Convolutional architectures are cortex-aligned de novo

Atlas Kazemian[1], Eric Elmoznino[1,2], and Michael F. Bonner[1]

[1]Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA
[2] University of Montreal and MILA - Quebec AI Institute

Please address correspondence to:
Atlas Kazemian (ataskazemian@gmail.com) or Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218

ABSTRACT

What underlies the emergence of cortex-aligned representations in deep neural network models of vision? The success of widely varied architectures has motivated the prevailing hypothesis that large-scale pre-training is the primary factor underlying the similarities between brains and neural networks. Here, we challenge this view by revealing the role of architectural inductive biases in models with minimal training. We examined networks with varied architectures but no pre-training and quantified their ability to predict image representations in the visual cortices of both monkeys and humans. We found that cortex-aligned representations emerge in convolutional architectures that combine two key manipulations of dimensionality: compression in the spatial domain and expansion in the feature domain. We further show that the inductive biases of convolutional architectures are critical for obtaining performance gains from feature expansion—dimensionality manipulations were relatively ineffective in other architectures and in convolutional models with targeted lesions. Our findings suggest that the architectural constraints of convolutional networks are sufficiently close to the constraints of biological vision to allow many aspects of cortical visual representation to emerge even before synaptic connections have been tuned through experience.

INTRODUCTION

Understanding the computations that transform patterns of light on the retina into representations that lead to complex behavior is a longstanding challenge in neuroscience. Early work focused on theory-guided, hand-engineered models for simulating image-evoked neural responses, but these models had limited success beyond primary visual cortex (Carandini et al., 2005; Jones & Palmer, 1987; Movshon et al., 1978; Riesenhuber & Poggio, 1999). Following the advent of deep learning, it was discovered that deep neural networks (DNNs) could achieve unprecedented performance in predicting image-evoked responses in the ventral visual stream, and they have since become the leading theoretical models of visual computation in the brain (Agrawal et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). However, despite over a decade of research, it remains an open question what key factors underlie the representational similarities between DNNs and visual cortex (Conwell et al., 2022; Elmoznino & Bonner, 2024; Saxe et al., 2020; Serre, 2019; Storrs et al., 2021).

A prominent theory proposes that cortex-aligned representations emerge in DNNs when their constraints and objectives match those of biological vision (Cao & Yamins, 2021a, 2021b). This theory builds on foundational work in neuroscience arguing that biological visual representations are strongly constrained by neuroanatomic architecture and optimized for specific ethologically relevant objectives (Cao & Yamins, 2021b; Konkle & Alvarez, 2022; Kriegeskorte, 2015; Richards et al., 2019; Yamins & DiCarlo, 2016). However, recent findings have challenged this constraint-based perspective. Most importantly, it has been shown that DNNs with widely varied architectures and training objectives predict ventral-stream representations equally well after being pre-trained on massive datasets (Conwell et al., 2022; Storrs et al., 2021). This suggests that for highly trained DNNs, architectural and objective-based constraints are relatively weak (Conwell et al., 2022; Elmoznino & Bonner, 2024). There is, thus, a crucial and unresolved question: How can we reconcile the apparent degeneracy of DNN designs with the strong role that architectures and objectives are thought to play in shaping the representations of biological vision?

In this article, we present a new perspective on this problem by examining how the encoding performance of untrained DNNs changes as a function of targeted architectural manipulations. By removing the influence of massive pre-training, our approach reveals the surprising extent to which architectural factors alone can promote the emergence of cortex-aligned representations. Specifically, we scaled the number of random features in convolutional, fully connected, and transformer architectures and quantified their ability to predict image-evoked responses in the visual cortices of monkeys and humans. Our results showed a strong effect of scaling that was specific to convolutional architectures. Scaling the deeper layers of a

convolutional network yielded striking performance gains—the best model even approached the performance of a classic pre-trained network in the monkey data. In contrast, scaling the features of fully connected and transformer architectures yielded only small improvements, resulting in substantially worse performance than the convolutional architecture despite being matched on dimensionality. Further, the performance of the convolutional architecture depended critically on spatial locality and nonlinear activation functions, again demonstrating that the benefits of scaling were highly architecture-dependent.

Together, these findings show that the architectural biases imbued into convolutional networks allow many aspects of cortical visual representation to readily emerge even before synaptic connections have been tuned through experience. Thus, while pre-training on large-scale datasets may be sufficient to induce cortex-aligned representations in many degenerate architectures, the representations of convolutional architectures exhibit a remarkable degree of cortical alignment in their initial random states.

RESULTS

Scaling untrained neural networks to predict visual cortex representations

We examined how the encoding performance of untrained neural networks changed when scaling the dimensionality of their features. We focused on convolutional, fully connected, and transformer architectures. The convolutional architecture, which we refer to as the Expansion model, contained a hierarchy of layers that each perform the following operations: convolution, nonlinear activation, and pooling (Figure 1a). The first layer contained a pre-defined set of wavelet filters, but all other layers contained randomly initialized weights. Our key manipulation was the number of random convolutional filters in the last layer. We also analyzed fully connected and transformer architectures and manipulated the number of randomly initialized features in their last layers. Detailed descriptions of network architectures can be found in the Methods.

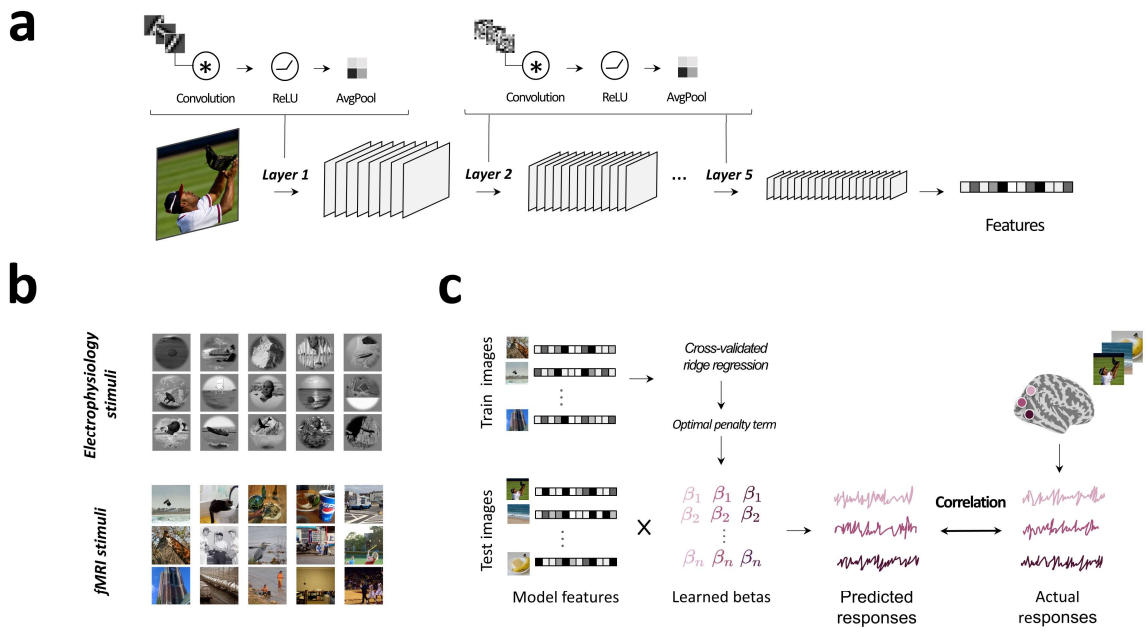

Figure 1. Convolutional model architecture and evaluation framework. a) Our convolutional architecture consists of a repeating set of modules that are stacked in a hierarchy of five layers. In the first layer, images are convolved with a set of pre-defined wavelet filters, then passed through a rectified linear unit (ReLU) activation function, and an average pooling operation. In the deeper layers, inputs are convolved with a large set of randomly initialized filters, followed by ReLU and average pooling. Features were extracted from the last layer of the network and used to evaluate cortical encoding performance. **b)** The top panel shows examples of stimuli from the monkey electrophysiology dataset (Majaj et al., 2015), and the bottom panel shows examples of stimuli from the human fMRI dataset (Allen et al., 2022). **c)** To evaluate cortical encoding performance, model image representations were linearly mapped to image-evoked cortical responses using ridge regression. Regression coefficients and the optimal ridge penalty parameter were estimated on a set of training images and performance was evaluated on held-out test images. Encoding performance scores were obtained by computing correlations between the predicted and actual cortical responses in the test set.

We quantified the ability of these networks to predict image-evoked responses in visual cortex in both monkey electrophysiology and human fMRI data. The monkey dataset consists of single-unit recordings from two macaques who viewed 3200 images of objects on synthetic backgrounds (Majaj et al., 2015). We considered all units in V4 and IT regions from both macaques. The second dataset, called The Natural Scenes Dataset (NSD) (Allen et al., 2022), is the largest existing human fMRI dataset on natural scene perception, containing responses to ~73,000 images across 8 participants. We considered fMRI responses from three regions of interest (ROIs) from all participants: early, midventral, and ventral visual streams. Example stimuli from both datasets are shown in Figure 1b. Using regularized linear regression, we fit encoding models that mapped the final layer of each neural network to the responses of electrophysiology units or fMRI voxels. These encoding models were estimated on a set of

training images, and their performance was evaluated on a held-out set of test images. Encoding performance was measured as the correlation between predicted and actual cortical responses in the test set (Figure 1c).

We examined the degree to which encoding performance increased as a function of dimensionality expansion. We expected to find some level of performance gain from dimensionality expansion. However, what was not known was how large of a performance gain could be obtained and whether these gains differed across architectures. Our goal was to characterize the point at which encoding performance converged and no longer improved from the addition of more dimensions. We, therefore, increased the number of random features in the last layer by multiple orders of magnitude and quantified encoding performance at each level of dimensionality. The results are shown in Figure 2. As expected, all three architectures obtained performance gains from dimensionality expansion. This was true across all ROIs in both the monkey and human data. However, the degree of the performance gain was markedly different for the convolutional architecture compared with the other two architectures, even though these architectures were matched on dimensionality. Specifically, we found that dimensionality expansion strongly improved the performance of the convolutional architecture while resulting in substantially smaller improvements for the fully connected and transformer architectures. These findings show that representations with high encoding performance can be readily discovered through simple linear transformations of a random convolutional feature space, and it indicates that the inductive biases of the convolutional architecture strongly promote the emergence of cortex-aligned representational dimensions.

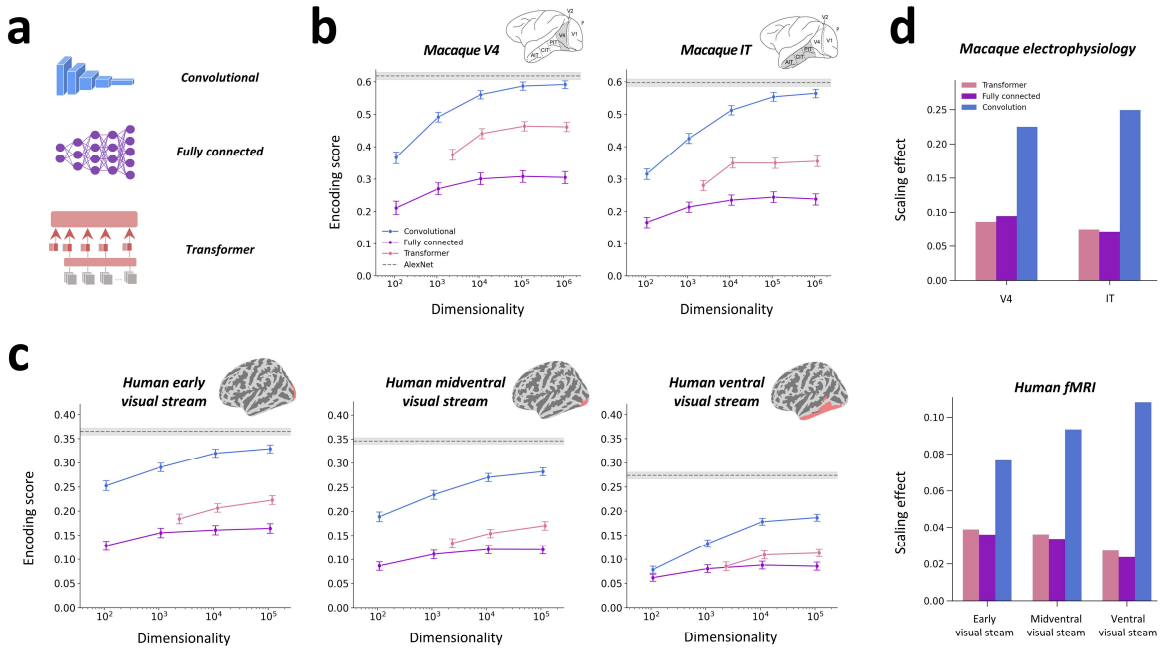

Figure 2. Dimensionality expansion strongly improves the encoding performance of convolutional neural networks. a) The effects of dimensionality expansion were examined for a convolutional neural network, a deep fully connected network, and a vision transformer. **b)** These plots illustrate the encoding performance for all three architectures as a function of dimensionality expansion. There was no pre-training for these networks. Encoding performance was evaluated for regions V4 and IT in the monkey electrophysiology data. The x-axis plots the number of random features in the output layer, and the y-axis shows the encoding score for predicting image-evoked cortical responses. The gray dashed line indicates the performance of the best performing convolutional layer of pre-trained AlexNet. The convolutional architecture without pre-training attained large performance gains as a function of dimensionality expansion and approached the performance of pre-trained AlexNet. In contrast, the other two architectures showed much less improvement as the number of output features was expanded. The encoding performance plots in this panel and in panel c show the mean performance across units/voxels from all participants, and the error bars denote 97.5% confidence intervals from 1,000 bootstrap samples. **c)** These plots show the same analyses as in panel b but for regions along the ventral stream in the human fMRI data. The encoding performance of the largest convolutional model approaches that of pre-trained AlexNet in the early visual stream and reaches ~70% of the performance of pre-trained AlexNet in the higher-level ventral stream. **d)** A summary statistic for the effect of dimensionality expansion was calculated as the difference in performance between the highest- and lowest-dimensional version of each architecture. The performance of the convolutional model was strongly modulated by the dimensionality of its random feature space. In contrast, the effect of dimensionality expansions was much weaker in the other two architectures.

For comparison, we also plotted the performance of a conventional CNN architecture (AlexNet) that was pre-trained on ImageNet classification. We found that in both the V4 and IT regions of monkey visual cortex, the largest untrained convolutional architecture approached the performance of pre-trained AlexNet, demonstrating the remarkable degree to which high-level object representations in monkey visual cortex can be predicted from random convolutional features without the need for goal-driven pre-training (Figure 2b). Interestingly, in the human

data, we observed a larger performance gap between the untrained convolutional architecture and AlexNet, and we found that this performance gap increased along the hierarchy of the ventral stream (Figure 2c). Thus, there appears to be a stronger emphasis on abstract, semantic information in the human fMRI data than in monkey electrophysiology data. One possible interpretation of this difference is that humans have semantically richer ventral stream representations than monkeys. However, another possibility is that it arises from differences in the temporal sampling of these datasets. The monkey electrophysiology data has high temporal resolution and largely captures the first wave of feedforward processing following image onset. In contrast, the human fMRI data is temporally coarse and, thus, has a stronger contribution from feedback and recurrent processes that may emphasize abstract, semantic content.

Emergent properties of untrained convolutional neural networks

The large performance gains of the expanded convolutional network suggest that brain-relevant representations readily emerge when natural images are processed with random convolutional filters. We further explored the emergent properties of untrained CNNs in two follow-up analyses. First, we performed principal component analysis (PCA) on the CNNs to identify their principal modes of variance for natural images, and we found that almost all the explained variance of the high-dimensional CNNs could be accounted for by a reduced set of principal components (PCs) (Figure 3). These analyses also show that there is a benefit from first performing dimensionality expansion using a large set of random convolutional filters before reducing the dimensionality through PCA. For example, when the model containing 10^6 random features was reduced to 10^2 PCs, its performance was still substantially better than a model with 10^2 random features, even though both models have the same number of regressors. Thus, the high-dimensional feature space appears to be critical for allowing the model to explore many possible directions of variance, but within this feature space, brain-relevant dimensions correspond to the principal modes of variance for natural images.

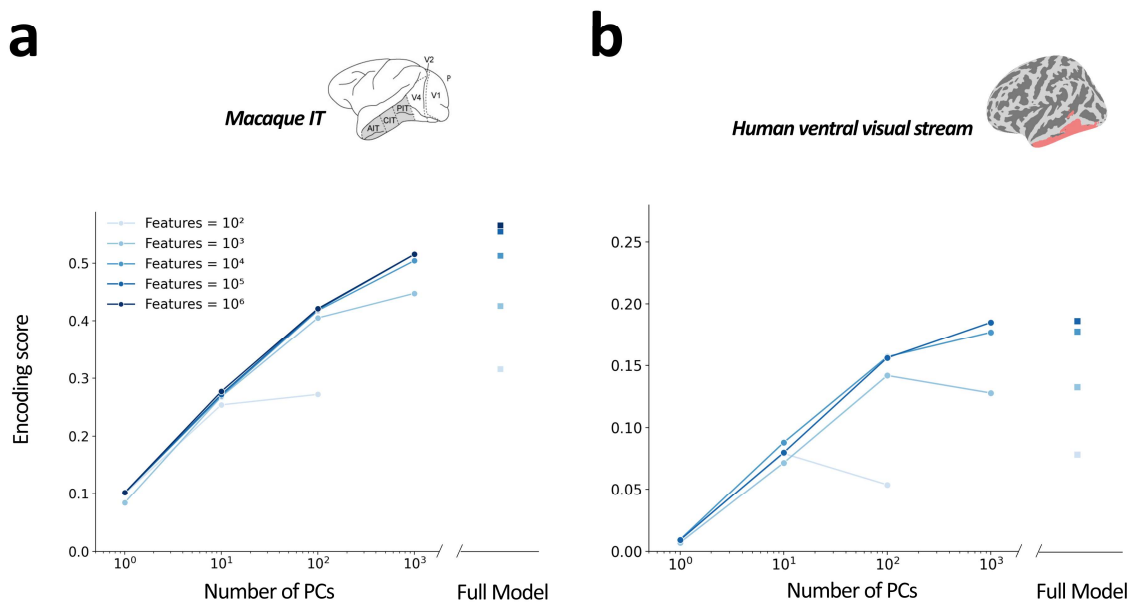

Figure 3. Encoding performance remains high after dimensionality reduction. These plots show encoding performance for models whose dimensionality was reduced through principal component analysis. Networks with high-dimensional random feature spaces in their last layer were reduced to a smaller set of principal components (PCs). PCs were identified using an independent set of stimuli, and they reflect the principal directions of variance for the natural image representations of each network. Results are shown for the monkey IT region and the human high-level ventral stream**.** Panels **a** and **b** show encoding performance for the convolutional networks from Figure 2 after their final layers were reduced to varying numbers of PCs. An inset on the right of each plot shows the performance of the original networks without dimensionality reduction. Color shades represent the number of random features in the last convolutional layer of each network. Both plots show that all networks can be reduced to a lower-dimensional subset of PCs with only a minor drop in encoding performance. For example, the dimensionality of the largest networks can be reduced by orders of magnitude but retain nearly the same level of encoding performance.

We next performed a visualization analysis to explore the representational space of our largest untrained CNN. Specifically, we visualized the image representations of the CNN in a 2D space using t-distributed stochastic neighbor embedding (tSNE). This visualization shows that despite having no pre-training for classification, the CNN forms many intuitively interpretable clusters of images, including clusters related to sports, vehicles, food, animals, and people (Figure 4). This suggests that there is a remarkable degree to which images naturally organize into semantically meaningful clusters in the representational space of high-dimensional untrained CNNs.

Figure 4: Two-dimensional visualization of image representations in a high-dimensional untrained convolutional network. Image representations from the final layer of a high-dimensional untrained convolutional network were embedded in two dimensions using t-distributed stochastic neighbor embedding. These embeddings show that even in an untrained network, images naturally form clusters that are intuitively interpretable. This visualization highlights several of these clusters, including clusters related to sports, vehicles, food, animals, and people.

Critical architectural components

Our analyses thus far show that strong encoding performance cannot be obtained through dimensionality expansion in arbitrary architectures but, instead, depends critically on the

inductive biases of the convolutional architecture. We next sought to determine which components of the convolutional architecture are most important for its performance. We performed a series of analyses in which we altered specific properties of the architecture and evaluated encoding performance in high-level visual cortex.

We first confirmed that nonlinearities were critical by removing the nonlinear activation function in all layers while keeping the rest of the architecture the same (Figure 5). As expected, we found that in both the monkey and human data, performance was worse in a network without nonlinearities and there was little-to-no benefit of dimensionality expansion. This result demonstrates that a nonlinear transformation is needed to explain the representations of high-level visual cortex.

We next examined whether the spatial locality of the convolutional filters was important for model performance (Figure 5). Spatial locality is one of the central inductive biases of convolutional neural networks, and we expected it to have a critical role in encoding performance. To examine this question, we applied a fixed permutation to image pixels, which made the inputs to the convolutional filters spatially distributed rather than spatially local. All other architectural properties remained unchanged. As expected, we found that in both the monkey and human data, there was a large drop in encoding performance in the network without spatially local image representations, demonstrating that spatial locality is critical for achieving strong encoding performance.

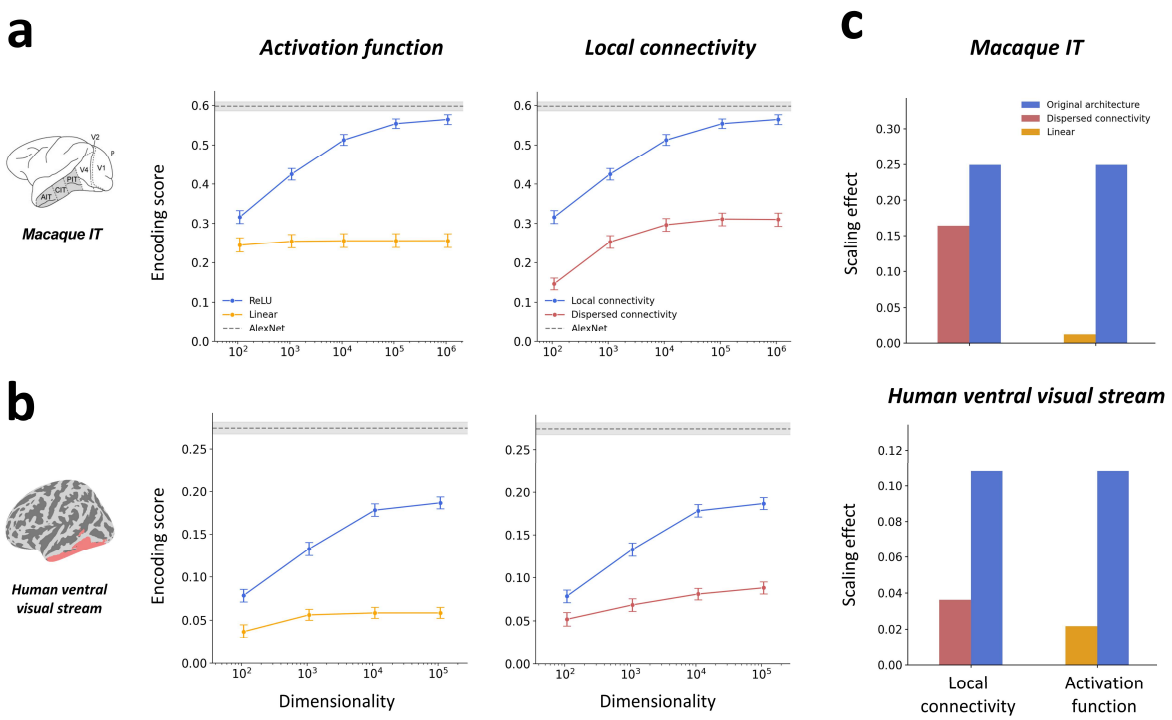

Figure 5: Analysis of critical architectural components of the convolutional network. a-b) These plots show the performance of our largest untrained convolutional network after removing key architectural components but keeping all other architectural properties intact. Panel a shows results for macaque IT, and panel b shows results for the human high-level ventral stream. The left plots show that removing nonlinearities results in a large drop in encoding performance in both the monkey and human data, and it prevents the model from improving as a function of dimensionality expansion. The right plot shows that removing the spatial locality of the convolutional filters results in an overall decrease in encoding performance, though networks with dispersed connectivity still have moderate benefits from dimensionality expansion. Plotting conventions are the same as in Figure 2. **c)** These plots show the effects of dimensionality expansion, as in Figure 2d. They demonstrate that without nonlinearities, the benefits of dimensionality expansion are removed and without spatial locality, they are diminished.

Further analyses of other architectural factors are reported in Supplementary Figures 1-3. Briefly, these analyses show that there is a small benefit from the use of pre-defined wavelet filters in the first layer and that multiple types of nonlinear activation functions and random initialization methods yield similar patterns of results.

In sum, the findings from these architectural manipulations highlight the effectiveness of the convolutional network's central inductive biases. Specifically, without nonlinear transformations and spatially local image representations, dimensionality expansion is insufficient to achieve strong encoding performance.

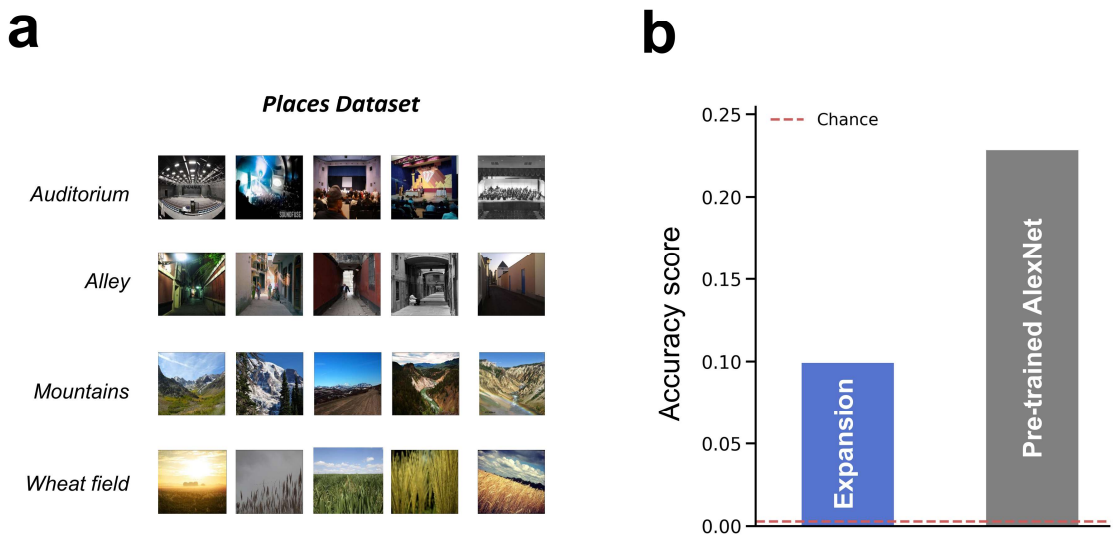

Figure 6: Image classification performance for the untrained convolutional network. a) Performance was evaluated for classification on the 365 scene categories in the Places dataset (Zhou et al., 2018). **b)** A 365-way linear classifier was fit to the principal components of features from the final convolutional layer of the Expansion model with 10^5 features, and AlexNet pre-trained on ImageNet. Classification accuracy was evaluated on held-out test images. This plot shows that the Expansion model performs poorly at image classification despite obtaining strong encoding performance scores in monkey visual cortex and early to mid-level regions of human visual cortex.

Encoding performance can be strong even when classification accuracy is weak

Pre-training neural networks on image classification improves their cortical encoding performance, and it has been argued that strong classification performance is a signature of high-performing neural network models of visual cortex (R. Cao & Yamins, 2021b; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). However, our findings show that even untrained convolutional networks can achieve good encoding performance when they are high-dimensional and have the right architectural properties, suggesting that classification performance may not be as tightly linked to encoding performance as previous work suggests. To address this question, we examined the classification performance of our largest untrained convolutional model, which we call the Expansion model here, and we compared it with AlexNet pre-trained on ImageNet. We used the Places365-Standard dataset, which contains images from 365 different scene categories (Zhou et al., 2018) (Figure 6a). We used Places rather than ImageNet so that we could evaluate both models on a dataset on which they had not previously been optimized. We fit linear classifiers to the final convolutional layer of each network, and we evaluated performance on held-out images. Figure 6b shows that the

Expansion model performs much worse than AlexNet at classifying scene categories. Thus, even though the encoding performance of the Expansion model is strong enough to rival AlexNet in monkey visual cortex, it is, nonetheless, much worse than AlexNet at classifying images. This suggests that optimization for image classification is not critical for obtaining highly predictive encoding models of visual cortex.

DISCUSSION

We show that dimensionality expansion in an untrained convolutional neural network achieves surprisingly strong performance at explaining image-evoked responses in the primate visual cortex, in some cases reaching the performance of a standard pre-trained network. This performance is not simply an effect of model scale, where any high-dimensional model would successfully predict brain responses. Instead, it reflects the effectiveness of the convolutional architecture's inductive biases, which greatly increase the likelihood of discovering brain-relevant representational dimensions through random feature sampling.

Previous work has argued that task-based pre-training is among the most important factors for obtaining brain-aligned representations in DNNs (R. Cao & Yamins, 2021b; Conwell et al., 2022; Khaligh-Razavi & Kriegeskorte, 2014; Storrs et al., 2021; Yamins et al., 2014; Yamins & DiCarlo, 2016). However, our findings show that with the right inductive biases, architectural manipulations alone can yield large improvements in encoding performance without pre-training or backpropagation. In fact, in the monkey visual cortex data, these architectural manipulations were sufficient to reach the performance of AlexNet pre-trained on ImageNet. Thus, architectural factors play a more powerful role in inducing brain-aligned representations in DNNs than previous work suggested.

Even though AlexNet is no longer state-of-the-art, it played a central role in early work that convinced neuroscientists to pursue task-based pre-training as a primary factor in model development (Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016), and it inspired over a decade of computational-neuroscience research that has focused heavily on the study of pre-trained DNNs from the engineering field. This focus on task-based pre-training was a major shift from earlier modeling work, which had a much stronger theoretical emphasis on architectural and computational motifs (Oliva & Torralba, 2001; Portilla & Simoncelli, 2000; Riesenhuber & Poggio, 1999). We wonder if untrained computational models might have received more attention had we known ten years ago that AlexNet-level encoding performance in macaque IT could be reached by an untrained architecture. Looking forward, we suggest that architecture optimization in untrained or minimally trained networks is a promising future direction for exploring the inductive biases that may underlie biological vison.

The performance advantage that we observed for our convolutional architecture compared with other architectures contrasts with recent work showing that pre-trained DNNs with highly distinct architectures can obtain similar levels of encoding performance in visual cortex (Conwell et al., 2022). However, it is important to note that massive pre-training may be sufficient to overcome a lack of brain-aligned inductive biases in diverse network architectures, such as in vision transformers. In fact, previous interpretability studies of vision transformers suggest that they implicitly learn to attend to increasingly large regions across layers, thus resembling the increasing receptive fields of convolutional networks (Cordonnier et al., 2020; Dosovitskiy et al., 2021). Furthermore, even fully connected neural networks trained on translation-invariant data automatically develop convolution-like operations, characterized by local receptive fields and weight sharing (Ingrosso & Goldt, 2022). We suggest that the role of architecture can be better understood by focusing on settings in which the contribution of pre-training has been minimized.

Our results also suggest that previous work does not paint a fair picture of how far we can get in predicting visual cortex responses without gradient descent optimization. This is because many of the previously studied architectures were specifically designed to obtain strong benefits from deep learning. It is, thus, not surprising that when comparing trained and untrained versions of these architectures, there is a major advantage of training. Our work examines architectural variations that would be difficult to train through gradient descent due to the large dimensionality of their feature spaces. In doing so, we show that untrained CNNs can perform far better than prior research indicated.

Previous work in computer vision has examined the surprising effectiveness of feature representations in randomly initialized convolutional networks. Jarret et al (2009) showed that a single-layer, randomly initialized CNN competes with an end-to-end trained network in classifying objects from the Caltech 101 dataset. This finding led Saxe et al. (2011) to investigate why random weights performed so well in such a case. To this end, they showed that a combined convolution and pooling architecture can yield spatial frequency selectivity and translation invariance using only random filters. More recently, Cao & Wu (2021) showed that a random, untrained CNN has an inductive bias to localize objects. They suggest that this phenomenon may be driven by the nature of images, where the background is relatively texture-less compared to the foreground objects, increasing the chance that the background will be deactivated by ReLU. Studying the potential of untrained networks has also gained traction in the field of neuroscience. In an effort to test the effectiveness of untrained CNNs in predicting neural responses in the visual cortex, Baek et al. (2021) evaluated the performance of untrained AlexNet in explaining fMRI responses to faces and showed that face-selective units

arise within the network in the absence of supervised training (Baek et al., 2021). In a different study, to improve the biological plausibility of current encoding models, Gieger et al. (2022) investigates model performance under minimized supervised training. Using a biologically inspired CNN with recurrent connections, they showed that training only around 5% of model synapses results in an 80% match to the performance of the fully trained network at modeling macaque ventral stream (Geiger et al., 2022).

We observed a notable difference in the encoding performance of our untrained CNN in the monkey data compared with the human data. While the untrained CNN rivaled AlexNet in monkey V4 and IT, it had a substantial performance gap relative to AlexNet in the human data, and this performance gap increased at higher levels of the ventral stream. This difference between the monkey and human data may be caused by several factors, such as the nature of the stimuli and experimental tasks or representational differences across species, with human vision placing a stronger emphasis on semantic associations (Bonner & Epstein, 2021; Doerig et al., 2022; Wang et al., 2023). However, another possible explanation is the differential contribution of feedback processes in these two datasets. In the monkey electrophysiology study, the images were presented for 100 ms to elicit purely perceptual processes, and neural responses were taken from a small temporal window immediately after stimulus onset to measure feedforward activations. In contrast, in the human fMRI study, the images were shown for three seconds, and due to the low temporal resolution of fMRI, the responses reflect at least several seconds of cortical activity, which means that they have a much greater contribution from feedback and recurrent processes that may emphasize abstract semantic information.

Note that we are not proposing that the Expansion model developed here reproduces the architectural details of visual cortex. In other words, we are not arguing that the hierarchy of visual cortex performs random dimensionality expansion. Though note that other work has indeed argued for random expansion as a biological computation (Babadi & Sompolinsky, 2014; Cayco-Gajic & Silver, 2019). In our work, dimensionality expansion serves not as a proposed computational mechanism but rather as a means of discovering brain-relevant representations, just as others have used supervised pre-training and backpropagation. Intriguingly, the process of distilling brain-relevant representation from a random, high-dimensional feature space has potential parallels with the neural pruning that occurs in visual cortex over the course of development (Sakai, 2020). Thus, an exciting direction for future work is to consider whether pruning in high-dimensional feature spaces could be incorporated into computational models of learning in human vision. Another potential direction is to consider whether our architecture could be combined with a small amount of pre-training to yield a highly performant model that is trained in a data-efficient manner.

In sum, our results show that many aspects of visual cortex representation naturally emerge in high-dimensional convolutional networks without the need for massive pre-training. Thus, contrary to the view that many diverse architectures explain visual cortex representations equally well (Conwell et al., 2022; Storrs et al., 2021), our findings show that convolutional architectures are exceptional in their ability to predict cortical representations de novo.

METHODS

Neural network architectures

Convolutional architecture. Our convolutional network, referred to as the Expansion model in this paper, is constructed by repeating a block of operations across a hierarchy of layers. Our goal was to predict image-evoked responses in higher-level visual cortex, so we constructed a deep network with five layers. The same base architecture was used for all analyses, with changes made to the dimensionality of the final layer. Each layer of the model consists of the following operations: convolution, ReLU, and average pooling. In layer 1, we used pre-defined wavelets with a curvature parameter. The use of these curved wavelets allows the model to detect both low-level straight contours and mid-level curved contours, and it was inspired by previous work that used similar wavelets to model mid-level representations in the visual cortices of humans and macaques (Yue et al., 2014, 2020). We constructed these wavelets using 3 degrees of curvature and 12 orientations, resulting in 36 filters, which were applied separately to each RGB channel to yield 108 output channels. We note that model performance was not strongly contingent on the specific number of curvature degrees and orientations. Convolution layers 2-5 consist of randomly initialized filters constructed with kaiming uniform initialization in PyTorch. We used a fixed number of filters for layers 2-4 and a varied number of filters in layer 5. The maximum output dimensionality for layer 5 differed across the two neural datasets due to memory limitations during the model-fitting procedure; the smaller size of the macaque electrophysiology data enabled us to explore the effects of expansion up to 10^6 features, whereas with the larger fMRI data, we could only expand up to 10^5 features. Similarly, to avoid memory issues during principal component analysis, a model with 10^5 features was used for image classification. Further details about this architecture are listed in Table S1.

Fully connected architecture. We created a five-layer fully connected network in which the number of units in each layer matched the number of channels in the corresponding layer of the Expansion model. As in the Expansion model, the architecture was fixed for layers 1-4, and we varied the number of features in the final layer. Further details about this architecture are listed in Table S2.

Transformer architecture. We used an untrained Vision Transformer (ViT) based on the architectural specifications of the "vit_base_patch16_224" model from the timm Python library (*Pytorch Image Models (Timm) | Timmdocs*, n.d.). To expand the network's dimensionality in a manner consistent with the other architectures, we focus on the final encoder block. Specifically, we inserted a fully connected layer between blocks 11 and 12 to modify the

embedding dimensionality of the last block. We then assess the encoding performance based on the number of output features from this final encoder block.

Neural datasets

Human fMRI data were obtained from a publicly available dataset, called the Natural Scenes Dataset (NSD) (Allen et al., 2022), that contains fMRI responses to 73,000 natural images from the Microsoft Common Objects in Context (COCO) database (Lin et al., 2015). These fMRI data were collected at ultra-high-field strength with high spatial resolution (7T field strength, 1.6-s TR, 1.8mm^3 voxel size). Each image was shown for three seconds, and participants were instructed to maintain fixation. We used the NSD single-trial betas, preprocessed in 1.8-mm volume space, and denoised using the GLMdenoise technique (version 3; "betas_fithrf_GLMdenoise_RR") (Kay et al., 2013). We converted the betas to z-scores within each scanning session and computed the average betas for each stimulus across repetitions. Each participant viewed up to 10,000 images, of which 1,000 were shown to all participants and 9,000 were unique to each participant. We examined responses to all images shown to all 8 participants, and we focused on voxels from the ventral visual pathway using regions of interest defined in the NSD data as early, midventral, and ventral visual streams. These regions were predefined by the original authors as part of the "streams" ROIs.

Macaque electrophysiology data were obtained from a publicly available dataset containing single-unit recordings from two macaques implanted with arrays of electrodes in V4 and IT (Majaj et al., 2015). Neural responses were recorded from a total of 88 sites in V4 and 168 in IT. For each site, the average response was obtained between 70 and 170 ms after stimulus presentation. The stimuli in this dataset consisted of 3,200 artificially generated gray-scale images composed from 64 objects belonging to 8 categories. The objects were positioned on top of natural scenes at various locations, orientations, and scales. All images were seen by both animals.

Encoding models

We used linear regression with L2 regularization to map model activations to unit and voxel responses. For these analyses, the data were split into sets of training and testing images. In the human fMRI data, the training set for each participant consisted of all images that were uniquely seen by that participant, while the test set consisted of the images that were seen by all participants. In the monkey electrophysiology data, the training and test sets were selected randomly with an 80:20 split.

To identify the best mapping for each model, we first found the optimal L2 penalty for each participant and each region through leave-one-out cross validation on the training images. We tested 10 L2 penalty values logarithmically spaced using base 10, ranging from 1 to 10^{10}. We selected the optimal penalty that results in the best average score across all units or voxels, and we used the optimal penalty to estimate regression coefficients on the entire training set. The estimated coefficients were then applied to the model activations for the held-out test images to generate predicted neural responses.

To assess the accuracy of these predictions, we calculated the Pearson correlation coefficient (r value) of the predicted and actual responses for each unit and voxel. We computed bootstrap distributions of these accuracy scores over 1,000 iterations in which the test images were resampled with replacement. The encoding performance score for each region was obtained by taking the average of these distributions over all units or voxels and all participants. For the untrained networks, image features were extracted from the last layer and used as regressors. For AlexNet, the top-performing layer for each region was identified through cross-validation on the training images, following the same procedure used to determine the optimal L2 penalty.

For analyses involving PCA, we learned PCs of network activations on the training images, and we projected the activations for both the training and test images onto these PCs. We then performed encoding model analyses on these reduced network representations using the procedure described above.

Image classification

Image features were extracted from the last convolution layer of each model using the Places365-Standard dataset (Zhou et al., 2018), which contains 365 categories of scenes. To remove any effect of feature dimensionality on classification, we learned the first 1,000 PCs from each model on a random subset of 100 images from each category in the training set (for a total of 36500 images), and projected model features from the entire validation set onto these learned PCs. Each model's PCs were then used for classification on the validation set, which contains 100 images per category. Classification accuracy was determined through 5-fold cross-validation using sklearn's K-Fold cross-validator, which ensures a consistent sample distribution across each class. Each fold was split into an 80-20 distribution, resulting in 80 train images and 20 test images per class. Using sklearn's Logistic Regression classifier, images were classified into categories in a 365-way classification, with chance performance at 1/365 or 0.0027.

DATA AVAILABILITY

The Natural Scenes Dataset is available here: https://naturalscenesdataset.org/.
The monkey electrophysiology dataset is available as part of the Brain-score GitHub package: https://github.com/brain-score.
The Paces dataset is available here: http://places2.csail.mit.edu/download.html.

CODE AVAILABILITY

The Expansion model, as well as code for all analyses is available on GitHub: https://github.com/akazemian/untrained_models_of_visual_cortex.

REFERENCES

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv:1407.5104 [Cs, q-Bio]*. http://arxiv.org/abs/1407.5104

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Babadi, B., & Sompolinsky, H. (2014). Sparseness and Expansion in Sensory Representations. *Neuron*, *83*(5), 1213–1226. https://doi.org/10.1016/j.neuron.2014.07.035

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature Communications*, *12*(1), 7328. https://doi.org/10.1038/s41467-021-27606-9

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1), 4081. https://doi.org/10.1038/s41467-021-24368-2

Cao, R., & Yamins, D. (2021a). *Explanatory models in neuroscience: Part 1 -- taking mechanistic abstraction seriously* (arXiv:2104.01490). arXiv. https://doi.org/10.48550/arXiv.2104.01490

Cao, R., & Yamins, D. (2021b). *Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility* (arXiv:2104.01489). arXiv. https://doi.org/10.48550/arXiv.2104.01489

Cao, Y.-H., & Wu, J. (2021). *A Random CNN Sees Objects: One Inductive Bias of CNN and Its Applications* (arXiv:2106.09259). arXiv. https://doi.org/10.48550/arXiv.2106.09259

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do We Know What the Early Visual System Does? *Journal of Neuroscience*, *25*(46), 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005

Cayco-Gajic, N. A., & Silver, R. A. (2019). Re-evaluating Circuit Mechanisms Underlying Pattern Separation. *Neuron*, *101*(4), 584–602. https://doi.org/10.1016/j.neuron.2019.01.044

Conwell, C., Prince, J. S., Alvarez, G. A., & Konkle, T. (2022). *Large-Scale Benchmarking of Diverse Artificial Vision Models in Prediction of 7T Human Neuroimaging Data* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.03.28.485868

Cordonnier, J.-B., Loukas, A., & Jaggi, M. (2020). *On the Relationship between Self-Attention and Convolutional Layers* (arXiv:1911.03584). arXiv. http://arxiv.org/abs/1911.03584

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). *Semantic scene descriptions as an objective of human vision* (arXiv:2209.11737). arXiv. https://doi.org/10.48550/arXiv.2209.11737

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Elmoznino, E., & Bonner, M. F. (2024). High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, *20*(1), e1011792. https://doi.org/10.1371/journal.pcbi.1011792

Geiger, F., Schrimpf, M., Marques, T., & DiCarlo, J. J. (2022, January 28). *Wiring Up Vision: Minimizing Supervised Synaptic Updates Needed to Produce a Primate Ventral Stream | OpenReview*. https://openreview.net/forum?id=g1SzIRLQXMM

Ingrosso, A., & Goldt, S. (2022). Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, *119*(40), e2201854119. https://doi.org/10.1073/pnas.2201854119

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211. https://doi.org/10.1152/jn.1987.58.6.1187

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience*, *7*, 247. https://doi.org/10.3389/fnins.2013.00247

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. https://doi.org/10.1038/s41467-022-28091-4

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). *Microsoft COCO: Common Objects in Context* (arXiv:1405.0312). arXiv. https://doi.org/10.48550/arXiv.1405.0312

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, *35*(39), 13402–13418. https://doi.org/10.1523/JNEUROSCI.5181-14.2015

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of Physiology*, *283*, 53–77.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*(1), 49–70. https://doi.org/10.1023/A:1026553619983

Pytorch Image Models (timm) | timmdocs. Retrieved May 6, 2024, from https://timm.fast.ai/

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), Article 11. https://doi.org/10.1038/14819

Sakai, J. (2020). How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences*, *117*(28), 16096–16099. https://doi.org/10.1073/pnas.2010281117

Saxe, A., Nelli, S., & Summerfield, C. (2020). *If deep learning is the answer, then what is the question?* 26.

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience*, 1–21. https://doi.org/10.1162/jocn_a_01755

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12), 1415–1426. https://doi.org/10.1038/s42256-023-00753-y

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yue, X., Pourladian, I. S., Tootell, R. B. H., & Ungerleider, L. G. (2014). Curvature-processing network in macaque visual cortex. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), e3467–e3475. https://doi.org/10.1073/pnas.1412616111

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, *222*. https://doi.org/10.1016/j.neuroimage.2020.117295

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

SUPPLEMENTARY

Appendix A: Model Architectures

Table S1: Expansion model architecture. The number of filters in the last layer is varied to study the effect of feature expansion. H, W and C stand for height, width and number of channels respectively. In each layer, N, K, S and P stand for the number of convolution filters, filter size, stride and padding respectively.

Input shape		Layers	Output shape
(H=224, W=224, C=3)		Conv(N=36,K=15,S=1,P=0), ReLU, AvgPool(K=2,S=None)	(H=112, W=112, C=108)
(H=112, W=112, C=108)		Conv(N=1000,K=7,S=1,P=0), ReLU, AvgPool(K=2,S=1)	(H=53, W=53, C=1000)
(H=53, W=53, C=1000)		Conv(N=3000,K=5,S=1,P=0), ReLU, AvgPool(K=2,S=1)	(H=25, W=25, C=3000)
(H=25, W=25, C=3000)		Conv(N=5000,K=3,S=1,P=0), ReLU, AvgPool(K=2,S=1)	(H=12, W=12, C=5000)
(H=12, W=12, C=5000)		Conv(N=3000,K=3,S=1,P=0), ReLU, AvgPool(K=5,S=1)	(H=6, W=6, C=varied)
(H=6, W=6, C=varied)		Flatten()	(O=varied)

Table S2: Fully connected architecture. The number of filters in the last layer is varied for comparison against the Expansion model. I and O stand for input and output size, and N stands for the size of the linear layer.

Input shape	Layers	Output shape
(I=224*224*3)	Linear(N=36* 3), ReLU	(O=108)
(I=108)	Linear(N=1000), ReLU	(O=1000)
(I=1000)	Linear(N=3000), ReLU	(O=3000)
(I=3000)	Linear(N=5000), ReLU	(O=5000)
(I=5000)	Linear(N=varied), ReLU	(O=varied)

Appendix B: Supplementary Figures

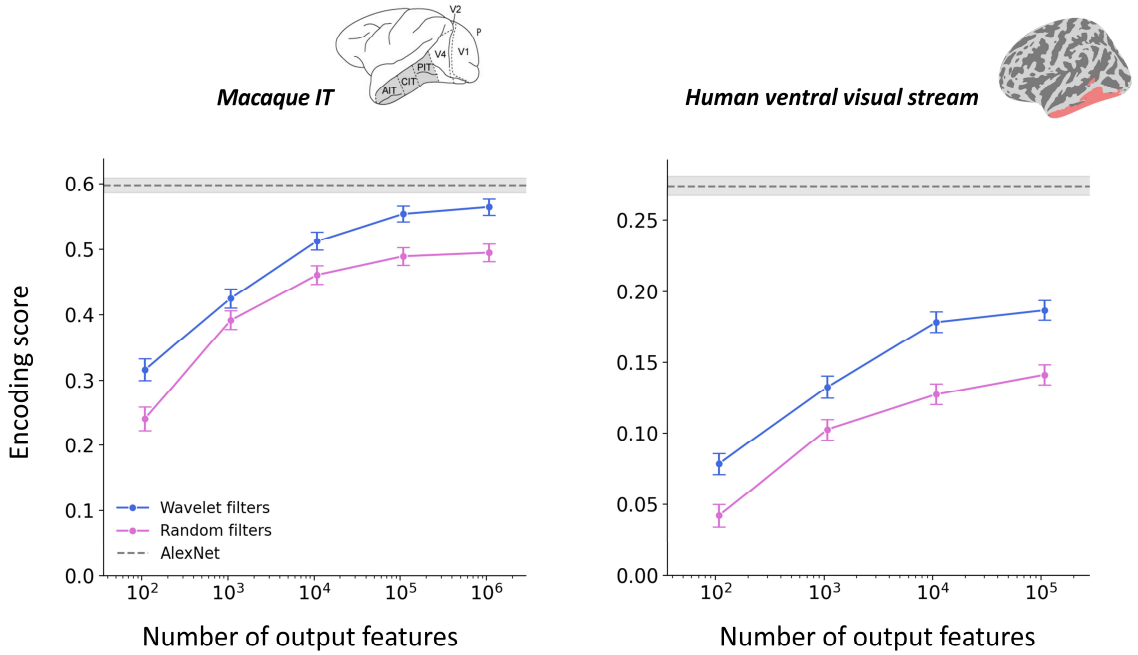

Figure S1. Effect of pre-defined wavelets in the first layer of the untrained convolutional neural network. These plots show the effect of using pre-defined wavelets in the first layer of the untrained convolutional network. Encoding performance is shown for macaque IT (left) and the human high-level ventral stream (right). For comparison, we examined a model with 3,000 randomly initialized filters in layer 1 (the number of random filters was maximized within computational memory limits). As in Figure 2, these plots show how encoding performance changes as a function of dimensionality expansion in the final convolutional layer. The x-axis plots the number of random features in the output layer, and the y-axis shows the encoding score for predicting image-evoked cortical responses. The gray dashed line indicates the performance of the best-performing convolutional layer of pre-trained AlexNet. There is a small, but consistent drop in performance for the fully random model across both datasets, demonstrating that overall, the network benefits from the implementation of pre-defined wavelets in its first layer.

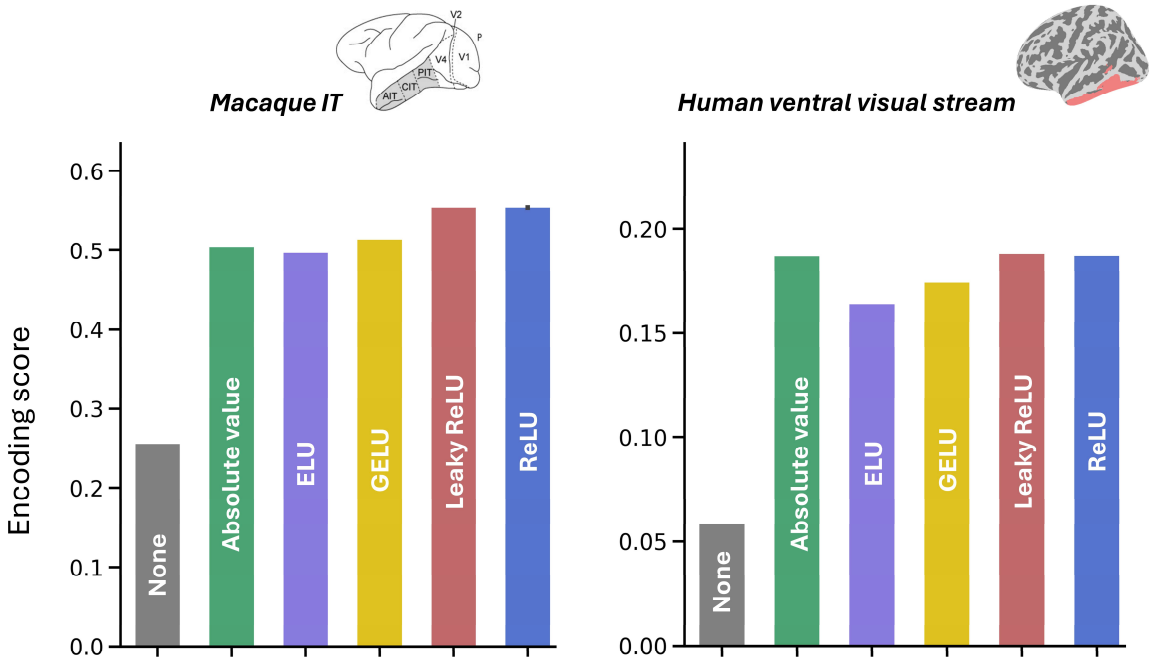

Figure S2. Different activation functions yield similar encoding performance for the untrained convolutional neural network. The effects of using different nonlinear activation functions in our untrained convolutional network were explored for monkey IT (left) and the human high-level ventral stream (right). These plots illustrate encoding performance for models with different nonlinear activation functions with otherwise identical architectures, all containing 10^5 features in their output layer. For comparison, this plot also includes a network without any nonlinear activation functions (the Linear model). The y-axis shows the encoding score for predicting image-evoked cortical responses. The results demonstrate that the inclusion of nonlinearities is critical, but various types of nonlinearities yield similar levels of performance. ReLU = rectified linear unit, GELU = Gaussian error linear unit, ELU = exponential linear unit

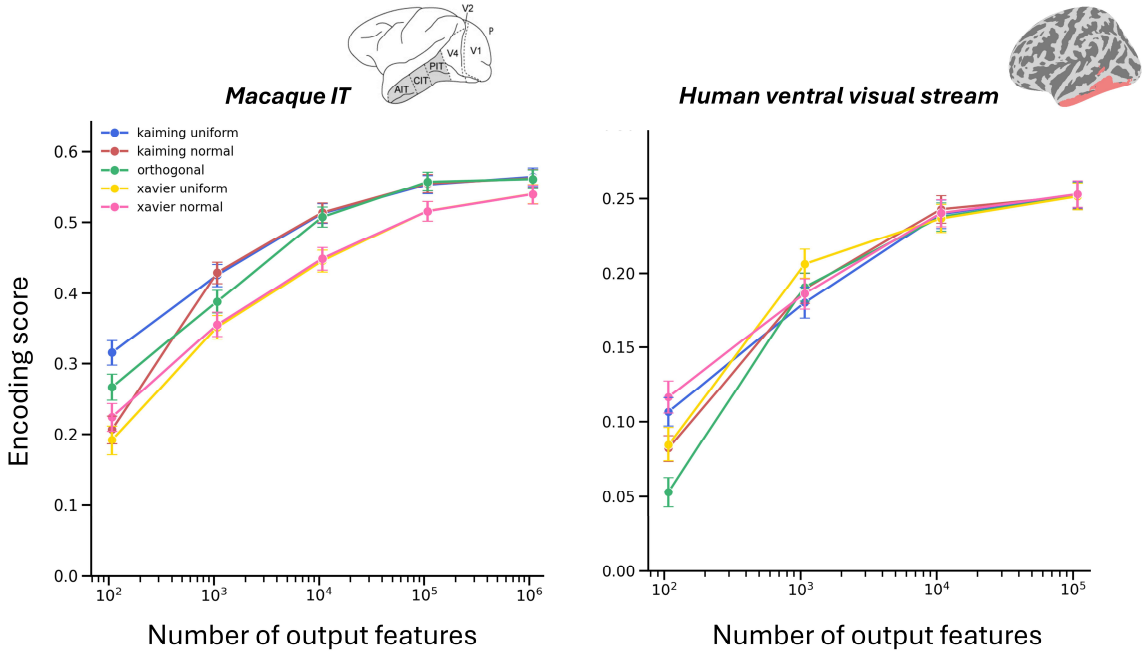

Figure S3. Different random initialization methods yield similar encoding performance for the untrained convolutional neural network. The effects of initializing the random features of our untrained convolutional network using different methods were explored for monkey IT (left) and human ventral visual stream (right). These plots illustrate encoding performance for identical architectures with different random initialization types. As in Figure 2, the x-axis plots the number of random features in the output layer, and the y-axis shows the encoding score for predicting image-evoked cortical responses. There is variation in encoding performance for different initialization methods in models with low dimensionality (on the left side of the x-axis). However, at higher levels of dimensionality, these performance differences diminish. This indicates that the type of initialization has minimal impact on encoding performance in the presence of model expansion.