voxel, banded ridge regression computes the weights $b_i^* \in \mathbb{R}_i^p$ (concatenated into $b^* \in \mathbb{R}^p$ with $p = \sum_i p_i$) defined as,

$$b^* = \underset{b}{\operatorname{argmin}} \, || \sum_i X_i b_i - y||_2^2 + \sum_i \lambda_i ||b_i||_2^2. \tag{2}$$

Similarly to ridge regression, the parameters $\hat{b}_i$ are learned on the training data set, while the hyperparameters λ_i are learned by cross-validation [22].

Related models Banded ridge regression is similar to the group lasso [41], in the sense that they both lead to a selection of groups of features. However, contrary to banded ridge regression, the group lasso does not optimize the group scalings. In fact, many fixed group scalings have been proposed in the past for the group lasso, using either the square root of the number of features [41], the trace [42] or the empirical rank of the feature kernels [43], or more exotic scalings [44, 45]. Instead, banded ridge regression learns optimal scalings from the data, by maximizing cross-validation performances.

Banded ridge regression can also be formulated as a special case of multiple-kernel learning. Multiple-kernel learning [42, 46] considers generalizations of kernel methods, where instead of a fixed kernel, a collection of kernels is combined to improve the flexibility of the model. Multiple-kernel learning methods traditionally learn the kernel combination jointly with the kernel model [47], yet learning the kernel combination via cross-validation has also been proposed [48, 49, 50]. Banded ridge regression is equivalent to multiple-kernel ridge when learning the kernel combination via cross-validation.

Banded ridge regression can also be linked with automatic relevance determination [51], which learns different hyperparameters to induce sparsity at the feature level. Applying this approach to Bayesian ridge regression leads to sparse Bayesian learning (also known as relevance vector machines) [52, 53], which is the direct Bayesian counterpart of banded ridge regression.

A.3 Model fitting

Dataset. The preprocessed dataset consisted of 73211 cortical voxels recorded every two seconds. Of the 3870 total time samples, 3600 samples were used in the train set (split into 12 runs of 300 samples) and 270 samples were used in the test set.

Hyperparameter search. For all models, a leave-one-run-out cross-validation scheme was used to select hyperparameters. For ridge regression, the regularization hyperparameter was selected with a grid-search over 30 log-spaced values ranging from 10^{-5} to 10^{20}. For banded ridge regression, the regularization hyperparameters were selected with a random-search; a thousand normalized hyperparameter candidates were sampled from a Dirichlet distribution, and were scaled by 30 log-spaced values ranging from 10^{-5} to 10^{20}. The best normalized candidate and the best associated scaling were selected on each voxel. Finally, models were refit on the entire training dataset with the selected best hyperparameters.

A.4 Significance testing

To test significance of the model prediction accuracy on each voxel, a permutation analysis was used. First, brain responses in the training set were permuted by blocks of 5 consecutive time samples. Then, a ridge regression model was fit on the permuted data, and prediction accuracy was computed on the test set. Repeating this procedure 1000 times led to a distribution of prediction accuracy, under the null hypothesis that there is no systematic relationship between model predictions and fMRI responses. Statistical significance was defined as any prediction accuracy that exceeded 99% of all of the permuted predictions ($p = 0.01$), and was calculated separately for each model. P-values were corrected for multiple comparisons using the false discovery rate [54]. To be able to compare the average prediction accuracy across models, the average was computed on all voxels significantly predicted by at least one of the compared models.

To test significance of the difference in prediction accuracy between two models, a paired t-test was computed over all voxels significantly predicted by at least on the two models.

Appendix references

[34] Paul J Hoffman. The paramorphic representation of clinical judgment. *Psychological bulletin*, 57(2):116, 1960.

[35] John W Pratt. Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international Tampere conference in statistics, 1987*, pages 245–260. Department of Mathematical Sciences, University of Tampere, 1987.

[36] D Roland Thomas, Edward Hughes, and Bruno D Zumbo. On variable importance in linear regression. *Social Indicators Research*, 45(1-3):253–275, 1998.

[37] Laura L Nathans, Frederick L Oswald, and Kim Nimon. Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research, and Evaluation*, 17(1):9, 2012.

[38] Ulrike Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152, 2015.

[39] Tarald O Kvålseth. Cautionary note about r 2. *The American Statistician*, 39(4):279–285, 1985.

[40] 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.

[41] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[42] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.

[43] Francis R Bach, Romain Thibaux, and Michael I Jordan. Computing regularization paths for learning multiple kernels. In *Advances in neural information processing systems*, pages 73–80, 2005.

[44] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[45] Qing Pan and Yunpeng Zhao. Integrative weighted group lasso and generalized local quadratic approximation. *Computational Statistics & Data Analysis*, 104:66–78, 2016.

[46] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *International conference on machine learning*, page 6, 2004.

[47] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011.

[48] Koji Tsuda, Shinsuke Uda, Taishin Kin, and Kiyoshi Asai. Minimizing the cross validation error to mix kernel matrices of heterogeneous biological data. *Neural Processing Letters*, 19(1):63–72, 2004.

[49] S Sathiya Keerthi, Vikas Sindhwani, and Olivier Chapelle. An efficient method for gradient-based adaptation of hyperparameters in SVM models. In *Advances in neural information processing systems*, pages 673–680, 2007.

[50] Teresa Klatzer and Thomas Pock. Continuous hyper-parameter learning for support vector machines. In *Computer Vision Winter Workshop (CVWW)*, pages 39–47, 2015.

[51] David JC MacKay. Bayesian methods for backpropagation networks. In *Models of neural networks III*, pages 211–254. Springer, 1994.

[52] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[53] David Wipf and Srikantan Nagarajan. A new view of automatic relevance determination. *Advances in neural information processing systems*, 20:1625–1632, 2007.

[54] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.