Directly interfacing brain and deep networks exposes non-hierarchical visual processing

Nicholas J. Sexton[*][†] Bradley C. Love [†][‡]

June 2021

Abstract

One reason the mammalian visual system is viewed as hierarchical, such that successive stages of processing contain ever higher-level information, is because of functional correspondences with deep convolutional neural networks (DCNNs). However, these correspondences between brain and model activity involve shared, not task-relevant, variance. We propose a stricter test of correspondence: If a DCNN layer corresponds to a brain region, then replacing model activity with brain activity should successfully drive the DCNN's object recognition decision. Using this approach on three datasets, we found all regions along the ventral visual stream best corresponded with later model layers, indicating all stages of processing contained higher-level information about object category. Time course analyses suggest long-range recurrent connections transmit object class information from late to early visual areas.

1 Introduction

Despite some shortcomings(*1, 2*), deep convolutional neural networks (DCNNs) have emerged as the best candidate models for the mammalian visual system. These models take photographic stimuli as input and, after traversing multiple layers consisting of millions of connection weights, output a class or category label. Weights are trained on large datasets consisting of natural images and corresponding labels.

The deep learning revolution in neuroscience began when layers of DCNNs were related to regions along the ventral visual stream in an early-to-early and late-to-late pattern of correspondence between brain regions and model layers (*3–5*) (fig 1A). This correspondence supported the view that the ventral stream is a hierarchy in which ever more complex features and higher-level information

[*] Correspondence to: n.sexton@ucl.ac.uk

[†] Department of Experimental Psychology, University College London, London, United Kingdom.

[‡] The Alan Turing Institute, London, United Kingdom

are encoded as one moves from early visual areas like V1 or V4 to inferotemporal (IT) cortex (*6–8*).

However, these correspondences between brain and model activity were based on total shared variance as opposed to task-relevant variance (fig 1B). Much of cortex-wide neural variance does not relate to the task of interest(*9*) and may co-vary with but not drive behaviour. Correspondences established by correlation alone do not necessitate that model layers and brain regions play the same functional role in the overall computation.

We propose a stronger test for evaluating how brain-like a model is. If, as is frequently claimed(*3–5*), a specific layer in a DCNN corresponds to a brain region, then it should be possible to substitute the activations on that layer with the corresponding brain activity and drive the DCNN to an appropriate output (cf. (*10–12*), fig 1C). For example, if we take V4 activity from a monkey viewing an image of a car and interface that brain activity with an intermediate DCNN layer hypothesised to correspond to V4, then the DCNN should respond "car" absent any image input. How well the DCNN performs when directly interfaced (through a simple linear mapping, see SI 6.5) with the brain provides a strong test of how well the interfaced brain region corresponds to that layer of the DCNN.

2 Driving model response with brain activity

We interfaced a pretrained DCNN(*13*) with data from two human brain imaging studies(*14, 15*) and a Macaque monkey study(*16*). All three studies involved viewing complex images. For a chosen model layer and brain region, we calculated a linear mapping from brain to model activity by presenting the same images to the model for which we had neural recordings (fig 1C). This simple linear mapping is a translation between brain and model activity. We evaluated the quality of this translation by considering held-out images and brain data that were not used in calculating the linear mapping (see SI 6.4).

Strikingly, for the two fMRI studies (figs 2A, 2B), the DCNN was most accurate at classifying novel images when brain activity across regions (both early and late along the ventral stream) was interfaced with later model layers. In contrast to previous analyses that focused on total variance, we did not find the early-to-early and late-to-late pattern of correspondence. Even primary visual cortex, V1, best drove the DCNN when interfaced with an advanced layer. For comparison, classifiers commonly used to decode information from fMRI data through multivariate pattern analaysis (MVPA) were at chance levels (fig 6), which highlights the useful constraints captured in the pretrained DCNN. After training on a million naturalistic images, the DCNN developed representations that paralleled those of the ventral stream, which made decoding object class possible by way of a linear mapping from brain activity to an advanced DCNN layer. The interpretation is that all brain regions contain advanced object recognition information, which conflicts with strict hierarchical views of the ventral visual stream.

To rule out any alternative explanation based on the indirect nature of fMRI recordings, we considered a third study consisting of direct multi-unit recording of spiking neurons implanted in the ventral visual stream of Macaque monkeys(*16*). These monkeys were shown images that did not readily align with the pretrained DCNN's class labels, so we evaluated neural translation performance by comparing the outputs of the DCNN when its input was a study image vs. when a DCNN layer was driven by brain data elicited by the same image. For the distance measure, KL divergence, lower values imply a better translation between brain and model activity. As in the fMRI studies, both relatively early regions (i.e., V4) and late regions (i.e., IT) best translated to later DCNN layers (fig 2C).

Across three diverse studies, we found a remarkably consistent pattern that strongly diverged from previous analyses — both early and late regions along the ventral visual stream best corresponded (i.e., translated) to late model layers. It is not that previous analyses were poorly conducted (see SI fig 5 for a successful reanalysis of data(*16*) finding the early-to-early and late-to-late canonical pattern). Rather, our novel analyses focused on task-relevant analysis, i.e., variance that can drive behaviour, provided a different view of the system than standard analyses focused on shared variance. Integrating these two views suggests a non-hierarchical account of object recognition marked by long-range recurrence transmitting higher-level information to the earliest visual areas.

3 Long-range recurrence as opposed to strict hierarchy

One way to reconcile the existing literature based on shared variance with our analyses based on task-relevant variance is to propose that long-range connections from IT transmit higher-level information to early visual areas. Even if most variance in lower-level visual areas is attributable to stimulus-driven, bottom-up activity, the majority of task-relevant information could be attributable to signals originating from IT (fig 3).

This view predicts specific patterns of Granger causality between early and late areas along the ventral visual stream. Do past values of one time series predict future values of the other? In terms of total spiking activity, lower-level areas should first cause activity in higher-level areas during the initial feed-forward pass in which stimulus-driven activity propagates along the ventral visual stream. Later in processing, the causality should become reciprocal as top-down connections from IT affect firing rates in lower-level areas, such as V4 (fig 3, bottom row). In contrast, Granger causality for task-relevant information should first be established from IT to V4 (i.e., the top-down signal) and only later in processing should recurrent activity lead to causality from V4 to IT (fig 3, top row). In this fashion, all areas are effectively "late" after long-range recurrent connections transmit information from IT to early visual areas along the ventral stream though most variance for these areas would be dominated by lower-level (bottom-up) stimulus information.

We tested these predictions using the monkey multi-unit spiking data(*16*) that has the temporal resolution to support the analyses. Images were presented one after the other, each visible for 100ms, with a 100ms period between stimuli. Figure 4A shows the mean firing rates (10 ms bins) with activity in V4 increasing shortly before IT, consistent with stimulus-related activity first occurring in V4. Figure 4B revisits our previous analyses (fig 2C) but with spike counts binned into 10ms intervals rather than aggregated over the entire trial. Even with only 10ms of recordings, neural translation from V4 and IT to an advanced DCNN network layer minimises KL divergence between model outputs arising from image input vs. when driven by brain activity.

Turning to the key Granger causality analyses, we evaluated whether early ventral stream regions become more like late-ventral stream regions over time due to recurrence (fig. 3). As processing unfolded, we found mutual causality between lower-level (V4) and higher-level (IT) areas for analyses conducted over spike counts (fig 4C) and for analyses on the KL divergence times series that assessed the ability of brain regions to drive DCNN response (fig 4D).

Critically, the specific predictions of the long-range recurrence hypothesis were supported with V4 first driving IT ($V4 \rightarrow IT$) for the analysis of spike counts but IT first driving V4 ($V4 \leftarrow IT$) for the task-relevant information analysis using the KL divergence time series (see SI for details). These results are consistent with stimulus-driven bottom-up activity proceeding from V4 to IT on an initial feed forward pass through the ventral stream with actionable information about object recognition first arising in IT. Then, recurrent connections from IT to V4 make task-relevant information available to V4. As this loop is completed and cycles, both areas mutually influence one another with the impact of bottom-up stimulus information maintained throughout the process.

4 Discussion

Computational models can help infer the function of brain regions by linking model and brain activity. Mulitlayer models, such as DCNNs, are particularly promising in this regard because their layers can be systematically mapped to brain regions. Indeed, the deep learning revolution in neuroscience began with analyses suggesting an early-to-early, late-to-late pattern of correspondence between DCNN layers and brain regions along the ventral visual stream during object recognition tasks(*3–5*).

However, as we have argued, correspondences based on total shared variance should be treated with caution. To complement these approaches, we presented a test focused on task-relevant variance that directly interfaced neural recordings with a DCNN model. If a brain region corresponds functionally to a model layer, then brain activity substituted for model activity at that layer should drive the model to the same output as when an image stimulus is presented. Of course, models and brains speak different languages, so a translation between brain and model activity must first be learned, which in our case was accomplished by a linear transformation. Once the translation function is learned, novel brain data and images can be used to evaluate possible brain-model correspondences.

Our approach, which focuses on task relevant variance within the overall computation, as opposed to local shared variance (fig 1), uncovered a pattern of correspondences that dramatically differed from the existing literature. We found that all brain regions, from the earliest to the latest of visual areas along the ventral stream, best corresponded to later model layers. These results indicate that neural recordings in all regions contain higher-level information about object category even when most variance in a region is attributable to lower-level stimulus properties (fig 3).

To resolve this discrepancy between our analyses focused on task-relevant variance and those based on shared variance, we evaluated the hypothesis that long-range recurrence between higher-level brain regions, such as IT, influenced activity in lower-level areas like V4. Analysing both firing rates of cells and information-level analyses using our brain-model interface approach, we found evidence that recurrent activity renders all areas functionally “late” as processing unfolds, even when total variance in some early visual regions is largely driven by bottom-up stimulus information. In this way, we integrate previous findings with our own and highlight how our method can be used to test hypotheses about information flow in the brain.

Our approach, which considers task-relevant variance, may help resolve conflicting interpretations on the function of brain regions. For example, the fusiform face area (FFA) responds selectively for faces, but its wider functional role in object recognition has been the subject of extensive debate(*17*). Here, we show that interfacing FFA into late model layers drives object recognition comparably to the lateral occipital complex (fig 2B) on non-face natural images. We suspect that the function of a region will only be fully understood by considering task-relevant variance across several tasks in light of activity in connected brain regions. The tight interface we champion between computational models and brain activity should prove useful in evaluating theoretical accounts of how the brain solves tasks over time.

Computational models that perform the tasks end-to-end, from stimulus to behaviour, should be particularly useful. In essence, translating between brain regions to layers of such models can make clear what role a brain region plays within the overall computation. In the case of object recognition, our results suggested that recurrent models may be best positioned to explain how the nature of information within brain regions changes as the computation unfolds.

This conclusion is in line with a growing body of modelling work in neuroscience that affirms the value of recurrent computation(*18–20*). Unlike the aforementioned work, we suggest that long-distance recurrent connections that link disparate layers should be considered (cf. (*21*)). We suspect such models will be necessary to capture time course data and the duality found in some brain regions, namely how most variance in a brain region can be attributable to lower-level stimulus properties while co-mingled with important higher-level, task-relevant signals.

As deep learning accounts in neuroscience are extended to other domains, such as audition (*22*), and language processing (*23*), the lessons learned here may apply. Our brain-model interface approach can help evaluate whether the brain processes signals across domains in an analogous fashion. By minding the distinction between shared and task-relevant variance, the role brain regions play within the overall computation may more readily come into focus.

Our approach may also have practical application in brain machine interfaces (BMI). Recent BMI developments have emphasised the readout of motor commands, neural processes taking place close to the periphery. In contrast, by leveraging the constraints provided by a pre-trained DCNN, we were able to gain traction on the 'stuff of thought', categorical and conceptual information in IT. Because we learned a general translation from brain to model, our approach applied to BMI would allow distant generalisation. For example, we were able to extrapolate to novel categories (see SI). For example, a translation from brain to model that never trained on horses, but trained on other categories, can perform zero-shot generalisation when given brain activity elicited by an image of a horse. The interface has the potential to produce a domain-general mapping rather than one dependent on specific training data. In the future, BMI approaches that address general thought without exhaustive training on all key elements and their combinations may be feasible.

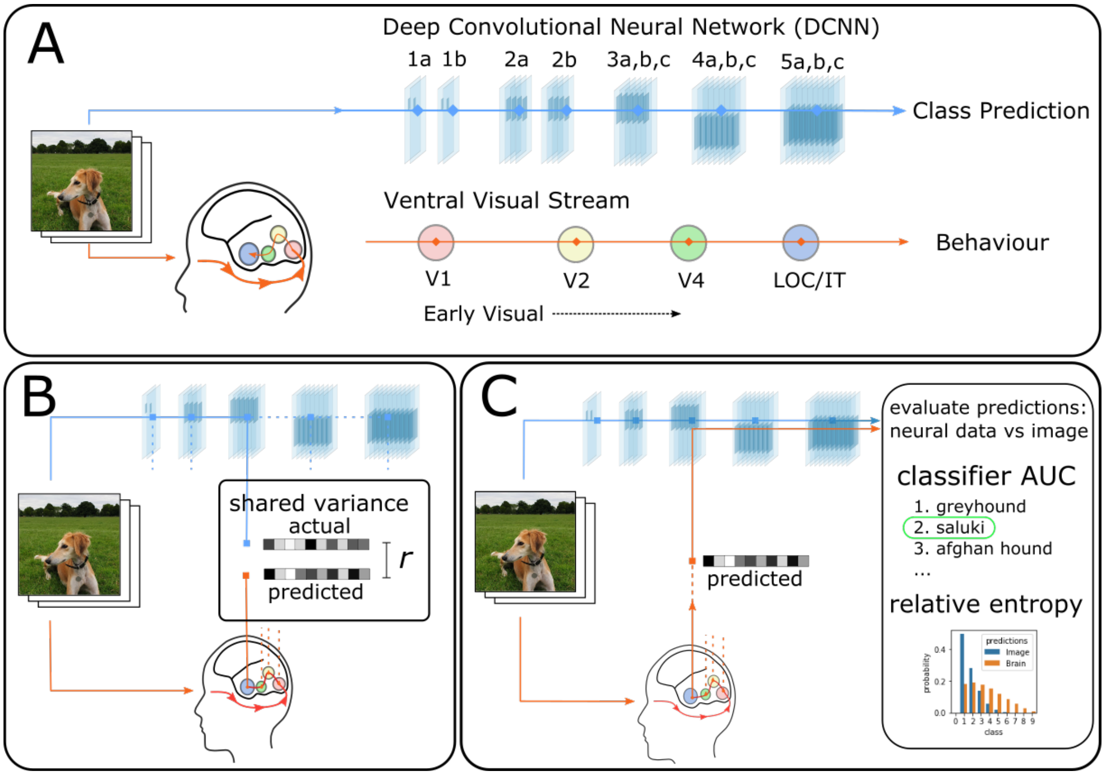

Figure 1: Deep Convolutional Neural Networks (DCNNs) trained on large naturalistic image datasets(*24*) have emerged as leading models of the mammalian ventral visual stream. (**A**) Typically, processing in DCNNs is hierarchical starting with the stimulus and proceeding across successive layers as higher-level information is extracted, culminating in predicting the class label(*13*). Numerous analyses(*3–5*) based on shared variance suggest the brain follows related principles with an early-to-early, late-to-late pattern of correspondence between the ventral visual stream and DCNN layers. (**B**) These shared-variance correspondences are evaluated locally, typically involving one brain region and one model layer, with no recourse to behaviour (i.e., the object recognition decision). (**C**) We propose a stronger test of correspondence based on task-relevant variance. If a model layer and brain region correspond, then model activity replaced with brain activity should drive the DCNN to an appropriate output (i.e., decision). The quality of correspondence is evaluated by comparing DCNN performance when driven by a stimulus image vs. interfaced with brain activity.

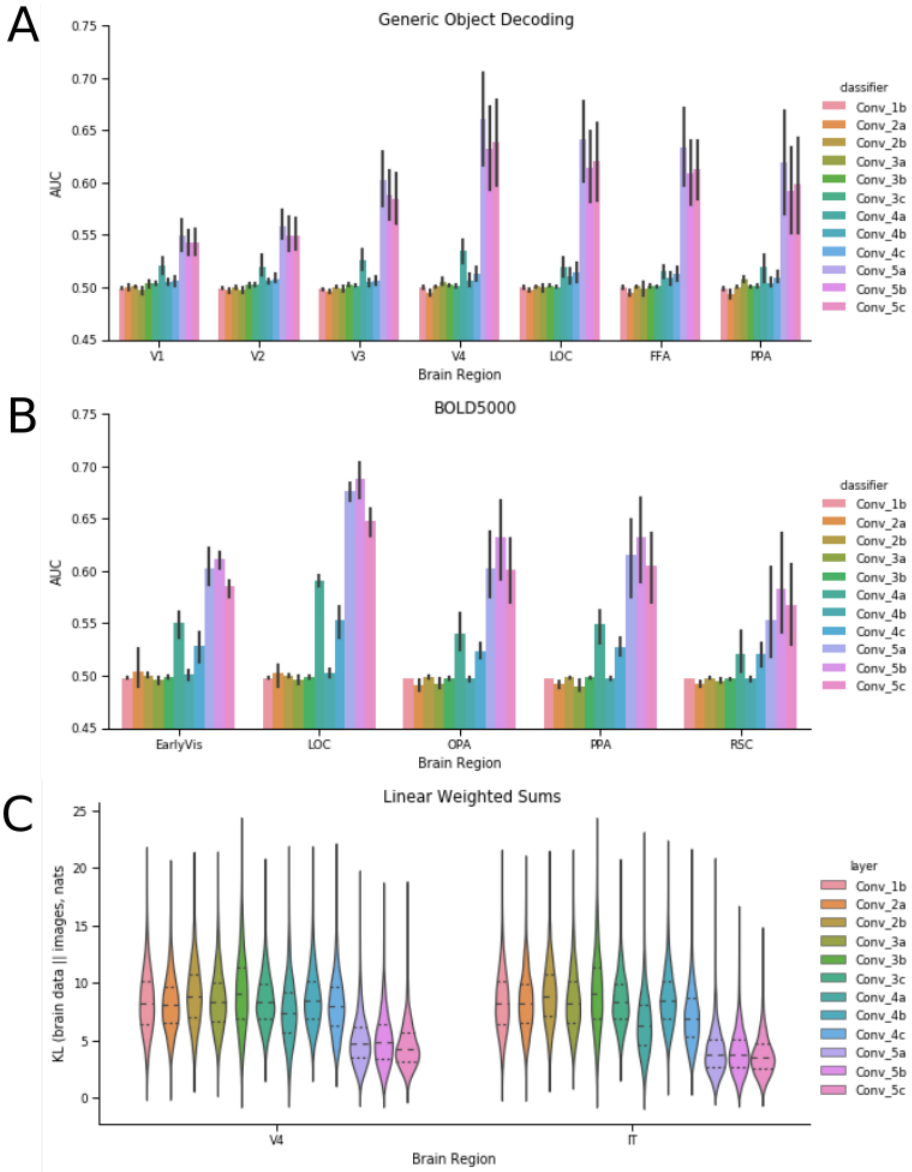

Figure 2: Results from interfacing neural data with a Deep Convolutional Neural Network (DCNN). Using the method shown in fig 1C, brain activity is directly inputted to a model layer to assess correspondence between a brain region and model layer. (**A**) For this human fMRI study(*15*), all brain areas drive DCNN object recognition performance to above chance levels. Performance is best for all brain areas when interfaced with later model layers. (**B**) The same pattern of results is found for a second human fMRI study(*14*). (**C**) In a third study, KL divergence is used (see main text and SI) to measure the degree of correspondence for when the DCNN is driven by image input vs. multi-unit recordings from macaque monkeys(*16*). For KL divergence, lower values indicate better correspondence. Once again, all regions best correspond to later network areas. These three analyses indicate that higher-level visual information is present at all stages along the ventral visual stream.

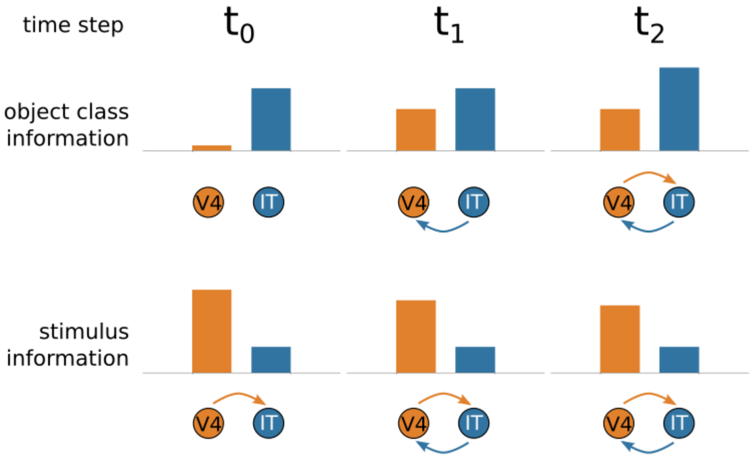

Figure 3: Hypothesised interactions between early (V4) and late (IT) regions along the ventral visual stream as processing unfolds. We hypothesise how stimulus and object-class information propagates between V4 and IT over time. At t_0, the forward pass reaches IT from V4, with V4 activity reflecting low-level stimulus properties but little information about object class. At t_1, object-class information from IT flows back to V4, increasing its task-relevant activity, which in turn influences IT at t_2. Notice that later in processing, V4 reflects object class information, but most of its activity remains tied to bottom-up stimulus properties. These hypothesised interactions would reconcile our results (fig 2) based on task-relevant information with previous results based on shared variance.

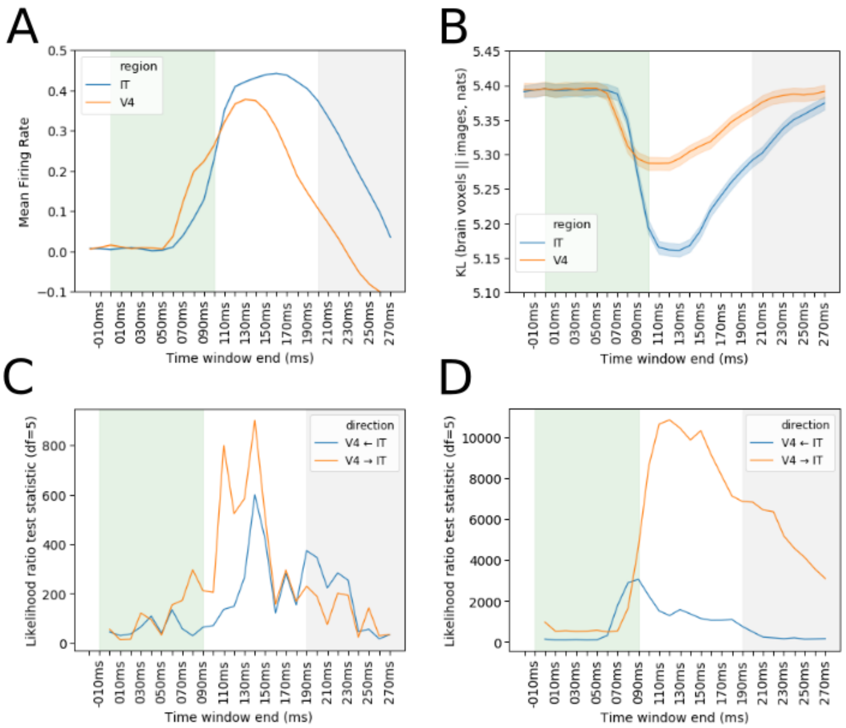

Figure 4: Analyses of monkey multi-unit recordings(*16*) time locked to stimulus presentation in 10ms time bins. Each visual stimulus was presented for 100ms (shaded green) with 100ms before the next (shaded grey). (**A**) Mean normalized spike counts for all electrodes for V4 and IT. (**B**) Task-relevant analysis (lower values imply closer correspondence with a late DCNN layer) show both V4 and IT can appropriately drive DCNN response (fig 1B), starting around 70ms after stimulus onset. (**C**) Consistent with our long-range recurrence hypothesis (fig 3), Granger Causal Modelling indicates that, while V4 first drives IT in terms of raw firing rates ($V4 \rightarrow IT$), (**D**) IT first drives V4 in terms of task-relevant information ($V4 \leftarrow IT$). These results are consistent with information about object category information (as assessed by interfacing with a late layer in a DCNN) first arising in IT and then feeding back to V4. At later time steps, Granger causality between V4 and IT becomes reciprocal ($V4 \leftrightarrow IT$) as the loop cycles.

References

1. Buckner, C. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence* **2,** 731–736 (2020).
2. Geirhos, R. *et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness* in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (2019).
3. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* **35,** 10005–10014 (2015).
4. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology* **10,** e1003915. ISSN: 1553-7358 (2014).
5. Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111,** 8619–8624 (2014).
6. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73,** 415–434 (2012).
7. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences* **17** (2013).
8. Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592,** 86–92 (2021).
9. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience* **22** (2019).
10. Kragel, J. E., Morton, N. W. & Polyn, S. M. Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **35,** 2914–2926 (2015).
11. Purcell, B. A. *et al.* Neurally constrained modeling of perceptual decision making. *Psychological Review* **117** (2010).
12. Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J. & Van Maanen, L. Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology. Model-based Cognitive Neuroscience* **76,** 65–79 (Feb. 2017).
13. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs].* arXiv: 1409.1556 (2014).
14. Chang, N., Pyles, J. A., Gupta, A., Tarr, M. J. & Aminoff, E. M. BOLD5000: A public fMRI dataset of 5000 images. *Scientific Data* **6** (2019).

15. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications* **8,** 1–15 (2017).

16. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience* **35,** 13402–13418 (2015).

17. McGugin, R. W., Gatenby, J. C., Gore, J. C. & Gauthier, I. High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences* **109** (2012).

18. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience* **22,** 974 (2019).

19. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences* **116,** 21854–21863 (2019).

20. Kubilius, J. *et al. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs* in *Advances in Neural Information Processing Systems* **32** (2019).

21. Nayebi, A. *et al. Task-driven convolutional recurrent models of the visual system* in *Advances in Neural Information Processing Systems* (2018), 5290–5301.

22. Kell, A. *et al.* Computational similarities between visual and auditory cortex studied with convolutional neural networks, fMRI, and electrophysiology. *Journal of Vision* **15** (2015).

23. Schrimpf, M. *et al.* The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv* (2020).

24. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. en. *International Journal of Computer Vision* **115,** 211–252. `https://doi.org/10.1007/s11263-015-0816-y` (2015).

25. Chollet, F. *Keras* 2015. `https://github.com/fchollet/keras`.

26. Roads, B. D. & Love, B. C. Enriching ImageNet with Human Similarity Judgments and Psychological Embeddings. *arXiv:2011.11015 [cs]* (2020).

27. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. Diverse deep neural networks all predict human IT well, after training and fitting. *bioRxiv* (2020).

28. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]* (2015).

29. Hastie, T. *The elements of statistical learning : data mining, inference, and prediction /* (Springer, 2009).

30. Schrimpf, M. *et al.* Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. en. *Neuron* **108** (2020).

31. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience* **37** (2014).

32. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12,** 2825–2830 (2011).

5 Acknowledgements

The authors thank colleagues in the LoveLab for discussion and comments on early versions of this manuscript. **Funding:** This research was supported by NIH Grant 1P01HD080679 (https://www.nih.gov/), Royal Society Wolfson Fellowship 183029 (https://royalsociety.org/), and a Wellcome Trust Senior Investigator Award WT106931MA (https://wellcome.org/) held by B.C.L. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Competing Interests:** The authors have declared no competing interests exist. **Author Contributions:** N.J.S: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing - original draft, writing - review & editing, visualization. B.C.L.: Conceptualization, methodology, resources, writing - review & editing, supervision, funding acquisition.

6 Methods and Materials

6.1 Datasets

We re-analysed three existing neural datasets. Two, BOLD5000(*14*) and Generic Object Decoding(*15*) consist of fMRI from human subjects who viewed images taken from Imagenet(*24*), a benchmark large dataset of natural images. We restricted the BOLD5000 dataset to only those images drawn from Imagenet (2012 ILSVRC) edition and to subject 1-3 who completed the full experiment. The analysis of Generic Object Decoding used the data from the 'training' portion of their image presentation experiment, consisting of 1200 images from 150 categories drawn from the Imagenet Fall 2011 edition. For both datasets, each image was presented once, thus each row represents individual trials.

The third dataset consists of neuron spike counts directly recorded from V4 and IT of two macaque monkeys (*16*), in a rapid serial visual presentation paradigm where each image is passively viewed for 100ms, with 100ms between images. We used the publicly available data processed as detailed in those publications. For the neural interfacing analysis of the spiking neural dataset, we used spike rates aggregated over multiple presentations of each of 3200 unique images, in the interval 70-170ms after stimulus onset, with the electrodes from the two subjects concatenated, as in the original analysis (*16*). For the Granger causal modelling analysis of the same dataset, we used spike rates at the level of the individual trial (i.e., no aggregation) for each 10ms time bin.

The neural data corresponding with each image was related to layer activations of a deep convolutional neural network (DCNN) trained on image classification, when processing the same pixel-level data. The three neural datasets contain data for various brain regions from ventral stream, including visual areas (V1, V2, V3 V4, included as 'EarlyVis' in (*14*)), areas responsible for processing shape and conceptual information (LOC, IT) and various downstream areas (OPA, PPA, FFA, RSC).

For details on neuroanatomical placement or functional localisation of each region, we refer readers to the original publications. Further details of brain regions and dimensionality of the data from each region are presented in supplementary information table 1.

6.2 DCNN

As the base DCNN for all simulations, we used a re-implemented and trained version of VGG-16 (*13*) (configuration D) using Keras(*25*) version 2.2.4 and TensorFlow version 1.12. This model was selected for its uncomplicated architecture, near-human level classification accuracy on ImageNet, and widely reported robust correspondence with primate or human data on various measures, including human behavioural (similarity judgements (*26*), human image matching (*20*)) and neural (*20, 27*). We implemented and trained a version of the architecture with 64×64×3 input size, with corresponding changes in spatial dimensions for all layers (table 2). For all analyses, images from all datasets were cropped to a square and resized to this resolution. For the monkey multi-unit dataset, where images are contained in a circular frame, the central 192 × 192 portion of the 256 × 256 original was cropped and resized, to decrease the proportion of image taken up by blank space in the corners. While the original authors trained their network in a two-stage process, beginning with a subset of the layers, the inclusion of batchnorm(*28*) between the convolution operation and activation function of each layer enabled training the complete network in a single pass. We used the authors setting for weight decay (ℓ_2 penalty coefficient of 5×10^{-4}) and a slightly different value for dropout probability (0.4). Model architecture details are presented in the supplementary information (table 2).

6.3 DCNN training

Our training procedure followed (*13*). The model was trained on ImageNet 2012 (1000 classes) for analyses of the BOLD5000 and monkey multi-unit datasets. For the Generic Object Decoding dataset, the model was trained to convergence on ImageNet Fall 2011 (21841 classes), before layer FC3 was replaced and retrained with 150 classes, corresponding with the classes used in our re-analysis of (*15*). For ImageNet Fall 2011 we randomly allocated 2% of each class including all images used in (*15*) to an in-house validation set that was not used for training. One image used by (*15*) was missing from our image dataset and was excluded from all analyses. All images were resampled from their native resolution to $64 \times 64 \times 3$ by rescaling the shortest side of the image to 64 pixels and centre-cropping.

Both versions of the model was trained using mini-batch stochastic gradient descent, with a batch size of 64, an initial learning rate of 0.001 and Nesterov momentum of 0.90561. The learning rate decayed by a factor of 0.5 when validation loss did not improve for 4 epochs, with training terminating after 10 epochs of no improvement. All layers used Glorot normal initialization. During training, images were augmented with random rescaling, horizontal flips and translations.

6.4 Cross-validation

Classifier-based methods require training classifier parameters, before evaluating it on data withheld from the training set. In all analyses, we use the standard approach of k-fold cross validation(*29*), in which the dataset is randomly allocated into k equally-sized partitions, and the analysis is iterated k times, each time training on $k-1$ partitions and evaluating on one. In this way, the classifier is evaluated over the entire dataset. For all analyses, except where otherwise specified, we use stratified 8-fold cross validation, that is to say dataset items are randomly allocated to partitions with the constraint that $1/k$ of each class be allocated to each validation partition. For the spiking neural dataset(*16*), each unique image was rendered from one of 64 objects, with varying position and orientation. Here, stratification was done at the object level.

For the out-of-training-class generalisation analysis, we used leave-one-class-out cross validation, where for m classes, the analysis is iterated m times, the evaluation set consisting only of the entirety of a single class, on each iteration.

6.5 Neural Interfacing analysis

Given a dataset D, consisting of an image matrix D_i of shape $(n, 64, 64, 3)$ where n is the number of images, and a corresponding neural data matrix D_r, of shape (n, d) where d is the number of neural features (electrodes, for multi-unit data, or voxels, for fMRI data), consider a DCNN computing a function f on D_i, mapping D to P_i, an (n, m) matrix of predictions, each row being a probability distribution over the m classes the DCNN was originally trained to classify.

$$f(D_i) = P_i \tag{1}$$

For an arbitrary intermediate model layer q, we may decompose f into g_q and g'_q, by computing intermediate activations, $g_q(D_i)$:

$$f(D_i) \equiv g'_q(g_q(D_i)) = P_i \tag{2}$$

The neural interface analyses proceeded by applying a linear transform W to the centered and column-normalized neural data, D_r and inputting the result into DCNN layer q, to compute a matrix of model predictions for the neural data, P_r.

$$g'_q(WD_r) = P_r \tag{3}$$

6.5.1 Linear transformation matrix training

The transformation matrix W was computed by partitioning image and neural datasets D_i, D_r into training and evaluation partitions using 8-fold cross-validation, and W was learned as a linear mapping from D_r to the layer q activations generated by the corresponding images, D_i, on the training partition:

$$g_q(D_i) = WD_r + \epsilon \tag{4}$$

For each cross-validation fold, the model predictions were computed for the evaluation partition. In practice, W was computed as a single-layer linear neural network with no bias or activation function, to minimise mean-squared error of supervision targets $g_q(D_i)$ using mini-batch stochastic gradient descent with momentum, (batch size 64, momentum of 0.9, l_2 regularization of 0.0003, initial learning rate of 0.1, decreasing by a factor of 0.5 when validation loss did not improve for 4 epochs and terminating after 400 epochs or after validation loss did not improve for 20 epochs.) For the analysis of the macaque dataset(*16*) on the level of the individual trial, prior to performing the GCM model, W was computed using the Adadelta optimizer (batch size of 128, initial learning rate of 0.04.)

We also considered an alternative mode for training W, by first assembling the model in the form of equation 3, composed of transformation matrix W initialised with small random weights, followed by DCNN layer q onwards, g'_q, thus mapping end-to-end from neural measures D_r to output. W was then trained by back-propagating the categorical cross-entropy error term from the softmax output layer, using the supervision target of the ground-truth labels for the neural dataset (D_r), with all other weights in the network frozen. This method produced a pattern of results that was qualitatively similar, although with lower absolute accuracy (SI fig 8).

6.5.2 Neural Interface Evaluation

The output of the model, P, is an (n, m) matrix of probability distributions over the m output classes the original DCNN was trained on, for each of n images in D. We computed this for the original DCNN on the image dataset, $f(D_i) = P_i$, and also for the neural dataset for each brain region r and model layer q, $g'_q(WD_r) = P_r$. The correspondence between r and q was evaluated by comparing the model predictions P_r either against the ground-truth classes (by computing the overall AUC of the classifier, via the equality between AUC and Wilcoxon-Mann-Whitney U) or against model predictions from the image dataset, by computing the KL divergence of P_r from P_i for each row n.

6.6 Shared Neural Variance Analysis

For comparison, we present an example of a shared neural variance analysis using the macaque spiking neuron dataset (*16*) and our re-implemented model. Conceptually, in common with the interfacing analysis (section 6.5), the analysis evaluates the correspondence between a brain region r and a model layer q. Layer q model activations, $g_q(D_i)$, were compared with a neural dataset obtained from the presentation of corresponding images, D_r. To establish our results are comparable to those previous, we used the neural predictivity method exactly as implemented in the Brainscore benchmark for DCNNs (*30*).

The dataset was iteratively partitioned using 8-fold cross-validation into training/validation partitions. Following the method of (*30*), we used the image stimuli from the training partition to generate model activations on each layer. We used PCA to calculate the first 1000 principal components of these activations, before training a PLS regression model (25 components) to predict, for each electrode, the firing rate across the validation partition. The predictivity for each electrode was computed as the Pearson correlation coefficient between the predicted firing rates across the dataset and the actual recorded values, with the overall predictivity given by the correlation coefficient of the median electrode.

6.7 Simple Classifiers on the Neural Datasets

To establish performance baselines for the interfaced fMRI datasets, which were evaluated in terms of classification performance, we applied various standard classifiers to the neural data directly, to predict the image class from the neural data from various brain regions. Known as multi-voxel pattern analysis (MVPA), evaluating the trained classifier's ability to predict class labels from fMRI or spiking neural data is now a standard approach to quantifying the categorical-level information within a brain region (*31*). Nevertheless, in the present analyses the number of different classes is unusually large, and the number of examples from each class unusually small, (1916 images from 958 classes(*14*), 1200 images from 150 classes(*15*)) for a straightforward MVPA analysis on these datasets. We report the AUC of the classifier computed in the same way as for the neural interfacing analysis 6.5.2. All classifiers were implemented as detailed below using version 0.20.3 of the Scikit-Learn library (*32*).

6.7.1 Multiclass Logistic Regression

Implemented as LogisticRegression with the 'multinomial' option, the lbfgs solver and a maximum of 10^3 iterations.

6.7.2 Nearest Neighbours Classifier

Implemented as KNeighboursClassifier. Given the structure of the BOLD5000 dataset, with only two examples per class (thus, either one or two examples in the training partition, test classification of each class on the basis of one correct training example) we classified on the basis of the single nearest neighbour under a Euclidean distance function.

6.7.3 Linear Support Vector Machine (SVM)

Implemented as LinearSVC, using a one-versus-rest multi-class strategy, with a maximum of 10^4 iterations and C parameter of 10^{-3}.

6.8 Granger Causal Modelling

In contrast to the previous neural interfacing analysis of the spiking neural dataset, which aggregated spike rates over multiple presentations of each image, in the interval 70-170ms after stimulus onset, here we trained and evaluated the model on data at the individual trial level. We conducted a separate decoding analysis for each 10ms time bin, from -20ms (i.e., prior to stimulus onset) to 270ms, with all time indices referring to the preceding 10ms bin. Training linear transformation matrix W is described in section 6.5.1. Prior to the GCM, we pre-processed the trial-level relative entropy data to ensure stationarity by, first, subtracting the temporal mean and standard deviation from each trial, and second, subtracting the mean signal and dividing by the signal's standard deviation, thus ensuring that each time step has zero mean and unit variance.

Given two regions, X and Y, separate Granger-Causal models were computed for each direction $X \rightarrow Y$ and $X \leftarrow Y$, where each model takes the form of a linear regression, where the univariate outcome:

$$KL(D_X||D_i)_n \tag{5}$$

the KL divergence of region X with θ, the base model predictions, is predicted by the Granger null model (6), or the Granger-causal model (7).

$$KL(D_X||D_i)_{n-1}, ..., KL(D_X||D_i)_{n-p} \tag{6}$$

$$KL(D_X||D_i)_{n-1}, KL(D_Y||D_i)_{n-1}, ..., KL(D_X||D_i)_{n-p}, KL(D_Y||D_i)_{n-p} \tag{7}$$

where p, the maximum number of previous time-steps is a hyperparameter that is determined using model-selection criteria such as BIC. The appropriate model was determined by comparing log-likelihood ratios, given the data, for the causal and null models.

A Appendix

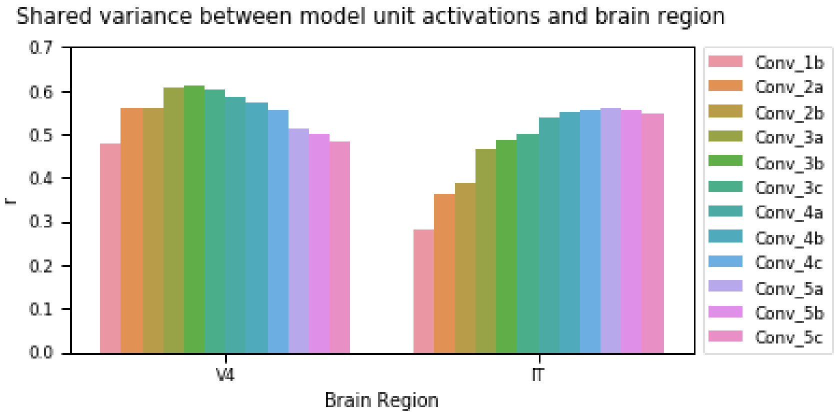

Figure 5: Standard approaches to relating primate ventral stream and DCNNs evaluate variance shared between data from each brain region and unit activations on each DCNN layer. They have been taken as evidence that earlier ventral stream regions (e.g., V4) correspond to earlier DCNN layers and later regions (e.g., IT) correspond to later DCNN layers. Here, we present a shared-variance based analyses of directly recorded spiking neural activity(*16*) and VGG-16 using an established method (*20*). Higher correlations reflect more shared variance between brain region and model layer.

531 B Supplementary Figures

532 C Tables

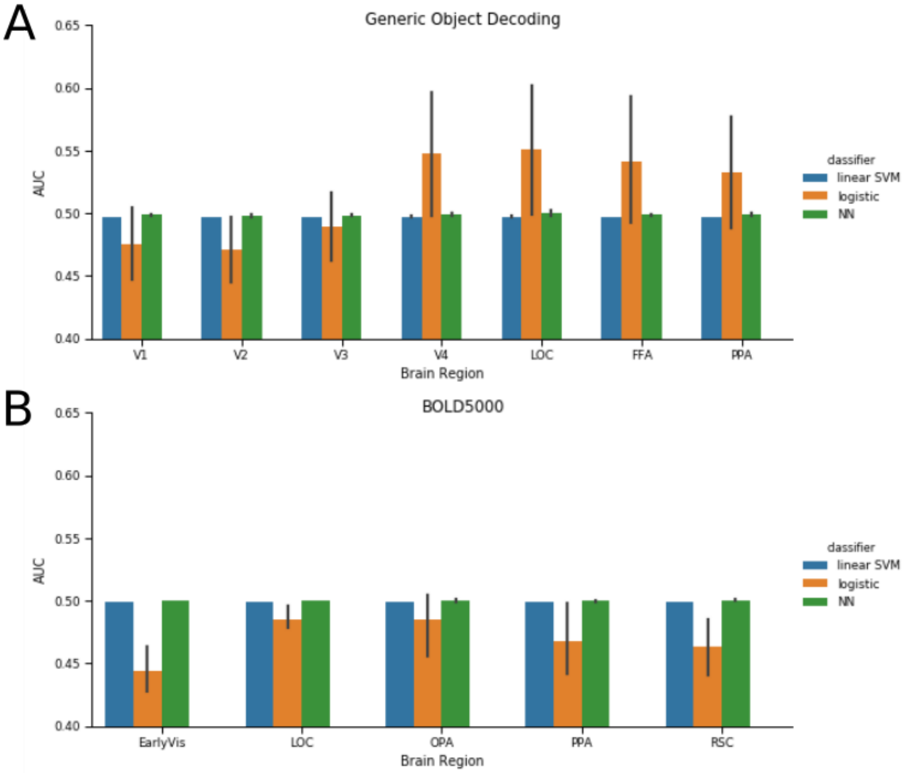

Figure 6: Comparison of the brain-DCNN interface as a neural pattern classifier, compared with standard linear classifiers typically used in multi-voxel pattern analysis (MVPA). We present classification performance (AUC) directly on neural patterns, on the Generic Object Decoding (**A**) and BOLD5000 (**B**) datasets, for simple classifiers (support vector machine with linear kernel, multiclass logistic regression, and a 1-nearest neighbour classifier) with results for interface with each layer of the DCNN presented for comparison. Performance of the simple classifiers is generally near chance (0.5), we attribute this to the large number of image classes (150, 958 respectively) and few available examples (2, 8 per class) which severely limit the available training data. Because the brain-DCNN interface learns a general mapping between brain region and model, it does not suffer this limitation, making it an appealing novel approach for MVPA.

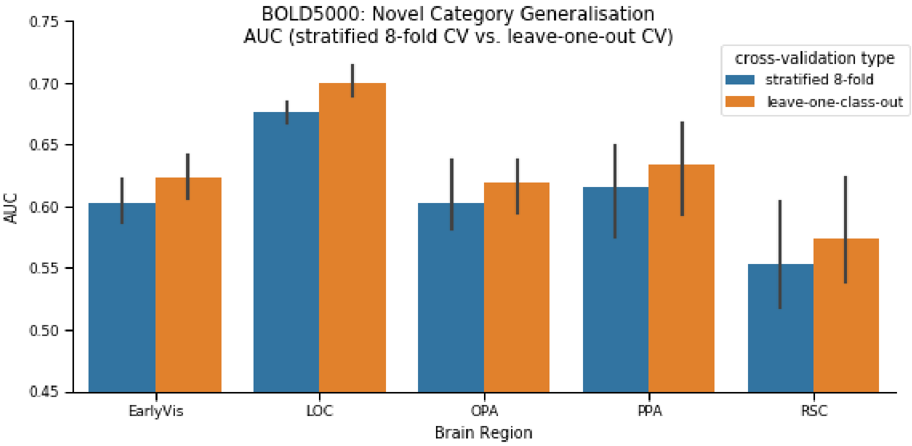

Figure 7: Learning a mapping directly from neural measures to DCNN activation space produces a general mapping, rather than being dependent on training examples. The neural interface has an in-built ability to generalise to novel classes. This is demonstrated by presenting classification performance (AUC) on the BOLD5000 dataset, by comparing cross-validation (CV) strategies. Error bars represent 95% confidence intervals across 3 subjects. Stratified 8-fold CV (the default, used in all other analysis) ensures each training partition contains at least one example of each class. Leave-one-class-out CV involves the same number of CV folds as there are classes, each time training on all data except one class, which is withheld for the validation set. Performance is equivalent or better (LOC) when generalising to novel classes, which we attribute to more training data per CV fold. Due to the training time, this analysis was restricted to layer 5a.

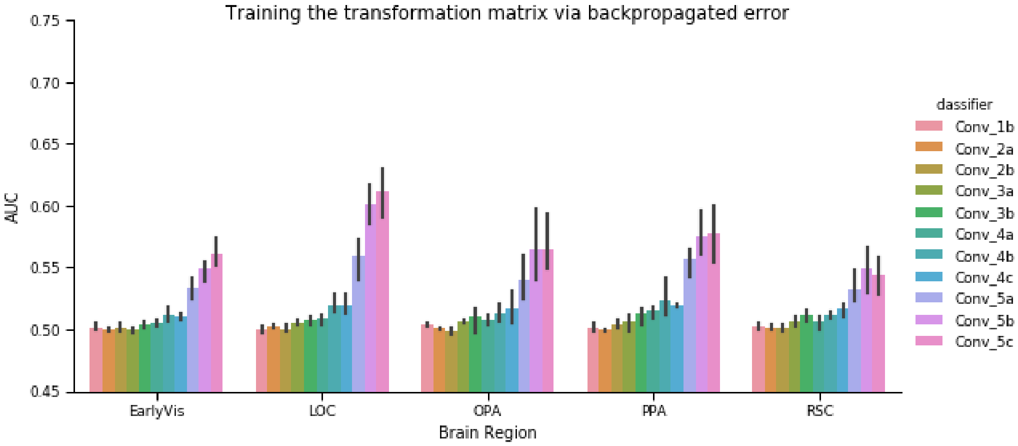

Figure 8: Alternative ‘backprop mode’ for training the transformation matrix W mapping from neural space to DCNN activation space. Classification performance (AUC) on the BOLD5000 dataset follows a qualitatively similar pattern to the main analysis (compare figure 2B), albeit with lower absolute accuracy. The default analysis trains W independently as a regression problem, using layer activations as supervision targets directly. Instead, this approach uses W as a weights matrix for a new neural network that takes neural data from a brain region as input, connected to the latter part of the DCNN, and training the network using the class labels as supervision targets, with all other DCNN weights frozen.

Dataset	Generic Object Decoding(*15*)	BOLD5000(*14*)	Linear Weighted Sums(*16*)
Stimuli	experiment 'train' phase: 1200 images from 150 categories (ImageNet Fall 2011)	1916 images from 958 categories (ImageNet ILSVRC 2012	3200 greyscale composite images, 64 objects in 8 categories,) non-congruent background
Task	one-back repetition detection	valence judgement ('like', 'neutral', 'dislike')	passive viewing, RSVP presentation 100ms/100ms
Subjects	5 human fMRI	3 human fMRI (partial data from subject 4 excluded)	2 Macaque monkeys (vectors concatenated) multi-unit recording
Time indices	full 9s of image presentation	TR3-4	70-170ms
Brain region (dimensionality per subject)	V1 (1004, 757, 872, 719, 659) V2 (1018, 944, 1031, 855, 891) V3 (759, 810, 861, 929, 907) V4 (740, 544, 754, 704, 860) LOC (540, 834, 996, 668, 566) PPA (356, 316, 496, 398, 550) FFA (568, 435, 928, 725, 929)	EarlyVis (495, 495, 1218) LOC (342, 888, 1027) OPA (288, 180, 392) PPA (331, 370, 273) RSC (229, 421, 394)	IT $(168 = 58 + 110)$ V4 $(88 = 70 + 18)$

Table 1: **Neural datasets** For further dataset details, such as how regions were defined, we refer readers to the original publications

Block	Layer	Dimensions ($h \times w \times c$)	Filter Size
Input		$64 \times 64 \times 3$	
1	1a	$64 \times 64 \times 64$	3×3
	1b	$64 \times 64 \times 64$	3×3
	max pool 1		2×2
2	2a	$32 \times 32 \times 128$	3×3
	2b	$32 \times 32 \times 128$	3×3
	max pool 2		2×2
3	3a	$16 \times 16 \times 256$	3×3
	3b	$16 \times 16 \times 256$	3×3
	3c	$16 \times 16 \times 256$	3×3
	max pool 3		2×2
4	4a	$8 \times 8 \times 512$	3×3
	4b	$8 \times 8 \times 512$	3×3
	4c	$8 \times 8 \times 512$	3×3
	max pool 4		2×2
5	5a	$4 \times 4 \times 512$	3×3
	5b	$4 \times 4 \times 512$	3×3
	5c	$4 \times 4 \times 512$	3×3
	max pool 5		2×2
FC	FC1	4096	
	dropout 1		
	FC2	4096	
	dropout 2		
	FC3 (output)	1000 softmax	

Table 2: **DCNN Architecture:** Layer configuration and dimensions of the DCNN used for all analyses.