Rapid unsupervised alignment with the natural image manifold

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

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 of America

Abstract

There is a stark contrast between the nature of feature learning in biological and artificial vision. While brains learn without explicit supervision and with little data, deep neural networks require supervised feedback and massive training sets. Here we show that a surprisingly simple unsupervised learning algorithm can yield large improvements in the brain alignment of a deep vision model. Specifically, we trained a network in which each layer learns to compress its representations onto the principal modes of variance for natural images—a form of local learning that does not require backpropagation or supervision. Using a relatively small sample of training images, this unsupervised learning algorithm strongly improves the network's ability to predict the image-evoked fMRI responses of visual cortex, and it makes downstream learning on an image-classification task more efficient. Remarkably, after an initial unsupervised-learning phase, the first half of the network's layers can be frozen with little impact on the ability to learn image classification. Together, these findings suggest that a parsimonious learning algorithm—operating locally and without supervision—may be sufficient to induce the features of early-to-mid-level vision and may accelerate the learning of downstream task-specific functions.

Keywords: convolutional neural networks; deep learning; fMRI; encoding models; random networks

Model Architecture

In order to simplify the pre-training process and reduce the number of learnable parameters, we deliberately chose an architecture that incorporates strong inductive biases, including convolution and band-pass filtering in all layers (Mallat, 1989). This design allowed us to isolate the contribution of unsupervised pre-training. We used a convolutional architecture in which spatial and channel-mixing are factored into separated operations (Guth et al., 2024). 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 100,000 images from the ImageNet training set (Krizhevsky et al., 2012). Through pre-training, we assigned the weights of the channel-mixing filters as the eigenvectors of the first K principal components, allowing each layer to compress its inputs onto the 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 be implemented sequentially in a deep hierarchy without the representations becoming overly compressed and low-dimensional. For comparison, we also examined an untrained model with random channel-mixing and a fully supervised model trained on ImageNet classification.

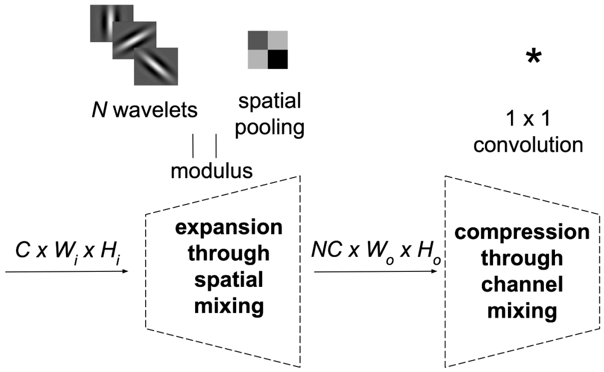
N wavelets
spatial pooling
| |
modulus
*
1 x 1 convolution
$C \times W_i \times H_i$
expansion through spatial mixing
$NC \times W_o \times H_o$
compression through channel mixing

Fig 1: Each layer of the deep vision model consists of expansion through spatial mixing followed by a learned compression through channel mixing. (*C: number of channels, W_i /H_i: width/ height of input feature maps, N: number of wavelet filters, W_o /H_o: width/ height of output feature maps*)

Methods and Results

To study brain alignment, we evaluated how well our model performed at predicting image-evoked cortical responses in human fMRI data from the ventral visual stream in the Natural Scenes Dataset (Allen et al., 2022). Feature vectors from the best performing layer 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 that our local unsupervised model matches the performance of conventional supervised learning up to intermediates layers. In contrast, the randomly initialized model performs substantially worse.

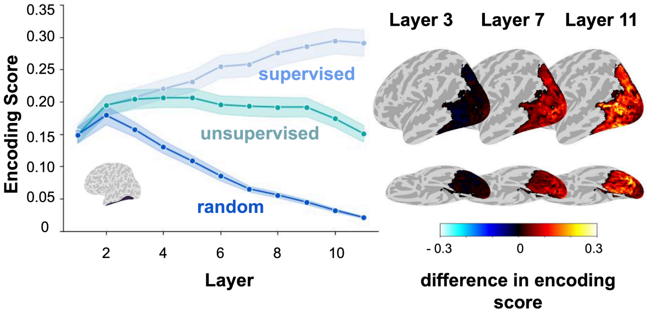

Fig 2: *(left)* Performance of random initialization, unsupervised and supervised learning. *(right)* Difference in performance between unsupervised learning and random initialisation.

To assess whether unsupervised initialization facilitates subsequent supervised learning, we performed supervised training on the mini-ImageNet image classification task (Vinyals et al., 2016). We evaluated both classification accuracy and encoding scores for models with and without unsupervised initialisation. Figure 3 illustrates the progression of these metrics over training epochs, revealing that models with unsupervised pre-training converge to higher accuracy and encoding scores more rapidly than a conventionally initialized network. This demonstrates that unsupervised pre-training can substantially enhance the efficiency of downstream supervised learning.

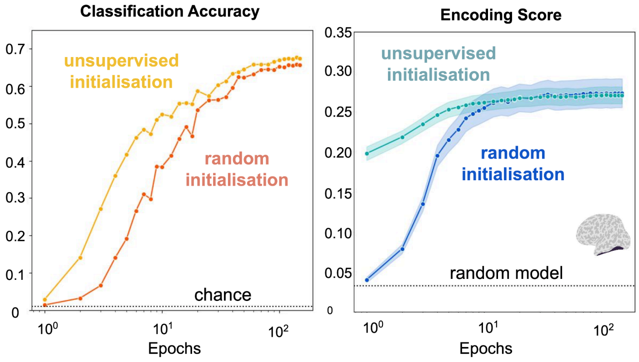

Fig 3: *(left)* Classification accuracy and *(right)* encoding score over training epochs in models with random initialization or unsupervised pre-training

To further examine how unsupervised pre-training influences early- to mid-level representations, we froze varying numbers of initial layers in models subjected to unsupervised pre-training and in randomly initialized models. We then applied supervised learning to the remaining unfrozen layers. As shown in Figure 4, models with unsupervised pre-training maintain robust classification accuracy and encoding scores—even when the first half of their layers are frozen—indicating that the learned representations in these early layers are sufficiently general and transferable to downstream tasks.

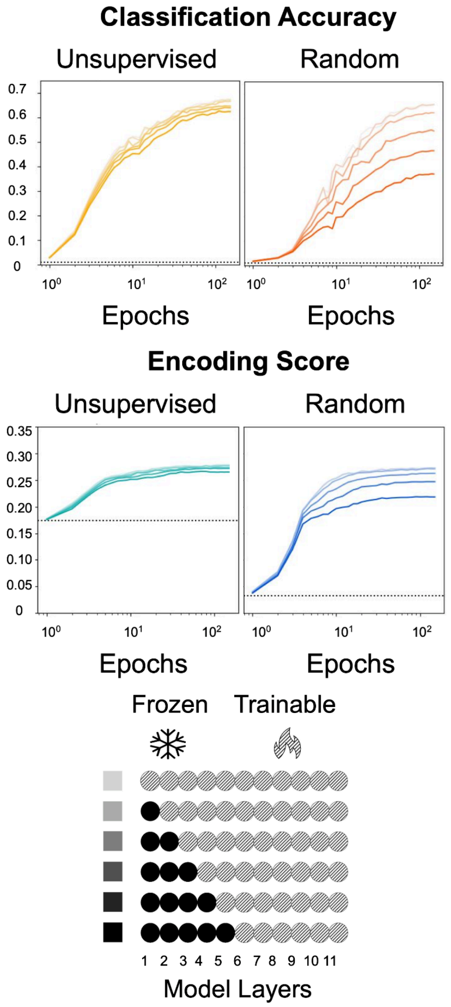

Fig 4: Classification accuracy and encoding scores across training epochs for models with varied frozen layers, in models with random initialization or unsupervised pre-training.

In sum, our findings show that a surprisingly simple unsupervised learning algorithm, which iteratively compresses and expands the representations in a deep hierarchy, enhances alignment with the human visual cortex and improves downstream task-learning efficiency.

References

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

Guth, F., Ménard, B., Rochette, G., & Mallat, S. (2024). *A Rainbow in Deep Network Black Boxes* (No. arXiv:2305.18512). arXiv. https://doi.org/10.48550/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*. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, koray, & Wierstra, D. (2016). Matching Networks for One Shot Learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html