Model Zoology and Neural Taskonomy for Better Characterizing Mouse Visual Cortex

Colin Conwell[1], **Michael A. Buice**[2], **Andrei Barbu**[3], **George A. Alvarez**[1]
[1]Dept. of Psychology, Harvard University; [2]Allen Institute for Brain Science; [3]CSAIL & CBMM, MIT

Abstract

What is the representational structure of mouse visual cortex and how is it shaped? Mice obviously interact with the world and recognize objects but unlike in primates the activity of their visual cortex is not well described by modern object-recognizing deep neural networks. Using a two-photon calcium-imaging dataset of the activity in more than thirty thousand neurons in mouse visual cortex recorded in response to a stimulus set of natural scenes (the Allen Brain Observatory Visual Coding dataset), we make a preliminary attempt to clarify the relative contributions of model task and architecture in characterizing the representational structure of rodent vision. Our method consists of comparing the neural recordings from the mouse brain to the responses of over 50 different networks: the same architecture of network trained individually on 21 different computer visual tasks in the Taskonomy project, and 30 different architectures all trained on the same task (object recognition) in the PyTorch model zoo. We also try and isolate which tasks are performed where, creating what we call 'taskonomic signatures' of different neural sites as an exploratory metric of functional specificity across mouse visual cortex. Object recognition does appear to be important — the best performing object recognizer in the PyTorch zoo is one of the most predictive models overall (and far outperforms a randomly initialized counterpart). At the same time, non-semantic tasks such as 2D segmentation also seem significant. This work suggests an avenue to discover critical visual tasks which may heretofore have been overlooked by computer vision and neuroscience alike, and constitutes a first step towards establishing the overall task-related (rather than purely anatomical) structure of the visual brain. A better understanding of mouse visual cortex, given the availability of data unmatched in quality and resolution compared to what can be recorded from primates, may provide new insights into fine-grained recognition and scene understanding beyond those which define even our most capable machines. Novel combinations of task and architecture inspired by this neural data may in turn help to explain the conspicuously large amount of reliable neural variance left unexplained by even the most predictive models we catalogue here.

1 Introduction

To date, the most successful models of biological visual cortex are the object-recognizing deep neural networks the modeling community have applied to the prediction of primate visual cortex — networks that are powerful enough even to act as generators for synthesizing stimuli that drive neural activity beyond its typical range (Bashivan et al. (2019); see also Appendix). Their success in the modeling of rodent visual cortex, however, has been a bit more meted. It appears that the types of features best matched from the feature spaces of rodent visual cortex are those extracted from the deeper layers of modern deep nets — a matching that contrasts the traditional schematic wherein "simple" and "complex" cells are best modeled by early convolutional and pooling layers (Shi et al., 2019). Some have even suggested that randomly initialized networks provide about as predictive a set of features as task-optimized neural networks, while still outperforming hand-engineered features (Cadena et al., 2019). Here, we re-examine the state of neural network modeling in rodent visual cortex, a survey we summarize in three findings: 1) that some sort of training (versus random initialization alone) is vital to the predictive power of a given neural network's features; 2) that different areas of mouse visual cortex are perhaps best described by different tasks; and 3) that even the most predictive models we test leave unexplained a vast majority of the variance in measurements from rodent visual cortex.

2 METHODS

2.1 DATASET

As our animal reference, we use the Allen Brain Observatory Visual Coding dataset (de Vries et al., 2020), a two-photon calcium-imaging dataset consisting of approximately 65,000 neurons collected from across the visual cortex of 221 awake, adult mice. The neurons sampled include 6 areas of visual cortex and 4 cortical layers — though some combinations of area and layer are absent from the assays. The sampled combinations of cortical area and layer constitute a total of 21 distinct neural sites. For the purposes of this analysis, we limit our dataset to the neural activity measured in response to a set of 118 natural images. Each of these images is displayed exactly 50 times over the course of an assay. To ensure a sufficient amount of signal, we whittle down our selection of neurons to include only those neurons with split-half reliability of 0.75 and above (the split halves in this case being constituted of 25 of the 50 presentations for each image). This conservative threshold still leaves over 8200 neurons for analysis and supports the construction of target representational dissimilarity matrices (RDMs) with split-half reliabilities as high as 0.93.

2.2 MODEL ZOOLOGY

To explore the influence of model architecture on predictive performance, we use 30 model architectures from the PyTorch Model Zoo, a collection of empirically established architectures pretrained on the ImageNet image classification challenge. For each model, we extract features from one pretrained and one randomly initialized version, creating RDMs at each layer of each architecture with the same 118 natural images used in the Allen Brain Observatory assays. In what is perhaps a significant limitation, all the modern object detection models available from the PyTorch zoo are feed-forward architectures. We establish a hierarchy of models in terms of their average predictive power across all neural sites assayed, though we note the overall hierarchy occasionally differs by neural site.

2.3 NEURAL TASKONOMY

In addition to the 30 model architectures from the PyTorch model zoo, we extract features from a single architecture trained on 21 different computer vision tasks in a project called Taskonomy (Zamir et al., 2018). Key to the engineering of Taskonomy is the use of an encoder-decoder architecture in which only the construction of the decoder varies across task. The encoder terminates in a latent space of 1024 dimensions, described by the authors as containing the most abstract, titrated representations specific to each task before those representations are transformed again for readout. Following recent, similar analyses on human visual cortex (Wang et al., 2019), we initially focus on RDMs constructed from this latent space before expanding our analysis to include representations extracted from all layers of the encoder. In the interest of clarity, we cluster the 21 tasks according to their 'taskonomic' category — a total of 5 clusters (2D, 3D, semantic, geometric or miscellaneous). These purely data-driven clusters are derived from how effectively a set of features learned for one task transfer to (or boost the performance in) another task.

2.4 COMPARISON TO NEURAL DATA

To compare the representational spaces extracted from the assembled deep neural network models with those of mouse visual cortex, we use a combination of RDMs and a cross-validated nonnegative least squares (NNLS) regression (Jozwik et al., 2016). For any given comparison, the regressors are *all* the RDMs extracted from a given model (one per layer); the regressand is the RDM computed from the neural activity of a given cortical area and layer (a neural site). For the purposes of cross-validation, we subdivide the RDMs with a k-fold, fitting the regression on $k - 1$ folds, then predicting the held-out fold, eventually constructing an RDM entirely of heldout predictions. We then correlate this predicted RDM with the RDM from the brain to obtain our similarity score.

2.5 TASKONOMIC SIGNATURES

To better resolve the relationship of task category to neural site — with the hypothesis that different neural sites may be best predicted by different tasks — we perform a series of ANOVAs. Noting that Bonferroni corrections in the post-hoc tests often leave us underpowered to find significant differences across the various combinations of neural site and task category, we supplement our comparisons with an analysis of the weights from a modified NNLS regression, simultaneously passing multiple models at a time as regressors and analyzing across 1000 random k-fold splits whether the regression assigns higher or lower weights to these different models when predicting different neural sites.

3 Results

3.1 Model Zoology

In contrast to previous findings suggesting that randomly initialized models provide as predictive a feature space as pretrained models, we find that 28 out of 30 ImageNet-pretrained architectures outperform their randomly initialized counterparts (15 significantly so; 6 with Bonferroni correction for multiple comparisons). The best performing architecture overall (significantly different from its randomly initialized counterpart) was an ImageNet-pretrained MobileNetv2 architecture with $R^2 = 0.1164$). The next best architectures were two ImageNet-pretrained SqueezeNet architectures with mean $R^2 = 0.1162$ and $R^2 = 0.109$, respectively. The worst performing ImageNet-pretrained architecture was a batch-normed VGG19 architecture with mean $R^2 = 0.032$. The worst performing architecture overall was exactly the inverse of the best performing model: randomly initialized MobileNetv2 resulting in a mean $R^2 = 0.011$. For the full hierarchy, see Figure 2 in Appendix A.1.

In addition to establishing the hierarchy of architectures, we tested two hypotheses regarding the composition of the hierarchy (limiting ourselves for now to the pretrained models): the first, inspired by the superlative performance of the MobileNet and SqueezeNet architectures (all designed to maximize the tradeoff between task accuracy and total number of trainable parameters) was whether there was a relationship between predictive performance and parameter count. A Pearson correlation suggests a slight, marginally significant negative relationship ($r = -0.346$, $p = 0.061$) — such that as total parameter count increases, average predictive performance (slightly) decreases. The second hypothesis we tested was whether there was a relationship between predictive performance and model depth (the total number of layers in a given architecture). Here, the Pearson correlation suggests a slight, but ultimately insignificant positive relationship ($r = 0.05$, $p = 0.783$).

3.2 Neural Taskonomy

The most predictive single task from the latent space of the Taskonomy encoder was 2D segmentation with mean $R^2 = 0.046$; the least predictive task was room layout (orientation and aspect ratio of the cubic room layout) with mean $R^2 = 0.000065$. The most predictive single task across all layers of the Taskonomy encoder was again 2D segmentation with mean $R^2 = 0.092$. The least predictive single task across all layers was neural inpainting with mean $R^2 = .022$. For the full hierarchy, see Figure 3 in Appendix A.2.

3.3 Taskonomic Signatures

In a two-way ANOVA examining the effects of task category (a data-driven cluster of tasks) and cortical area on representational similarity (limiting our analysis first to the latent space of the Taskonomy encoder), we find a significant main effect of task category ($p < 10^{-11}$, $\eta^2 = 0.13$) and a marginally significant interaction ($p = 0.056$, $\eta^2 = 0.07$). Expanding our analysis to all layers of the Taskonomy encoder, we partially recapitulate the same pattern of results, finding again a significant main effect of task category ($p < 10^{-6}$, $\eta^2 = 0.054$), but without a significant interaction of task category and neural site. Figure 1 provides an example of one of the taskonomic signatures that may be driving the interaction effect in the latent space of the Taskonomy encoder. (For the full set of signatures from which this example was extracted, see Figure 4 in Appendix A.3.)

As caveat to any conclusion drawn from such an example, we note that many of the taskonomic signatures derived in the latent space — including the example in Figure 1 — are absent when taking into account all layers, a similar trend to that we saw in terms of overall predictive performance. Just as the difference between the best and worst performing tasks is reduced from a roughly 7500% difference ($0.046 - 0.00061$) to a roughly 350% difference ($0.92 - 0.022$), so too are many of the differences that drive the taskonomic signatures in the latent space rendered trivial across all layers.

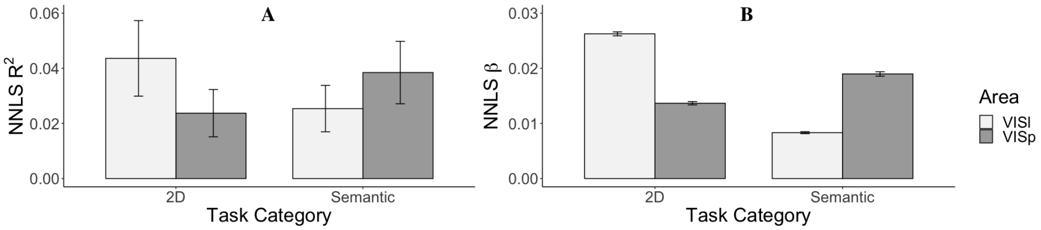

Figure 1: A representative example of a ‘taskonomic’ signature (a task-driven dissociation) between two cortical areas (VISp and VISl) and two clusters of task (2D and Semantic tasks). 2D tasks include 2D edge-detection and 2D segmentation; Semantic tasks include object classification and semantic segmentation. **A** is the dissociation in terms of variance explained. **B** is the dissociation in terms of relative weights in a combined nonlinear least squares regression analysis repeated across 1000 random 6-fold splits, where all task RDMs are used simultaneously as predictors and weights are assigned according to how well each feature space contributes to the overall prediction of neural data. The similarity of the pattern verifies the interaction in the absence of a statistically significant difference in terms of R^2 alone, and holds for 93.9% of random splits. Error bars in **A** are 95% confidence intervals of the mean R^2 across the four cortical layers. Error bars in **B** are permuted 95% confidence intervals from the cross-validation procedure.

3.4 Overall Performance

The overall best performance of any single model in any single neural site was Taskonomy’s 2DSegmentation in cortical layer 4 of area VISal, with $R^2 = 0.281$. This constitutes approximately a third of the total explainable variance in this site (its split-half reliability), with mean $R^2 = .789$.

4 Discussion

The results we have presented here are perhaps best interpreted as a preliminary sampler of the possibility space for the modeling of rodent visual cortex with off-the-shelf neural network models, or perhaps as a menu for new combinations of architecture and task that might thusfar have been neglected. Most promising in this domain seems to be the use of smaller, more computationally efficient networks combined with training on early and intermediate feature detection tasks for the modeling of brains in which vision is less central to behavior, and organizing principles like hierarchy are absent. The primary visual systems of mice may in the end be more akin to the peripheral visual systems of monkeys; mice lack a fovea, have a retina dominated by rods for vision under low light, and spatial acuity bordering on 20/2000 (Huberman & Niell, 2011). It is possible that mice rely on vision as a sort of broad bandpass filter for lower-frequency, dynamic stimuli that the mouse can then flee, fight or further investigate with its whiskers — perhaps its most sophisticated sensory organ. Indeed, the hierarchies seen in primate visual cortex may be better recapitulated in the rodent trigeminal system, where information from whisking is processed to remarkable depth (Zhuang et al., 2017). Combining task-optimized neural networks with the unparalleled access, resolution, and control afforded by rodent neuroimaging could put object recognition in context relative to the broader scene understanding of which it is a subcomponent or to other ethologically relevant tasks that drive the organization of the perceptual brain. An accounting of the full range of tasks carried out in scene understanding, the temporal and causal relationships between tasks, and how information is integrated from other senses could provide an invaluable roadmap for computer vision. (See Merel et al. (2019) for a promising example of this in the domain of embodied control).

In addition to a better understanding of vision as a whole, developing models which account for the relatively massive amount of unexplained variance our models leave in mouse visual cortex could help discover new organizing principles for all perceptual tasks. Clearly convolution and depth alone are not enough, at least not when applied to many tasks considered central to vision. Further innovations (perhaps the same that could solve fundamental computer vision problems like the learning of invariance to 3D rotations) will be necessary to more fully model the rich diversity and fiendish complexity of biological brains at scale – even the very smallest ones.

REFERENCES

P. Bashivan, K. Kar, and J. J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

N. Y. Bilenko and J. L. Gallant. Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics*, 10:49, 2016.

S. A. Cadena, F. H. Sinz, T. Muhammad, E. Froudarakis, E. Cobos, E. Y. Walker, J. Reimer, M. Bethge, A. Tolias, and A. S. Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46): 10577–10597, 2005.

S. E. de Vries, J. A. Lecoq, M. A. Buice, P. A. Groblewski, G. K. Ocker, M. Oliver, D. Feng, N. Cain, P. Ledochowitsch, D. Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020.

U. Güçlü and M. A. van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017.

A. D. Huberman and C. M. Niell. What can mice tell us about how vision works? *Trends in neurosciences*, 34(9):464–473, 2011.

K. M. Jozwik, N. Kriegeskorte, and M. Mur. Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, 83:201–226, 2016.

S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11), 2014.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019.

N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

J. Merel, D. Aldarondo, J. Marshall, Y. Tassa, G. Wayne, and B. Ölveczky. Deep neuroethology of a virtual rodent. *arXiv preprint arXiv:1911.09451*, 2019.

A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

J. Shi, E. Shea-Brown, and M. Buice. Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. In *Advances in Neural Information Processing Systems*, pp. 5765–5775, 2019.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

A. Wang, M. Tarr, and L. Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. In *Advances in Neural Information Processing Systems*, pp. 15475–15485, 2019.

D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. 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.

A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

C. Zhuang, J. Kubilius, M. J. Hartmann, and D. L. Yamins. Toward goal-driven neural network models for the rodent whisker-trigeminal system. In *Advances in Neural Information Processing Systems*, pp. 2555–2565, 2017.

A Appendix

A.1 Hierarchy of Neural Architectures

See Figure 2.

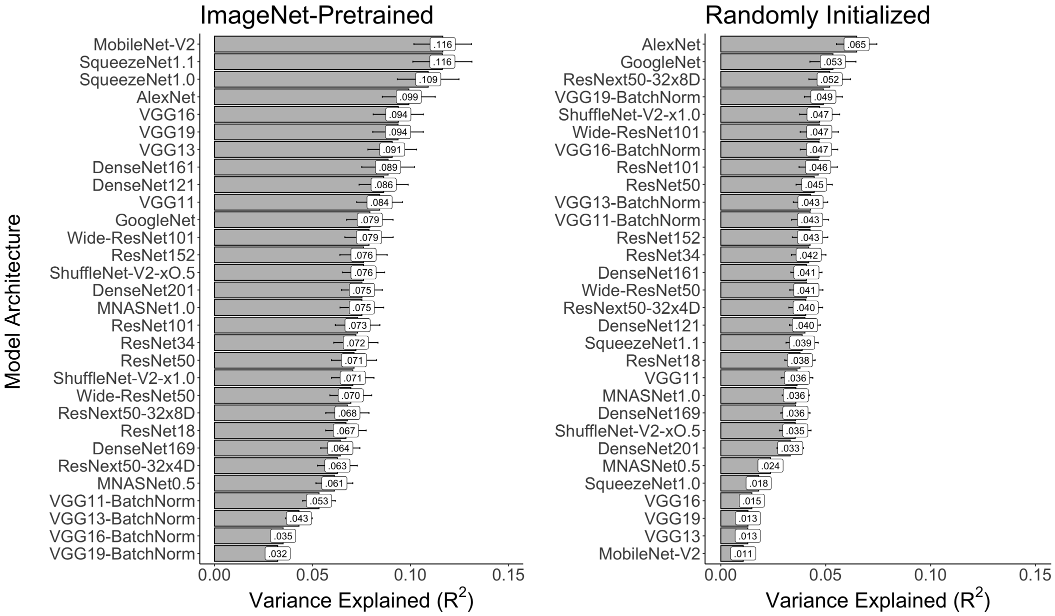

Figure 2: The full hierarchy of performance in terms of architecture. Error bars are the 95% confidence intervals when taking the mean predictive performance across cortical area and layer.

A.2 Hierarchy of Taskonomy Features

See Figure 3.

A.3 Taskonomic Signatures by Cortical Area

See Figure 4.

A.4 Comparison to Other Methods for Benchmarking Models of Brain Data

A variety of methods exist for comparing the responses of deep neural networks to neural responses recorded from brain tissue. The predominant – but by no means the only – methods might roughly be divided into two categories: regression and representational similarity analysis. Some of the first brain-to-network comparisons availed themselves of both; Yamins et al. (2014) citing Carandini et al. (2005) and Kriegeskorte et al. (2008) used linear regression for mapping responses in individual neural sites and representational similarity analysis for populations. Other seminal work comparing deep nets to primate visual cortex pioneered distinctive variants of these approaches. Güçlü & van Gerven (2017) employed regression in the form of encoding models to assess the hierarchical correspondence between earlier and later layers of processing across brain and machine. Khaligh-Razavi & Kriegeskorte (2014) built representational dissimilarity matrices by “remixing” and “reweighting”

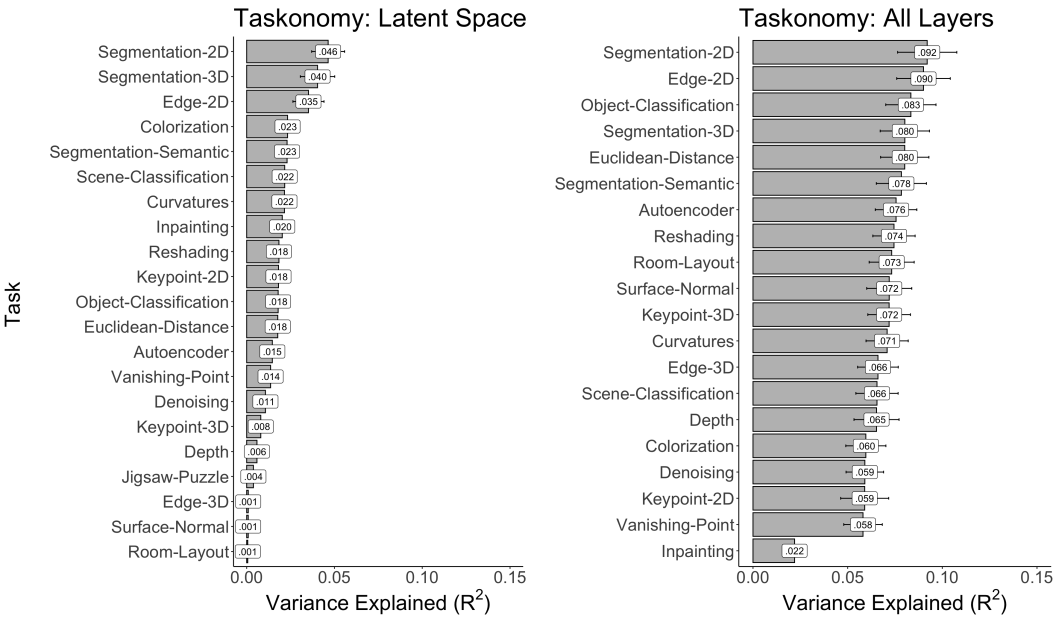

Figure 3: The full hierarchy of performance in terms of task. On the left are the predictions from only the latent space of the taskonomy encoder. On the right are the predictions leverage all layers in the taskonomy encoder. Error bars are the 95% confidence intervals when taking the mean predictive performance across cortical area and layer.

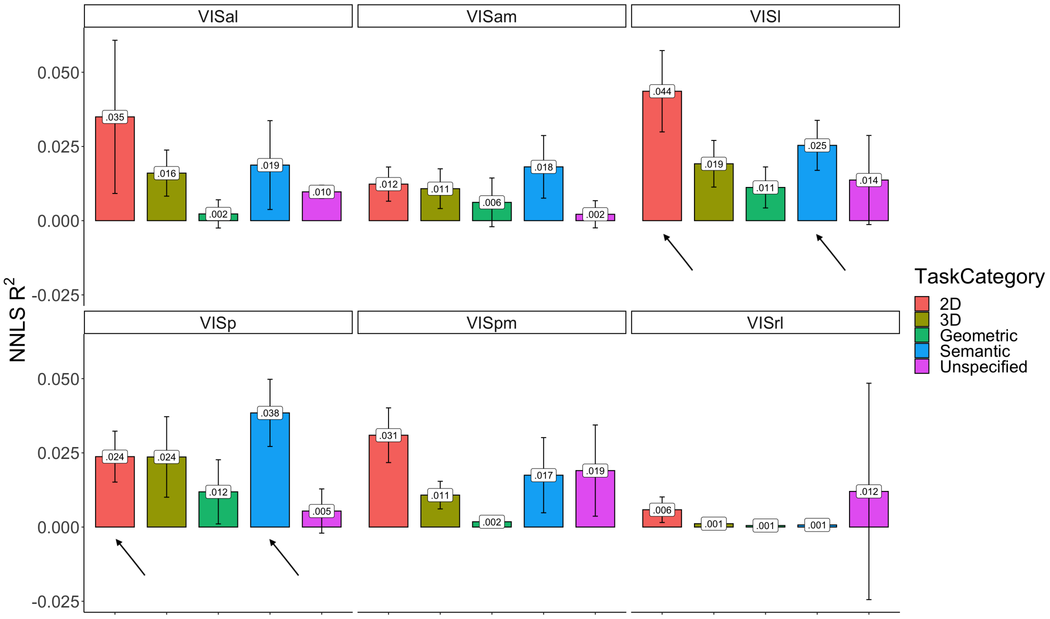

Figure 4: The full range of taskonomic signatures in the comparison of taskonomy's latent space with cortical area. The black arrows demarcate the data used in the example of the dissociation between areas VISl and VISp in the main body of the analysis (see Figure 1). Error bars are the 95% confidence intervals when taking the mean predictive performance across each cortical layer.

model features according to their performance in a support vector machine classifier trained on major categorical divisions in the stimulus set. A possible third strain of methods that doesn't fit so neatly into the binary of regression versus representational similarity are canonical correlation and alignment methods. These techniques leverage what is often assumed to be an underlying latent space of similarity shared across divergent high-dimensional datasets to assess (via projection) the shared variance between them. Canonical correlation and alignment methods are popular in both the machine learning (Morcos et al., 2018; Kornblith et al., 2019) and neuroimaging communities (Bilenko & Gallant, 2016), but have so far been applied mostly to comparison within, rather than across, domains and neural substrates. In the context specifically of comparisons to rodent neurophysiology, Cadena et al. (2019) use a modified regression analysis, predicting spike rate with a core feature model (VGG16) in tandem with a "shifter" network and "modular" network that correct for extraneous influences on recorded brain activity (including eye movements and running speed). The relative advantages of these various approaches as they pertain to characterizing the representational structure of biological brains is largely uncertain, with a comprehensive comparison of techniques on the same dataset seemingly absent from the literature.

The current standard for high-throughput benchmarking of neural data on neural models is perhaps that of Schrimpf et al. (2018) in BrainScore, a method that consists of a partial least squares (PLS) regression fit individually to each neural site (in their case, a cluster of neurons around a given electrode in a microarray), wherein the regressand is the responses from that site and the regressors are the principal components of a target model's feature space. The end product of this process is a reliability-corrected R^2 quantifying how much of the "explainable" variance in a given neural site is captured by a given model. While thorough in its granularity and extensive in its coverage, this combination of principal components analysis and partial least squares regression tends to be a computationally expensive process – often prohibitively so in the absence of cloud and cluster computing. Our method in this paper – a compromise that combines both penalized regression and representational similarity – trades some granularity of prediction for computational traction. This tradeoff was especially necessary in the context of working with optical physiology (calcium imaging) data, which provides a quantity of neural sites (individual neurons) at least an order of magnitude larger than the quantity provided by the electrophysiology that predominates in primate neuroimaging. In future work, we plan to more directly mirror the methods of Schrimpf et al. (2018) in BrainScore.

A.5 On Divergent Results with VGG16

Recent work has suggested that randomly initialized VGG16 provides as predictive a set of features as VGG16 pretrained on ImageNet (Cadena et al., 2019). In the hierarchy we produce, on the other hand, Imagenet-pretrained VGG16 strongly and significantly outperforms its randomly-initialized counterpart (with mean $R^2 = .079$ and 009, respectively). While the differences in methodology and dataset make a direct comparison of these results infeasible, one nuance in our data suggests a possible point of reconciliation. In addition to the standard VGG16 described in Simonyan & Zisserman (2014) (the same reference supplied by Cadena et al. (2019)), our catalogue includes a VGG16 with batch normalization. While batch normalization seemingly decreases the predictive accuracy of the network overall, it also eliminates the difference in predictive power between the Imagenet-pretrained and randomly initialized versions (with mean $R^2 = .027$ and .033, respectively). While it does seems that Cadena et al. (2019) used the standard VGG16 based on their reference, if in fact they used the version with batch normalization, this may explain the discrepancy between our results and theirs, with the relative predictive parity across pretrained and randomly initialized models only true in the case of batch-normalized VGG16. It is unclear at this time whether the effect of batch normalization is simply some artifact of the modeling process or is conceptually relevant to the neuroscience, but it is an effect worth considering, and another example of how multimodel comparisons might highlight idiosyncrasies of direct relevance to prediction.