A local unsupervised learning algorithm for building a visual hierarchy

Ananya Passi (apassi1@jh.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD 21218 United States of America

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

Abstract

Deep neural networks (DNNs) are the leading computational models of visual cortex, but they are trained using a biologically implausible mechanism that backpropagates a learning signal through the entire network hierarchy. We developed an approach for building a hierarchy of visual features using only local unsupervised learning, without the need for backpropagation. In our algorithm, each layer of a DNN contains a bottleneck in which representations are compressed and then expanded again. Learning is fully unsupervised, with each layer learning only to compress its inputs. This parsimonious algorithm yields representations that are competitive with conventional DNNs at predicting visual cortex representations up to intermediate layers. This work identifies a new approach for learning a visual hierarchy that is consistent with principles of learning in biology, requires no image labels or tasks, and may be sufficient to account for large fraction of visual cortex representations.

Keywords: convolutional neural networks; deep learning; fMRI; encoding models, visual cortex, vision

Introduction

The visual cortex extracts increasingly complex representations of visual input across a hierarchy of processing stages (Van Essen & Maunsell, 1983). DNNs provide the leading computational models for explaining these hierarchical visual representations (Conwell, Prince, Alvarez, & Konkle, 2022; Richards et al., 2019; Yamins & DiCarlo, 2016; Yamins et al., 2014). However, current DNNs rely on backpropagation during training—sending an error signal backward through the full network to update connection weights. This learning procedure is widely considered implausible for visual cortex, as there is no known biological mechanism for backpropagation. Thus, an alternative account is needed for how hierarchical visual representations could be learned in the brain.

We hypothesized that a hierarchical system could be built through local unsupervised learning rules operating within each processing stage, without requiring backpropagation of error signals across the full hierarchy. Specifically, we developed an algorithm where each layer of a DNN contains a bottleneck, compressing its input into a lower-dimensional representation before re-expanding it. We used a DNN trained with this algorithm to predict image representations in human visual cortex and compared the performance with a conventional backpropagation-trained DNN. We found that local unsupervised learning yielded visual representations that rivaled the performance of a conventional DNN up to intermediate layers of the network hierarchy. Our work identifies a novel framework for understanding how hierarchical visual representations could be learned through biologically plausible principles, without requiring training labels or tasks. This unsupervised deep learning approach may explain the development of early-to-intermediate level visual representations in the brain.

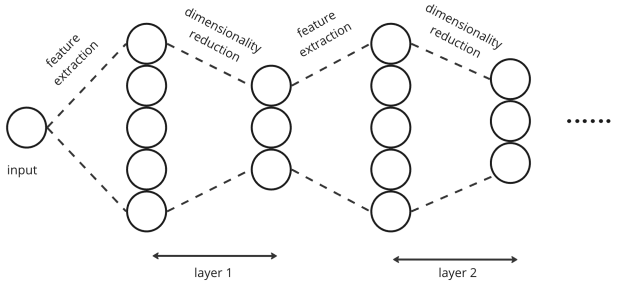

Figure 1: Model architecture consisting of sequential feature extractions and dimensionality reductions

Methods

We used a convolutional architecture in which spatial- and channel-mixing are factorized into separated operations (Guth, Ménard, Rochette, & Mallat, 2023). We used a fixed set of spatial-wavelet filters and learned the channel-mixing filters. Each layer learned the principal components of its input activations for one million images from the ImageNet training set (Krizhevsky, Sutskever, & Hinton, 2012). Thus, the weights of the channel-mixing filters corresponded to the eigenvectors of the first K principal components, and they were used to compress each layer's inputs onto its dominant modes of variance for natural images. The dimensionality of these compressed representations was then expanded again during the spatial convolution operation, which included a nonlinear activation function. This approach balances dimensionality compression and expansion, allowing the learning procedure to implemented sequentially in a deep hierarchy without the representations becoming overly compressed and low-dimensional.

We evaluated how well our model performed when predicting image-evoked cortical responses in human fMRI data from

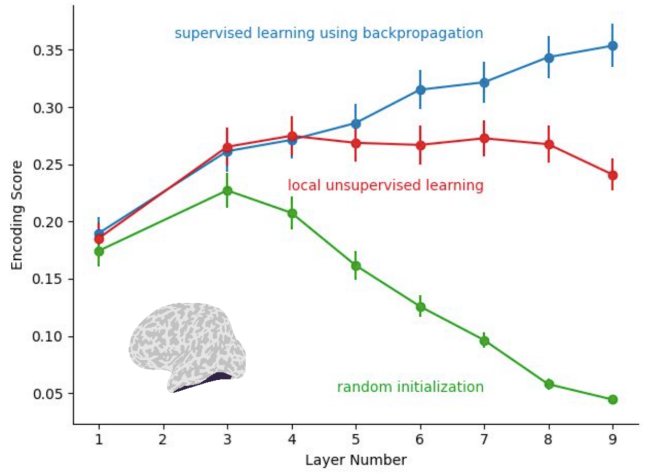

Figure 2: Mean voxel-wise encoding scores for supervised learning using backpropagation, local unsupervised learning and random initialization. The error bars denote the standard error across the encoding scores for five subjects. The brain rendering displays the region of interest - the ventral visual stream.

the ventral visual stream (Allen et al., 2022). To evaluate performance, feature vectors from different layers of our model were mapped to cortical responses using a regression procedure, which we validated on held-out test data. The encoding score of each model was obtained by measuring the correlation between the predicted and actual neural responses.

Figure 2 shows the mean voxel-wise encoding score of our model as a function of layer number. For comparison, we also show the performance of a conventional supervised model trained with backpropagation on ImageNet classification, and we show a randomly initialized model. All three of these models have the same architecture and differ only in how or if they were trained. The findings show that our local unsupervised model matches the performance of the conventional backprop up to the fourth layer, at which point these two models diverge. In contrast, the randomly initialized model performs substantially worse, and its performance drops dramatically at later layers. Together, these findings reveal the potential for using local, unsupervised learning to build a visual computational hierarchy that captures many aspects of representation in human visual cortex.

Conclusions

This work introduces an unsupervised algorithm for deep learning of visual representations without backpropagation. In contrast to existing methods that rely on image labels and tasks for supervised and self-supervised learning, our method uses only unsupervised learning for feature extraction in each layer. The approach presented paves the way toward a new deep learning framework for visual neuroscience that is grounded in parsimonious and biologically plausible theories of learning.

References

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

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. *BioRxiv*.

Guth, F., Ménard, B., Rochette, G., & Mallat, S. (2023). A rainbow in deep network black boxes. *arXiv preprint arXiv:2305.18512*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Van Essen, D. C., & Maunsell, J. H. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, *6*, 370–375.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

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.