What governs the emergence of brain-like specialized neurons in artificial neural networks?

Brian S. Robinson (brian.robinson@jhuapl.edu)
Johns Hopkins University Applied Physics Laboratory
Laurel, MD 20723, United States

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD 21218, United States

Abstract

Neurons with specialized properties have been widely characterized across the brain. It is a broad question as to why this occurs, with computational theories often assuming a central role of nonlinear neuron activation functions and connectivity constraints. In this work, we characterize neuron-specialization in artificial neural networks by extending regression-based approaches for predicting experimentally recorded neural activation patterns. When investigating a range of performant artificial neural network architectures, we demonstrate that (1) brain-aligned specialized neurons can emerge in layers *without* nonlinear neuron activation functions, and (2) the emergence of brain-aligned specialized neurons depends on training properties, not strictly on architecture. Overall, this work suggests that new and complementary explanations for the emergence of specialized neurons in biological brains may be needed, such as processes underlying learning and optimization. Furthermore, this work motivates brain-to-model comparison techniques that respect and further investigate properties of neuron specialization. These results may additionally inform general interpretability approaches for artificial neural networks, where methods for obtaining units for inspection is an active area of research.

Keywords: Neural Network Representations; fMRI; Privileged Basis; Neural Coding; Model Interpretability

Introduction

Neurons with specialized properties have been widely observed and characterized across the brain (e.g. neurons responding to specific complex visual features, place cells, etc.). At a fundamental level, there is a larger question of whether to consider neuron-specific properties at all or to to to look at a population level (Averbeck, Latham, & Pouget, 2006; Barack & Krakauer, 2021; Haxby et al., 2011; Gauthaman, Ménard, & Bonner, 2024). After all, the linearly decodable information and the relative distances between activation patterns can be preserved when linearly transforming from an original neural basis with a geometry-preserving rotation matrix. Indeed, this perspective dominates approaches to compare the activation patterns of artificial neural network models with the brain. However, in brains, there are additional biological constraints that violate the equivalence between different population-level encoding realizations, even if they share the same geometrical properties. A predominant constraint that can affect the specialization of neurons is their nonlinear activation, which impacts the energy efficiency of population pattern generation and transmission. In computational models, it is an active area of research to find how specialized neurons may emerge with a focus on hypotheses around nonlinear activation, such as sparse coding (Olshausen & Field, 1997), disentanglement of representations (Higgins et al., 2021; Soulos & Isik, 2024), wiring costs (Blauch, Behrmann, & Plaut, 2022; Margalit et al., 2024), and the importance of neuron activation nonlinearities (Khosla, Williams, McDermott, & Kanwisher, 2024). As a case study, we investigate population activation patterns in artificial neural networks in layers without output nonlinear activation functions. We demonstrate that (1) brain-aligned specialized neurons can emerge in layers *without* output nonlinear neuron activation functions, and (2) the emergence of brain-aligned specialized neurons is controlled by properties of training and not strictly architecture dependent.

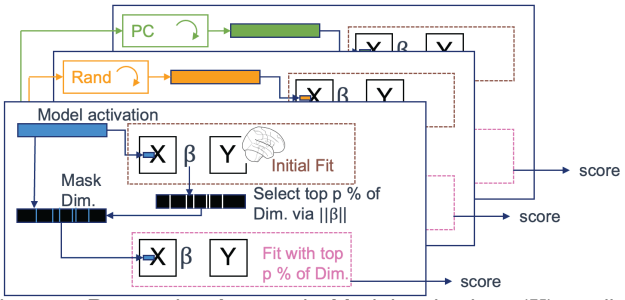

Figure 1: Regression Approach. Model activations (X) predict voxel-wise fMRI data (Y) via an initial regression yielding coefficients β. The L2 norm of β selects the top $p\%$ of model dimensions, used in a second regression. Cross-validated Pearson correlations are computed for original, randomly rotated, and PCA-rotated model activations.

Results

As a proof of principle, we investigate a selection of artificial neural network architectures, including transformers. We perform comparisons between model activation patterns and fMRI stimulus-evoked responses from a representative subject of the Natural Scenes Dataset (Allen et al., 2022). To characterize the neuron-specific coding properties in these networks, we compare brain similarity properties of activa-

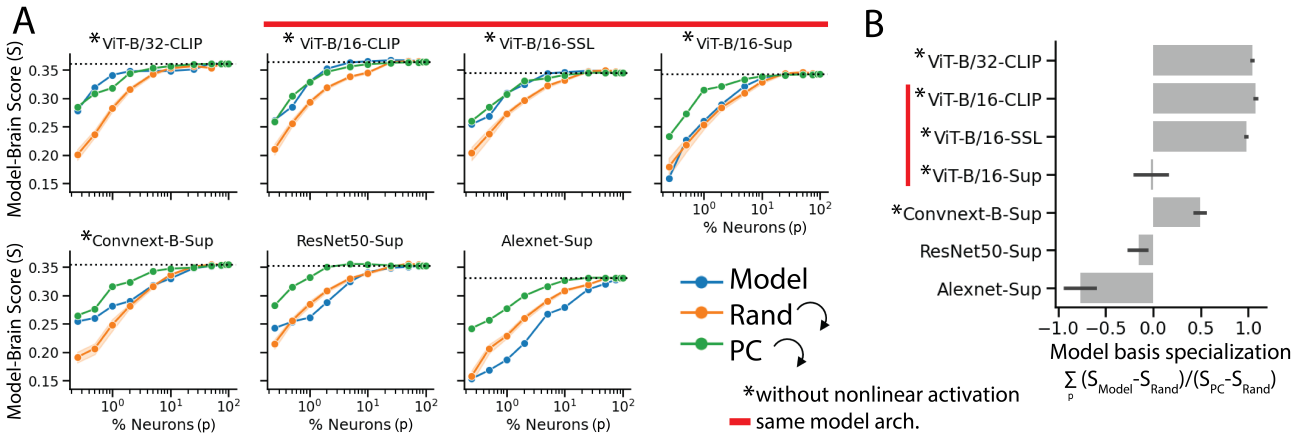

Figure 2: Neuron specialization occurs in models without nonlinear activations and varies between training method. A. Model-brain score curves are calculated when varying the top $p\%$ of model dimensions included. In many models, differences in score curves are observed between model and random bases. B. An aggregate neuron-specialization score for each model is computed by normalizing the difference between score curves of Model-Random by PC-Random. Training methods include supervised (Sup.), self-supervised learning (SSL), and contrastive language-image pretraining (CLIP). Error bars are bootstrapped from 10 random rotation samples.

tion patterns in the original model (model bases) with rotated versions of these activity patterns that are random (random bases) or based on principal components (PC bases). Note that traditional model to brain comparison methods such as representational similarity analysis and regression-based approaches are rotation-invariant and lead to the same brain similarity scores for these rotated models. However, for regression-based approaches (where model activations are used to predict brain activity and scored on held-out data), the relative weighting and contributions of each artificial neuron to the overall fit will be different. This relative weighting can be inspected via the learned coefficients, where with regularized regression, the magnitude of coefficients is a direct measure of the contribution of the artificial neuron to the similarity measure. By re-fitting the regression model while using only a fraction of the overall artificial neurons, we can measure how concentrated the brain-predicting activity is on a subset of specialized neurons (Fig. 1). To characterize the overall distribution of neuron-specialization, we refit the regression model for a range of percentages of neurons included. Neuron-specialization with a random rotation (sampled from the special orthogonal group) is a baseline that reflects how the specialization of a small subset of neurons can be attributed to chance. Alternatively, with the PC bases, the artificial neurons are explicitly computed to maximize the explained variance, where a higher neuron-specialization is expected.

We find that neuron-specialization occurs across some but not all models, and contrary to previous predictions, it is not dependent on including a nonlinear activation function (Fig. 2). The artificial neural network models span architecture classes of vision transformers (ViT-B-16, ViT-B-32) and convolutional neural networks (AlexNet, ResNet-50, ConvNext-B) and vary in training approach (supervised, self-supervised (Caron et al., 2021), and language alignment (Radford et al., 2021)). For a subset of transformer models, (which lack nonlinear neuron activations at layer outputs), we find a surprisingly high degree of neuron specialization where the neuron-specialization in the model basis can match the principal component basis (Fig. 2A). That is, the subset of neurons in the PC basis that are ranked to explain the overall signal variance are equal to the model bases (unlike random rotations). Additionally, we find that the emergence of brain-aligned specialized neurons depends on training properties, not strictly architecture (Fig. 2A-B). With the same ViT model architecture, training with language alignment and self-supervised approaches leads to enhanced neuron specialization compared to supervised training on ImageNet classification.

Discussion

In this work, we characterize neuron-specialization in artificial neural networks and find that brain-aligned neuron-specialization exists in model layer outputs without activation nonlinearities or other constraints such as wiring costs. These results suggest additional explanations for the emergence of specialized neurons in biological brains may be needed. More broadly, observations consistent with neuron specialization (or a "privileged basis") are observed between layers in transformer networks, such as certain dimensions with large activations (Kovaleva, Kulshreshtha, Rogers, & Rumshisky, 2021; Dettmers, Lewis, Belkada, & Zettlemoyer, 2022), but there is no consensus view on why this neuron specialization may emerge (Elhage, Lasenby, & Olah, 2023). A leading explanation is that utilizing an optimizer which stores momenta weight-wise in the model's neural basis (Elhage et al., 2023; He, Noci, Paliotta, Schlag, & Hofmann, 2024) may be the primary contributing factor. The possibility that neuron specialization can arise through weight-wise optimization processes in artificial neural networks offers a new and complementary explanation for how specialization might emerge in biological brains.

Acknowledgments

This research was supported by internal funding from the Johns Hopkins University Applied Physics Laboratory.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022, January). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, *7*(5), 358–366.

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, *119*(3), e2112566119.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, *35*, 30318–30332.

Elhage, N., Lasenby, R., & Olah, C. (2023). Privileged bases in the transformer residual stream. *Transformer Circuits Thread*, 24.

Gauthaman, R. M., Ménard, B., & Bonner, M. F. (2024). Universal scale-free representations in human visual cortex. *arXiv preprint arXiv:2409.06843*.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., ... Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416.

He, B., Noci, L., Paliotta, D., Schlag, I., & Hofmann, T. (2024). Understanding and minimising outlier features in transformer training. *Advances in Neural Information Processing Systems*, *37*, 83786–83846.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, *12*(1), 6456.

Khosla, M., Williams, A. H., McDermott, J., & Kanwisher, N. (2024). Privileged representational axes in biological and artificial neural networks. *bioRxiv*, 2024–06.

Kovaleva, O., Kulshreshtha, S., Rogers, A., & Rumshisky, A. (2021). Bert busters: Outlier dimensions that disrupt transformers. *arXiv preprint arXiv:2105.06990*.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, *112*(14), 2435–2451.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, *37*(23), 3311–3325.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Soulos, P., & Isik, L. (2024). Disentangled deep generative models reveal coding principles of the human face processing network. *PLOS Computational Biology*, *20*(2), e1011887.