Scaling Laws for Task-Optimized Models of the Primate Visual Ventral Stream

Abdulkadir Gokce (abdulkadir.gokce@epfl.ch)
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Martin Schrimpf (martin.schrimpf@epfl.ch)
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Abstract

When trained on sufficiently large object classification datasets, particular artificial neural network models provide a reasonable match to core object recognition (COR) behaviors and the underlying neural response patterns across the primate visual ventral stream (VVS). Recent findings in machine learning suggest that training larger models on larger datasets with more compute budget translates into improved task performance, but how scale affects brain alignment is currently unclear. We here investigate the scaling laws for modeling the primate VVS with respect to the compute-optimal allocation of dataset and model size across over 300 models trained in a controlled manner. To evaluate models' brain alignment, we use a set of benchmarks spanning the entire VVS and COR behavior. We find that while increasing the number of model parameters initially improves brain alignment, larger models eventually lead to diminishing returns. Increasing the dataset size consistently improves alignment empirically, but we extrapolate that scale here also flattens out for very large datasets. Combining our optimal compute budget allocation for model and data size into scaling laws we predict that scale alone will not lead to substantial gains in brain alignment with current architectures and datasets.

Keywords: primate visual ventral stream; neural alignment; behavioral alignment; scaling laws; computer vision

Introduction

Certain artificial neural networks are the current most precise quantitative models of the primate visual ventral stream (Yamins et al., 2014; Rajalingham et al., 2018; Schrimpf et al., 2018; Cadena et al., 2019; Schrimpf et al., 2020). These models are typically optimized on (labeled) image datasets with objectives such as object classification or self-supervised representation learning. After training, the best models produce outputs that resemble human behavioral choices, and internal activity that resembles the underlying neural response patterns in cortical regions V1, V2, V4, and IT – but all models currently fall well short of explaining all the data (Schrimpf et al., 2018).

In machine learning, recent performance advances are increasingly driven by larger volumes of training data and larger model architectures (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai, Kolesnikov, Houlsby, & Beyer, 2022; Bahri, Dyer, Kaplan, Lee, & Sharma, 2022). We here ask if *scale* is also a key factor for model similarity to neural and behavioral data, and if scaled-up neural network architectures optimized on scaled-up image datasets could vastly improve model alignment to the primate visual ventral stream. Specifically, we explore the impact of the scale of model parameters and available training data to describe the compute-optimal allocation of model and dataset size ("scaling laws", Figure 1).

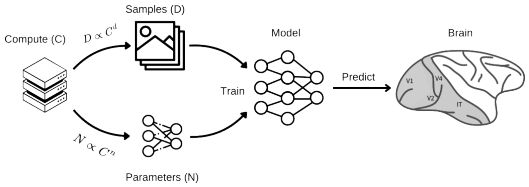

Figure 1: For a given compute budget (C), we determine the scaling laws for maximal neural and behavioral alignment to the primate visual ventral stream. We estimate the optimal allocation to training dataset ($D \propto C^d$) and model ($N \propto C^d$) as $d = 0.76$ and $n = 0.24$.

Methodology

We systematically train over 300 models from various architecture families on various image datasets. Since the current most brain-like models are trained on image classification, we focus on this task. We train models from the `ResNet` (He, Zhang, Ren, & Sun, 2016), `EfficientNet` (Tan & Le, 2019), `AlexNet` (Krizhevsky, Sutskever, & Hinton, 2012), and `CORnet-S` (Kubilius et al., 2019) families on versions of ImageNet (Deng et al., 2009) and ecoset (Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021). To independently manipulate model complexity and data volume, we train all neural networks with the same recipe; minimizing a cross-entropy loss for 100 epochs using an SGD optimizer with a decaying cosine learning rate (5 epochs of linear warm-up, reaching a peak learning rate of 0.1).

To assess the similarity to the primate visual ventral stream, we use five publicly available benchmarks from Brain-Score (Schrimpf et al., 2018, 2020) spanning V1, V2 (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013), V4, IT (Majaj, Hong, Solomon, & DiCarlo, 2015), and object recognition behavior (Rajalingham et al., 2018). These benchmarks test model alignment by presenting models with the same images that were shown to primate subjects, measuring internal activations or behavior. Neural alignment scores are computed with a linear predictivity metric – fitting a PLS linear regression (Yamins et al., 2014) from model activations to brain data for 90% of images and estimating similarity on the held-out 10% via Pearson correlation, cross-validated ten times. Behavioral alignment scores are computed with an i2n metric – testing if a model makes the same image-level mistakes as humans (Rajalingham et al., 2018). All scores from these benchmarks are adjusted to their respective noise ceilings, and we report the overall alignment score as the mean of all five sub-scores.

Following machine learning literature (Hoffmann et al., 2022), we capture scaling laws as parametric curves $S = \hat{S} - AX^{-\alpha}$ where S is the brain alignment score (plateauing at $\hat{S}$) as a function of the number of model parameters $X = N$ and training samples $X = D$. We then fit a function of the form $S = \hat{S} - \frac{A}{N^{\alpha}} - \frac{B}{D^{\beta}}$ to estimate the optimal allocation of compute budget to data and model scale. Curve parameters $(\hat{S}, A, B, \alpha, \beta)$ are learned using L-BFGS with Huber loss $(\delta = 10^{-3})$.

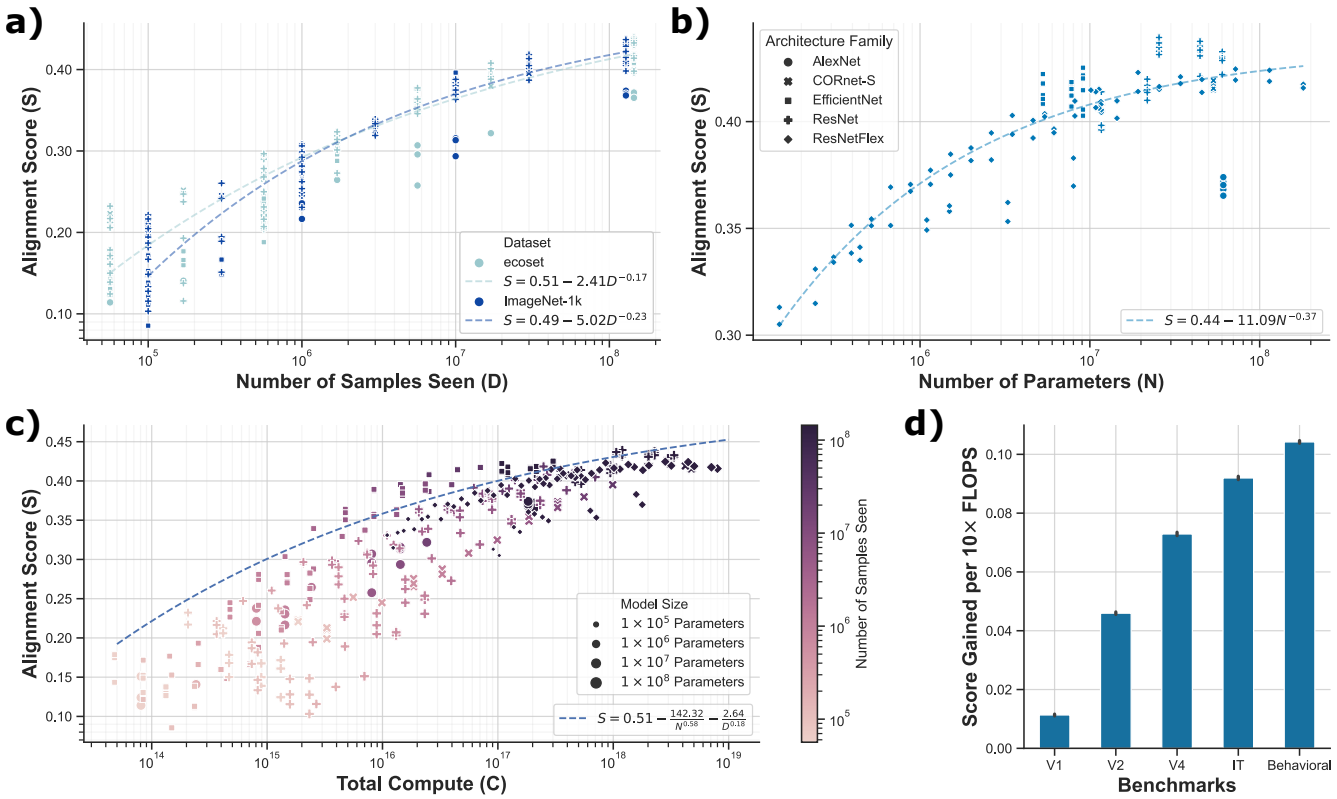

Figure 2: **Scaling laws for models of the primate visual ventral stream. a)** Training models with more data samples (D) improves alignment to brain and behavioral data (S). **b)** Training larger models improves alignment, but benefits saturate at $N \approx 10^8$ parameters. **c)** Combining dataset size (darker colors indicate larger datasets) and number of model parameters (larger markers indicate larger models) with respect to compute budget (C). Models following scaling laws $N \propto C^{0.24}$, $D \propto C^{0.76}$ yield optimal alignment scores while other allocations of compute result in poor model scores. **d)** Increases in compute scale (floating point operations per second, "FLOPs") most benefit model alignment to higher neural regions and behavior.

Results

Alignment scales logarithmically with dataset size. We generate subsets of ImageNet and ecoset by uniformly sampling d images per category where $d \in \{1, 10, 100\}$ (trained for 3 seeds) and $\{3, 30, 300\}$ (1 seed). Figure 2a shows the logarithmic dependency between model alignment and the number of dataset samples it has been trained on. This trend is highly comparable between the two base datasets and across model sizes and architectures. Our estimated scaling laws indicate that model alignment can be improved by training with more samples, but also predict a plateau at scores $\sim 25\%$ higher than the best model ($\hat{S} = 0.51$).

Alignment scales logarithmically with model size but diminishing returns. To evaluate the effect of model size, we compare architecture instantiations with different numbers of parameters (N). We cover low-N regimes by building `ResNetFlex` models with varying numbers of layers and widths relative to `ResNet-18`. Figure 2b shows model alignment for models with different numbers of parameters trained on full datasets. Although alignment increases with larger models, this trend saturates at $N \approx 10^8$ parameters. Architecture plays a vital role in parameter-efficient alignment: e.g., `AlexNet` models score well below similar-sized `ResNets`.

Optimal compute allocation favors data over parameters. Combining the number of training samples and model parameters, we establish the optimal allocation of compute budget to maximize the model alignment score (Figure 2c). We estimate that the allocation of compute to data and parameters should roughly follow $D \propto C^{0.76}$ and $N \propto C^{0.24}$. In other words, when increasing the compute budget 10-fold, the dataset size should be scaled by 5.7 and model parameters by 1.7. Again, the effect of scale saturates at $\hat{S} = 0.51$. Broken down into benchmarks (Figure 2d), we observe that scale most improves model alignment to later VVS regions and especially behavior.

Conclusion

The number of data samples and model parameters are key factors determining the brain and behavioral alignment of task-optimized models. We here establish compute-optimal scaling laws that favor samples over parameters and indicate that scale alone will be insufficient for perfect models of the primate visual ventral stream.

References

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2022). *Explaining scaling laws of neural network generalization.* Retrieved from https://openreview.net/forum?id=FvfV64rovnY

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019, April). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, *15*(4), e1006897. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.1006897 doi: 10.1371/journal.pcbi.1006897

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981. doi: 10.1038/nn.3402

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep Residual Learning for Image Recognition. In *Proceedings of 2016 ieee conference on computer vision and pattern recognition* (pp. 770–778). IEEE. Retrieved from http://ieeexplore.ieee.org/document/7780459 doi: 10.1109/CVPR.2016.90

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems.* Retrieved from https://openreview.net/forum?id=iBBcRUlOAPR

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). *Scaling laws for neural language models.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc.

Kubilius, J., Schrimpf, M., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Neural information processing systems (neurips)* (pp. 12785—-12796). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/9441-brain-like-object-recognition-with-high-performing-shallow-recurrent-anns

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. *The Journal of Neuroscience*, *35*(39), 13402–13418. doi: 10.1523/jneurosci.5181-14.2015

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8). doi: 10.1073/pnas.2011417118

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience*, *38*(33), 7255–7269. doi: 10.1523/jneurosci.0388-18.2018

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*. Retrieved from https://www.biorxiv.org/content/10.1101/407007v2

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*. Retrieved from https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X

Tan, M., & Le, Q. (2019, 09–15 Jun). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 6105–6114). PMLR. Retrieved from https://proceedings.mlr.press/v97/tan19a.html

Yamins, D. L., 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. doi: 10.1073/pnas.1403112111

Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022, June). Scaling vision transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 12104-12113).