fMRI Encoding of Visual Cortex with Contrastive Learning

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

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

UNIVERSITY OF DELAWARE

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^r 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 $Y^r \in \mathbb{R}^{n_{\text{train}} \times v_r}$ as the set of corresponding training fMRI responses for the v voxels in ROI r. The contrastive learning network consists of:

- $f_A^r(x) : x_i \in \mathbb{R}^{3\times224\times224} \mapsto a_i^r \in \mathbb{R}^{1000}$, an AlexNet architecture that has been pretrained on ImageNet
- $g_A^r(a)$ and $g_Y^r(y)$, linear mappings that map the output a^r of $f_A^r(x)$ and y^r, respectively, to a shared latent representation $h \in \mathbb{R}^{v_1}$, $v_1 \leq v_r$
- $g_H^r(h) : h \in \mathbb{R}^{v_1} \mapsto z \in \mathbb{R}^{v_2}$, $v_2 \leq v_1$, 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

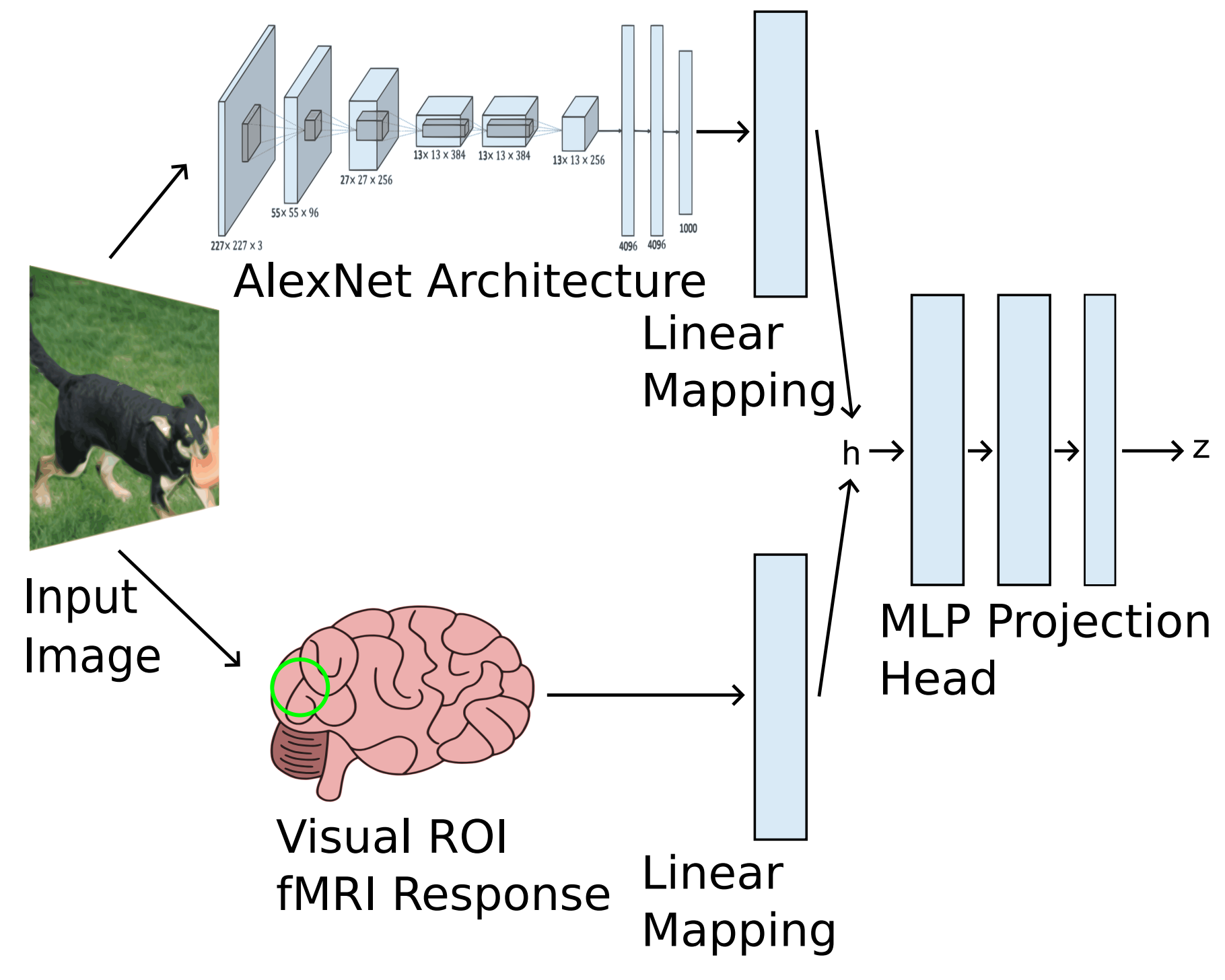


Figure 1. CL Model Architecture

A temperature-normalized contrastive loss function as in SimCLR is used to train $f_A^r(x), g_A^r(a), g_Y^r(y)$, and $g_H^r(h)$ [5]. The loss for a given trial i is:

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

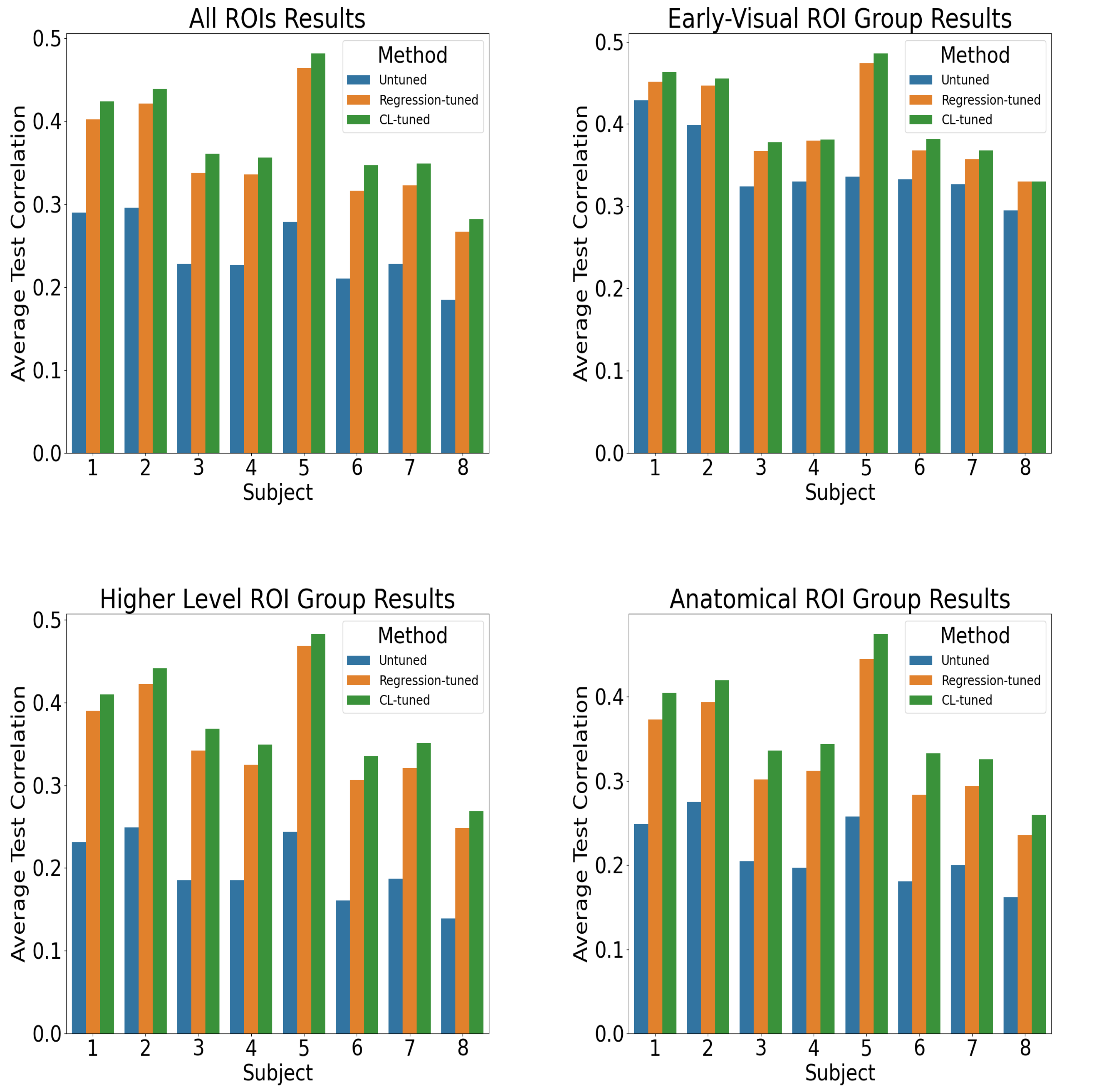
where z_i^x and z_i^y are the outputs of the MLP projection head for x_i and y_i^r, respectively, τ is the temperature parameter, and n_{batch} is the batch size. The losses are computed across mini-batches, with a batch size of 1024. Each trial i in the batch therefore has 1023 ”negative pairs” to contrast against the one positive pair.

Linear encoding models

The features from the tuned AlexNet encoder $\hat{f}_A(x)$ 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

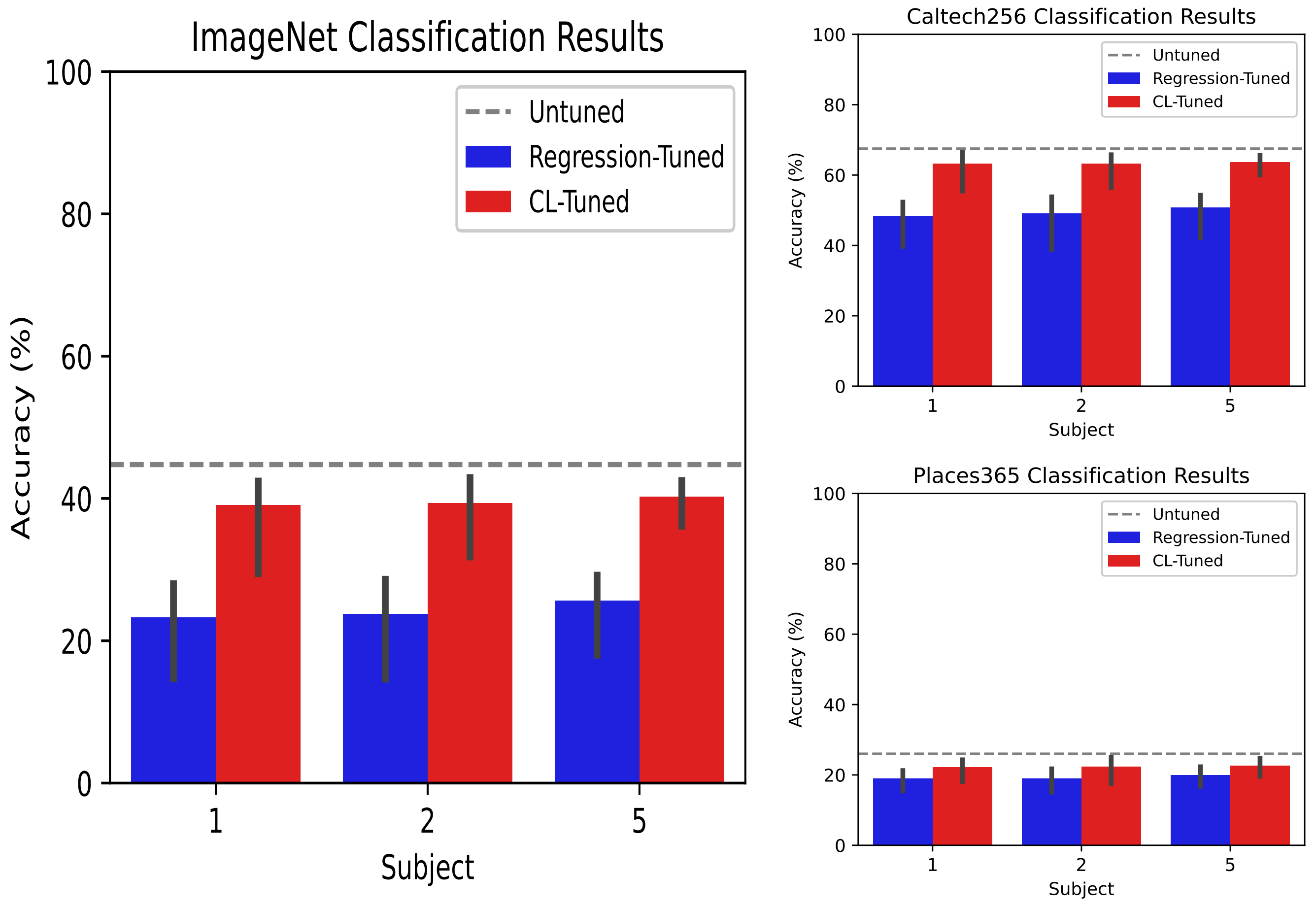


The above charts show the average test set correlation for each method across all subjects, broken up by groups of ROIs. Early-visual ROIs include V1, V2, V3, V4, and hV4, with the final metrics being an average of the performance on all of those ROIs. Higher level ROIs include body, face, place, and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

Image Classification Tasks

We assess the tuning of the AlexNet network using simple classification tasks with the ImageNet dataset (the original training task) and the Caltech256 and Places365 image classification datasets (not seen during training) [6, 7, 8].

We use the final 1000-node output of the AlexNet networks as feature representations of the images, fit a logistic regression model to predict the labels of the training images, and then evaluate the accuracy of the classifier on predicting the labels of the testing images given their AlexNet output features. We do this for the untuned AlexNet, the CL-tuned AlexNet, and the regression-tuned AlexNet for all ROIs in the 3 subjects with the highest encoding accuracies.



Discussion

- CL method produces significant improvements in encoding correlation across all ROIs and for each group of ROIs
- Increases in correlation for the tuned networks are greater for higher level ROIs than early-visual ROIs
- CL models experience minimal drops in accuracy on image classification datasets
- CL models perform moderately better than regression-tuned models on classification tasks

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.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[7] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” 2007.

[8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

Github: https://github.com/alexmul1114/fmri_encoding_contrastive_learning (in-progress)

Delaware Data Science Symposium 2023

alexmul@udel.edu, ajbrock@udel.edu