fMRI Encoding of Visual Cortex with Contrastive Learning

Alex Mulrooney [1] Dr. Austin J. Brockmeier [2]

[1]UD ECE [2]UD ECE and CIS

Introduction

- Traditional approaches to predicting BOLD fMRI responses to images have utilized fixed feature maps such as category labels or features from a pretrained CNN as inputs to linear encoding models [1, 2]
- We propose a novel method utilizing a contrastive loss function that tunes a CNN such that its image representations are more similar to image representations in the brain, which we then use to predict fMRI responses to unseen images

Methodology

Dataset

- We use images and corresponding fMRI responses from the Natural Scenes Dataset, which contains BOLD responses to about 9,000-10,000 distinct images from the COCO database across 8 subjects [3, 4]

Contrastive Learning Network

For the set of n_{train} training trials and a region of interest (ROI) with v voxels, we have $X \in \mathbb{R}^{n_{\text{train}} \times 3 \times 224 \times 224}$ as the set of pre-processed training images and $Y \in \mathbb{R}^{n_{\text{train}} \times v}$ as the set of corresponding training fMRI responses for the v voxels in the ROI. The contrastive learning network consists of:

- $f_A(x) : x_i \in \mathbb{R}^{3\times224\times224} \mapsto a_i \in \mathbb{R}^{1000}$, an AlexNet architecture that has been pretrained on ImageNet
- $g_a(a)$ and $g_y(y)$, linear mappings that map the output a of $f_A(x)$ and y, respectively, to a shared latent representation $h \in \mathbb{R}^{v_1}$, $v_1 \leq v$
- $f_M(h) : h \in \mathbb{R}^{v_1} \mapsto z \in \mathbb{R}^{v_2}$, $v_2 \leq v_1$, a MLP projection head with three fully connected layers, with the first two layers each being followed by a batch normalization step and a ReLu activation function

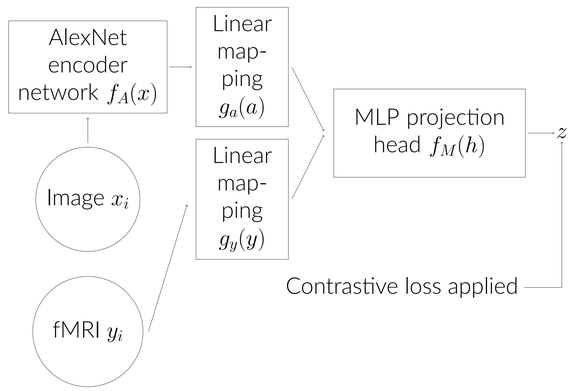


Figure 1. Contrastive Learning Network Architecture

A temperature-normalized contrastive loss function as in SimCLR is used to train $f_A(x), g_a(a), g_y(y)$, and $f_M(h)$ [5]:

$$l_i = -\log \frac{\exp\left(\text{sim}(z_i^x, z_i^y)/\tau\right)}{\sum_{j=1, j\neq i}^{n_{\text{batch}}} \exp\left(\text{sim}(z_i^x, z_j^y)/\tau\right)} \tag{1}$$

where $z_i^x and z_i^y$ are the outputs of the MLP projection head for x_i and y_i, respectively,τ is the temperature parameter, and n_{batch} is the batch size. The total contrastive loss for a batch computed from (1) as:

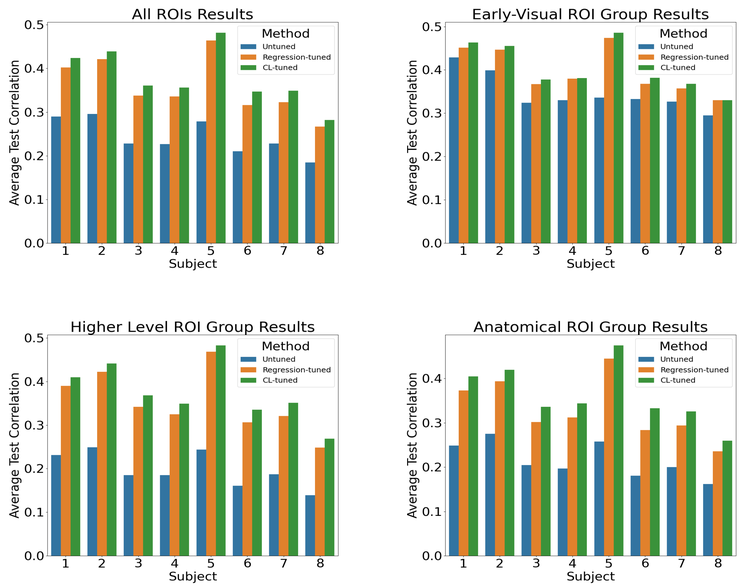
$$\mathcal{L}_{\text{batch}} = \sum_{n=1}^{n_{\text{batch}}} l_n \tag{2}$$

Linear encoding models

The features from the tuned AlexNet encoder $f_A\hat{(x)}$ from the CL network are compared against the features from the untuned AlexNet (control model) and the features from an AlexNet that has been tuned by regressing directly onto the voxels with a MSE loss (regression model). For each version of AlexNet, we fit a L_2 -penalized linear regression model to predict the activations of each voxel. We evaluate the accuracy of the predictions by calculating the mean Pearson correlation coefficient between the predicted and the true test data for each voxel in the ROI, and average across all voxels. We also compute the percentage of voxels in the ROI whose correlation is improved by CL versus the control and regression models.

Results

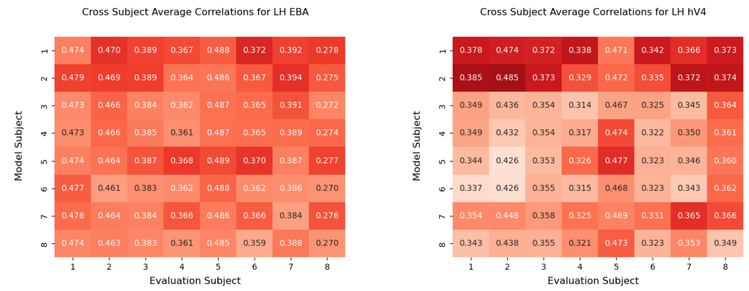
Subject-Specific Results



The preceding charts show the average test set correlation for each method across all subjects, broken up by groups of ROIs. Early-visual ROIs include V1, V2, V3, V4, and hV4, with the final metrics being an average of the performance on all of those ROIs. Higher level ROIs include body, face, place, and word selective regions such as EBA, FFA, RSC, and OWFA, respectively. Anatomical ROIs are divided by anatomical regions (early, ventral, parietal, etc.)

Cross-Subject Results

The heat maps below demonstrate the cross-subject encoding accuracies between subjects for two of the ROIs (left-EBA and left-hV4, where the element in row i and column j is the test set encoding accuracy for $\hat{f}_t^{i,j}$, the cross-subject linear encoding model that uses the tuned AlexNet features from subject i to predict the fMRI responses for subject j.



Discussion

- CL method produces significant improvements in encoding correlation across all ROIs and for each group of ROIs
- Increase in correlation for the tuned networks are greater for higher level ROIs than early-visual ROIs
- The learned representations for given ROIs are typically transferable across subjects, with minimal drops in average correlation for using a different subject's trained encoding network, or even better performance using a different subject's learned image features

Acknowledgments

This research was supported in part through the use of Information Technologies (IT) resources at the University of Delaware, specifically the high-performance computing resources.

References

[1] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[2] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, "Seeing it all: Convolutional network layers map the function of the human visual system," *NeuroImage*, vol. 152, pp. 184–194, 2017.

[3] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest *et al.*, "A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence," *Nature neuroscience*, vol. 25, no. 1, pp. 116–126, 2022.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.