THESIS

NEURALATOR 5000: EXPLORING AND ENHANCING THE BOLD5000 FMRI DATASET TO IMPROVE THE ROBUSTNESS OF ARTIFICIAL NEURAL NETWORKS

Submitted by

William Augustus Pickard

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2023

Master's Committee:

Advisor: Nathaniel Blanchard

Chuck Anderson
Michael Thomas

ABSTRACT

NEURALATOR 5000: EXPLORING AND ENHANCING THE BOLD5000 FMRI DATASET TO IMPROVE THE ROBUSTNESS OF ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) originally drew their inspiration from biological constructs. Despite the rapid development of ANNs and their seeming divergence from their biological roots, research using representational similarity analysis (RSA) shows a connection between the internal representations of artificial and biological neural networks. To further investigate this connection, human subject functional magnetic resonance imaging (fMRI) studies using stimuli drawn from common ANN training datasets are being compiled. One such dataset is the BOLD5000, which is composed of fMRI data from four subjects who were presented with stimuli selected from the ImageNet, Common Objects in Context (COCO), and Scene UNderstanding (SUN) datasets. An important area where this data can be fruitful is in improving ANN model robustness. This work seeks to enhance the BOLD5000 dataset and make it more accessible for future ANN research by re-segmenting the data from the second release of the BOLD5000 into new ROIs using the vcAtlas and visfAtlas visual cortex atlases, generating representational dissimilarity matrices (RDMs) for all ROIs, and providing a new, biologically-inspired set of supercategory labels specific to the ImageNet dataset. To demonstrate the utility of these new BOLD5000 derivatives, I compare human fMRI data to RDMs derived from the activations of four prominent vision ANNs: AlexNet, ResNet-50, MobileNetV2, and EfficientNet B0. The results of this analysis show that the old, less-advanced AlexNet has a higher neuro-similarity than the much more recent, and technically better-performing models. These results are further confirmed through the use of Fiedler vector analysis on the RDMs, which shows a reduction in the separability of the internal representations of the biologically inspired supercategories.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. Nathaniel Blanchard for his mentorship and support during this project. I would also like to thank the members of my thesis committee, Prof. Chuck Anderson and Prof. Michael Thomas.

Various versions and sections of this work have been submitted as separate works for peer review, including a journal article that is still under review at the time of submission. I would like to thank my co-authors who have contributed to the project and its constituent papers including Kelsey Sikes, for her editorial contributions and figure design, Huma Jamil, for her expertise and code contributions in Fiedler vector analysis, and Mark Hinds for his work on the project.

DEDICATION

To my wife Sara.

Thank you for being by my side in our journey together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In the beginning, artificial neural networks designed for machine vision tasks drew their inspiration directly from the behavior of biological neurons [1]. Today, the link between the design of cutting-edge machine vision models and neurobiology is tenuous at best. Although researchers may no longer be looking to the natural world for their inspiration, that does not mean the connection between the artificial and the biological ends there.

Investigations into similarities between artificial neural networks and biological brains have been unfolding since the early days of the neural network boom [2–5]. Using representational similarity analysis (RSA) [6], researchers have shown that modern artificial neural networks (ANNs) develop internal geometric representations to visual stimuli similar to the neuro-biological activations seen in the brains of primates and other mammals [7–17].

It has been demonstrated that ANN models with a greater neuro-similarity to the mammalian brain perform better at some tasks than models with lesser neuro-similarity. Blanchard et al. demonstrated that unsupervised predictive coding networks — a form of DNN composed of convolutional long short-term memory (LSTM) units — with greater neuro-similarity to functional magnetic resonance imaging (fMRI) brain scans of human subjects performed better in next-frame prediction and object matching tasks [18]. Li et al. were able to improve the robustness of a deep convolutional neural network (DCNN) performing an image classification task to injected image noise by modifying the model training with an additional loss function that favored greater neuro-similarity to a dataset derived from two-photon excitation microscopy (2PEF) of mice brains [19]. These new advances show the utility of leveraging neuro-biologic data for use in machine learning applications.

Following the methodology pioneered by Jamil et al. [20], we demonstrate that network representations drift further away from biological representations when networks are optimized for task performance. We posit that our findings mirror critiques of prominent research groups like

Google's DeepMind — [21] identified the viability of an adversarial typographic attack were simply writing the incorrect word on object sufficed for causing misclassifications (see Figure 1.1). In a blog post discussing the attack, [21] suggested:

> "this attack exploits the way image classification tasks are constructed. While images may contain several items, only one target label is considered true, and thus the network must learn to detect the most 'salient' item in the frame. The adversarial patch exploits this feature by producing inputs much more salient than objects in the real world. Thus, when attacking object detection or image segmentation models, we expect a targeted toaster patch to be classified as a toaster, and not to affect other portions of the image."

Figure 1.1: An adversarial typographic attack where the name of an incorrect class is written on an object. The attack causes a classifier to misclassify the apple with a confidence of 99.7%. Image source: [21].

Prior work has shown that representations closer to the biological brain are more robust to adversarial attacks [19], are adaptable to new tasks in a zero-shot context [13, 18], and have gains in task performance that emerge quicker than when learning representations without biological

similarity [18,22]. Given this pedigree, one would be remiss not to wonder why research into these comparisons is so rare. Unfortunately, most biological datasets are proprietary or too small [23] and without this resource, neither researchers nor practitioners can further investigate this phenomenon. Further, state-of-the-art models have traditionally been assessed by their accuracy on key datasets while evaluations of how well-embedded representations generalize to new tasks is a relatively recent phenomenon [24].

Indeed, assessing state-of-the-art models has always been important for both practitioners adapting those models to their own tasks and researchers seeking to understand and push towards better models [25]; however, the advent of works like CLIP, from [24], have ushered in a new era driven by evaluating neural networks on how adaptable their learned representations are to new tasks in a zero-shot context. This work provides the tools for researchers to take this idea further providing biologically viable target representations that can be factored into the optimization of networks.

Additionally, many of these works simply focus on post-hoc evaluations. There are relatively few works investigating how to optimize networks to achieve biological representations [26–30]. Even recent efforts to learn strong representations focus on unsupervised methods that allow massive amounts of data to be used for training, with the hope that stronger representations will emerge [24]. We hypothesize that a large biological dataset would facilitate a deeper investigation into the viability of biological representations for artificial neural networks. Of particular interest to this community is the potential for deeper investigations into how to optimize for biologically grounded manifolds.

1.1 Introduction to the BOLD5000

To this end, human fMRI brain scan datasets are being collected using stimuli images pulled from existing machine vision image datasets to make using fMRI data in machine learning research both easier and more fruitful [31,32]. Of particular interest, and the subject of the work performed in this paper, is the BOLD5000 [23].

BOLD5000, one of the largest, publicly available fMRI datasets, was created to address three areas of neural dataset design: 1) create a dataset of sufficient size to enable fine-tuning an ANN, 2) have a greater diversity of images and image categories than is normally present in a neural study, and 3) provide an overlap between the stimulus images used in the fMRI trials and the training image datasets of ANNs to allow for a more direct comparison of ANNs and human brain activation's [23].

The BOLD5000 is composed of stimuli images pulled from existing machine vision image datasets [31,32]. In total, it consists of fMRI brain scans from four participants (CSI1-4) presented with 5,000 real-world images from three commonly used computer vision datasets: 1,916 from ImageNet [33], 2,000 from Common Objects in Context (COCO) [34], and 1,000 custom images of scenes from categories inspired by Scene UNderstanding (SUN) [35]. Collectively, these datasets span a wide variety of categories and consist of images of real-world indoor and outdoor scenes, and objects either centered in or interacting with complex, real-world scenes.

To collect this brain data, all selected images were resized, cropped to 375 $\times$ 375, and adjusted for even luminance. For each input dataset, exemplar images were hand-selected by the BOLD5000 authors on a per-category basis. Subjects then engaged in 15 functional MRI sessions, where all images were presented on a single trial basis, except for a subset of 113, for which unique neural representation data was collected. During this, one participant (CSI4) did not complete the entire experiment. As a result, their data is typically discarded from studies using the BOLD5000 [17]. However, in this work, we incorporate their data into our final preprocessed work.

Recently, a second revision of the BOLD5000 data has been made available that uses the GLMSingle toolbox to enhance the reliability of the beta estimates [36]. However, this second release only contains raw voxel beta values for the whole brain of each participant. The data has not yet been broken out into functional regions nor has any further analysis, such as RSA, been made publicly available that may aid in further machine learning research. As illustrated in Figure 1.2 this work seeks to remedy this by augmenting the BOLD5000 dataset in the following ways:

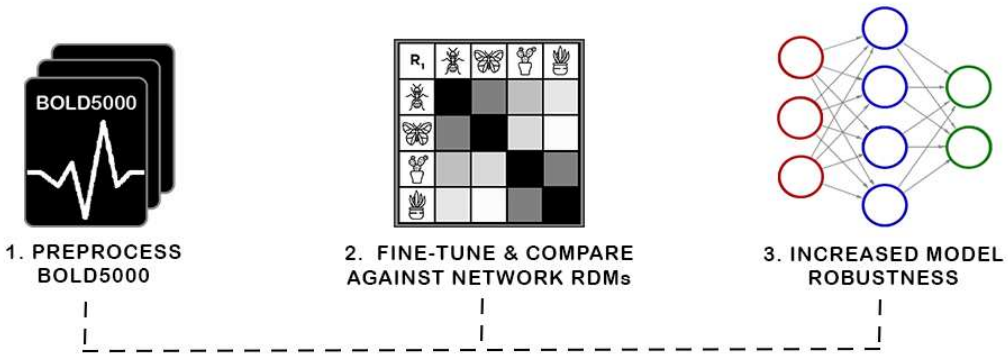

Figure 1.2: This work presents a new, biologically grounded representation for the evaluation and optimization of neural representations. Prior work has shown such representations correspond with robustness to adversarial attacks and task generalization. The curation of this new benchmark required preprocessing the BOLD5000 data into representational dissimilarity matrices (RDMs) and establishing a framework for investigating biological representations. The viability of the discovered representation was investigated with a novel application of Fiedler partitioning on the data to demonstrate the potential of the biological representation for adversarial robustness.

- Split the raw voxel beta values into functional regions of interest (ROIs) using the masks from the original release
- Re-segment each subject's cortical surface using two new brain atlases that map the ventral visual stream and other functional areas
- Calculate pre-computed representational dissimilarity matrices (RDMs) for each subject
- Leverage metadata available from the input datasets to demonstrate how additional insights can be gleaned using an interdisciplinary approach
- Apply a previously unexplored graph-based technique, the Fiedler algorithm, to this preprocessed dataset, demonstrating its versatility as an evaluation metric
- Introduce a framework that allows researchers to fine-tune, evaluate, and select models for robustness.

Ultimately, the products of this work will facilitate future research into how robust representations manifest and methods for optimizing networks to achieve trustworthy and adversarially robust results.

Chapter 2

Literature Review

2.1 Neuro-Similarity

Here, we detail prior works that investigate biological representation benchmarks. In particular, we focus on methods that investigate "neuro-similarity," i.e., the similarity of an artificial neural network's (ANN) learned representation to a benchmark of the biological brain. First, we examine metrics of neuro-similarity, then, efforts to increase neuro-similarity, and conclude with an investigation of works that link biologically consistent ANNs and robustness.

2.1.1 Metrics of Neuro-Similarity

Most works that evaluate ANNs for neuro-similarity utilize methods from representational similarity analysis (RSA). In particular, researchers derive metrics from representational dissimilarity matrices (RDMs) — either an ANN or neural data can be abstracted into an RDM for a set of stimuli. If two RDMs are created using the same stimuli set, they can be directly compared to one another by measuring the similarity of the consistency across that stimuli set. Two established metrics that capitalize on RSA for measuring the neuro-similarity of ANNs are human-model similarity (HMS) [18] and the Brain Score [13].

HMS [18] evaluates the neuro-similarity between fMRI data and ANNs as the Spearman correlation between the averaged fMRI RDM and an ANN's RDM. They validated their metric on self-supervised predictive coding networks — a form of ANN composed of convolutional long short-term memory (LSTM) units designed to mimic predictive coding employed by biological visual systems. They found that models with higher HMS exhibited higher performance on next-frame prediction (the self-supervised task the networks were trained on) and were more robust to other tasks that networks were not trained for, such as object matching. They also found that

HMS could be accurately measured early in the training process, and they proposed that it could be utilized for "early stopping" i.e., training could be abandoned before the weights fully converged.

Similar to HMS, Brain-Score [13] is a composite neural and behavioral benchmark set, which uses multiple evaluation metrics to score and rank ANNs according to how brain-like their visual object-recognition mechanisms are. To accomplish this, the internal representations of ANNs trained on ImageNet were compared for similarity against neural recordings taken from the V4 and IT cortical areas of macaque monkeys. From this, Dense-Net169, COREnet-S, and ResNet-101 were found to be the most brain-like, though Brain-Score was unable to reveal why.

HMS is the most similar to our methodology since we too use publicly available fMRI data, but a major limitation of HMS is that it only utilizes 92 stimuli, making it unsuitable to train with since networks quickly overfit to the small sample. These metrics are a great starting point for measuring neural similarity — however, to improve model robustness, more specific metrics need to be created. To effectively achieve this, datasets similar to this one must have as little noise in them as possible, something we address with BOLD5000.

2.1.2 Increasing Neuro-Similarity

The methods utilized for increasing the neuro-similarity of a DCNN can be split into two broad categories: the tailoring of image training datasets to achieve a distribution of input stimuli that more closely matches what may be experienced in nature [32, 37–39], and directly influencing the training of DNNs through the addition of a loss function that penalizes neuro-dissimilarity.

The former method has several examples in recent literature, and the approach is based on observations that training datasets designed for machine vision applications are crafted for domain-specific applications, or otherwise contain internal biases in their distribution of subject matter that do not match what is in nature [40]. A specific example of such a bias is the fact that ImageNet [33], one of the most widely used image classification datasets in the field, contains 120 categories of dog breeds but lacks any categories for humans. By creating datasets with more natural image

distributions, researchers have been able to significantly improve the neuro-similarity of the DNNs trained on these datasets.

While this approach does improve neuro-similarity in the trained models and demonstrates the potential of DNNs achieving higher levels of neuro-similarity, it may not always be feasible or desirable to augment every dataset with a great enough volume of images, or images of the correct type, to achieve a distribution that matches the natural world. For example, domain-specific datasets, such as for medical imaging research, don't have a complementary input set in nature to draw from. Datasets for machine vision research are also growing in size constantly and it may not be cost-effective or efficient to increase their size to a point where a natural distribution is achieved. However, these domain-specific models can potentially still benefit from greater neuro-similarity.

It has been demonstrated that DNN models with a greater neuro-similarity perform better at some tasks than models with lesser neuro-similarity. One exciting example of this, and the inspiration for this paper, was work done by [19], who improved the robustness of a deep convolutional neural network (DCNN) to image noise via fine-tuning with an additional loss function that favored greater neuro-similarity. These experiments were conducted using a dataset derived from two-photon excitation microscopy (2PEF) of mice brains — they released the code to enable the fine-tuning but did not release the data itself. The fine-tuning was enabled via RDM comparisons — however, unlike Brain Score and HMS, they approximated complete RDMs during training by only creating an RDM for a subset of stimuli. Constructing an entire RDM during training is computationally expensive because activations for each of the stimuli must be collected and compared.

2.1.3 Linking Neuro-Similarity to robustness

Despite [19] initial findings that improving neuro-similarity could increase robustness, none of the known evaluation metrics explicitly measure this improvement. We think this is an area where some could be created. We propose that robustness should be measured via Psychophysics [41,41]. This evaluation focuses on evaluating robustness across a range of different noise levels. It also focuses on explainable and trustworthy evaluations of networks — by exploring a multitude of

different noise types, the evaluation reveals specific weaknesses that networks are susceptible to e.g., in the domain of face recognition, [41] found that FaceNet was surprisingly susceptible to brown noise, while other methods were not.

Research like this shows the utility of leveraging neuro-similarity for use in machine learning applications and illustrates why more datasets, accessible to researchers outside of neuroscience, are needed. Further, the scarcity of evaluation metrics to apply to datasets like this means a notable research gap exists, worthy of future work.

2.2 Ventral Visual Stream

The ventral visual stream is a series of hierarchical cortical regions in the primate brain responsible for object recognition that has been termed the "what" pathway of the brain's vision center [42]. In humans, it extends from the primary visual cortex (V1), which is responsible for low-level visual features, to the inferior temporal (IT) cortex, which contains high-dimensional representations of object shape and category [43]. Due to its importance in object recognition, the ventral visual stream is the subject of many of the recent neuro-similarity papers, either as a whole [10, 14, 16], or for specific regions within the ventral stream such as V1 [44], or the IT cortex [7, 9, 15]. Due to the importance of the ventral visual stream in neuro-similarity research, and because the original BOLD5000 dataset does not include ROIs derived from the ventral visual stream, vcAtlas was selected as a way to add this context to the dataset.

Chapter 3

Materials and Methods

This work contains a synthesis of data from multiple sources and multiple fields of study. This section briefly covers some of the details of each of the datasets used in this work, how they were used, and how they were processed them into representational dissimilarity matrices (RDMs). Figure 3.1 give a high-level overview of the processing pipeline used to enhance the BOLD5000 dataset.

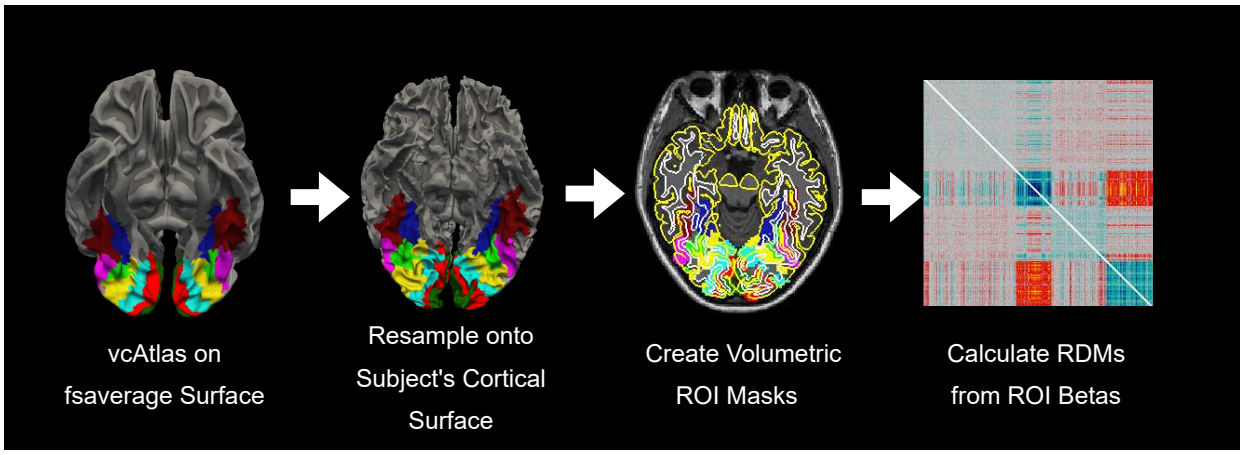

Figure 3.1: Overview of the processing pipeline used to enhance the BOLD5000 dataset. New cortical atlases are mapped onto each participant's cortical surface, new volumetric ROI masks are generated, vectors are extracted for each of the stimuli presentations, and finally, RDMs are created from comparisons of the stimuli response vectors.

3.1 Datasets

3.1.1 ImageNet in BOLD5000

The use of images from the ImageNet dataset in the BOLD5000 presents a unique opportunity because the ImageNet Large Scale Visual Recognition Challenge (ILSVC) benchmark remains the standard benchmark and training dataset for image classification models such as those included in this paper [45]. Before the BOLD5000 data, representations of neurological data tended to be col-

lected for simple stripped-back stimuli such as a clearly cut-out image against a grey background. While these simple stimuli enabled research comparing biological representations to artificial representations (e.g., like [18]), they had limited additional uses. For example, these stimuli were too simple and too few for fine-tuning networks to exhibit biologically consistent embeddings. The use of complex images like those within the ImageNet dataset may be non-ideal for traditional fMRI research, but they enable a wealth of experiments examining artificial neural networks (ANNs).

ImageNet classes are based on the WordNet synset hierarchy. In theory, this synset hierarchy can be used to establish the relationships between image classes. In practice, however, there are known deficiencies in the WordNet structure and most researchers resort to creating custom “supercategories" for the the image classes. Figure 3.2 illustrates a small portion of the synset hierarchy for the 1000 ILSVRC image classes. As can be seen in the far left of the synset tree, the classes for the two stringed instruments, "violin" and "acoustic guitar", are correctly placed near each other and therefore have a high path similarity. A counter-counter example is seen with the "spotlight", and "traffic light" pair of synsets. Both should fall under the hypernym of artifact, i.e. a man-made object. Instead, "traffic light" is labeled as an abstraction because of its use as a "signal" (there is no other hypernym path available for "traffic light"). Another problematic example is that of "orange", which has multiple hypernym paths. One path proceeds up through "natural object", while the other, confusingly leads from "food", through "solid", to "matter". This is a problematic series of hypernyms because while oranges and foodstuffs, in general, are indeed "solid matter", almost all other labels, including vertebrates and invertebrates, are not derived from this branch. "Natural object" was the supercategory selected for food objects because of this incongruity.

To overcome the deficiencies in the synset hierarchy new supercategories were manually selected for this paper. They were selected for their similarity to other categories used in fMRI research. The five supercategories used are Vertebrate, Invertebrate, Natural Object, Artifact, and Place. Table 3.1 summarizes the supercategories created for this project and the hypernyms of each supercategory. Each of the ImageNet labels was sorted into a supercategory by searching

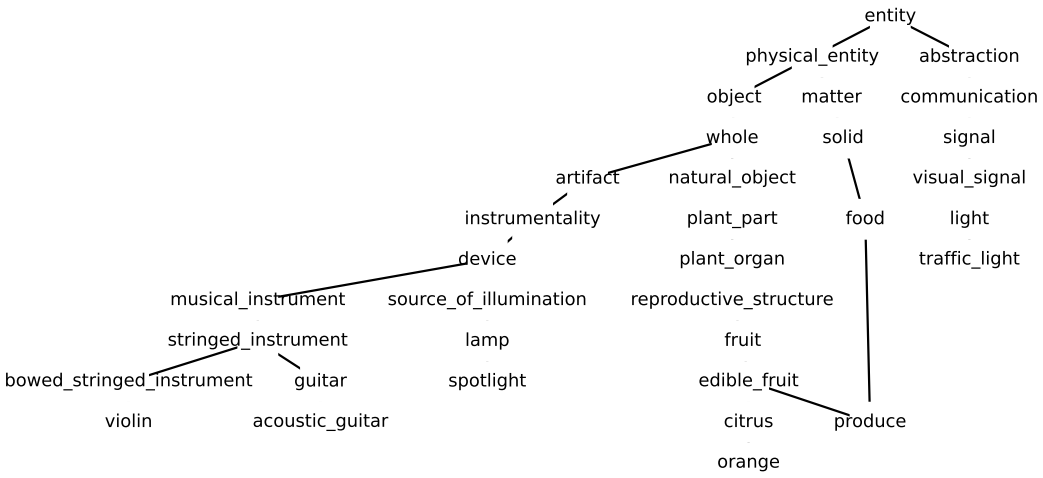

Figure 3.2: Graph representing the hypernym paths for the four synsets "violin", "acoustic guitar, "traffic light", "spotlight" and "orange". All ImageNet labels are nouns, and all nouns in WordNet have "entity" as their root.

its hypernym paths for one of the hypernyms listed in the table. Once a matching hypernym was found, the image was sorted into the corresponding supercategory.

Table 3.1: Five supercategories were created by combining the synset labels from the ImageNet stimuli. Each supercategory is made up of a set of hypernyms.

Supercategory	Hypernyms	Num. Images
Vertebrate	[animal, person]	646
Invertebrate	[invertebrate]	96
Natural Object	[food, plant, fungus, plant_part]	128
Artifact	[artifact]	912
Place	[structure, geological_formation]	134

Challenges with ImageNet Images

Multiple challenges were encountered with the ImageNet images used as BOLD5000 stimuli that likely contribute to the overall level of noise in the final RDM data and detract from the accu-

racy of the categorical analysis. The first challenge is that while ImageNet images are theoretically supposed to be singular images of just the entity described in the label without much additional context, many ImageNet images contain far more than just the entity in the image. This differs significantly from typical fMRI stimuli, which usually have the entire background whited out to focus the test subjects' attention. The second major challenge is related to the first, and is the fact that quite frequently the additional entity in the image is a human face. This is doubly problematic because the human brain has highly tuned areas of its visual cortex dedicated to the detection and decoding of human faces, such as the fusiform face area [46]. A number of the more egregious examples of this are presented in Figure 3.3. While the BOLD 5000 authors claimed to have selected "exemplars" of each label, it appears likely that two images were merely selected at random. Images that are supposed to be of artifacts or places, and are therefore expected to have a large dissimilarity to images of vertebrates, are going to be significantly affected by this source of noise.

3.1.2 vcAtlas Cortex Atlas

vcAtlas is a cross-validated cytoarchitectonic atlas of the human ventral visual stream [47]. Cytoarchitectonic regions of interest (cROIs) are defined by the spatial arrangement and type distribution of neural cells in the cortical ribbon. For the vcAtlas study, the borders of these regions are defined by statistically significant changes in the cellular structure using the gray level index (GLI). Data from 11 postmortem adult brains was used to define 8 cROIS; four in the occipital lobe (hOc1-4), and four in the fusiform gyrus (FG1-4). Each of the postmortem brains was aligned to Freesurfer's common fsaverage surface using cortex-based alignment. Once aligned, maximum probability maps (MPM) were created for each cROI based on the proportion of subjects for which a given vertex was contained within that cROI. The final atlas was created by taking the overlapping MPMs and assigning each vertex of the fsaverage surface to the cROI with the highest probability.

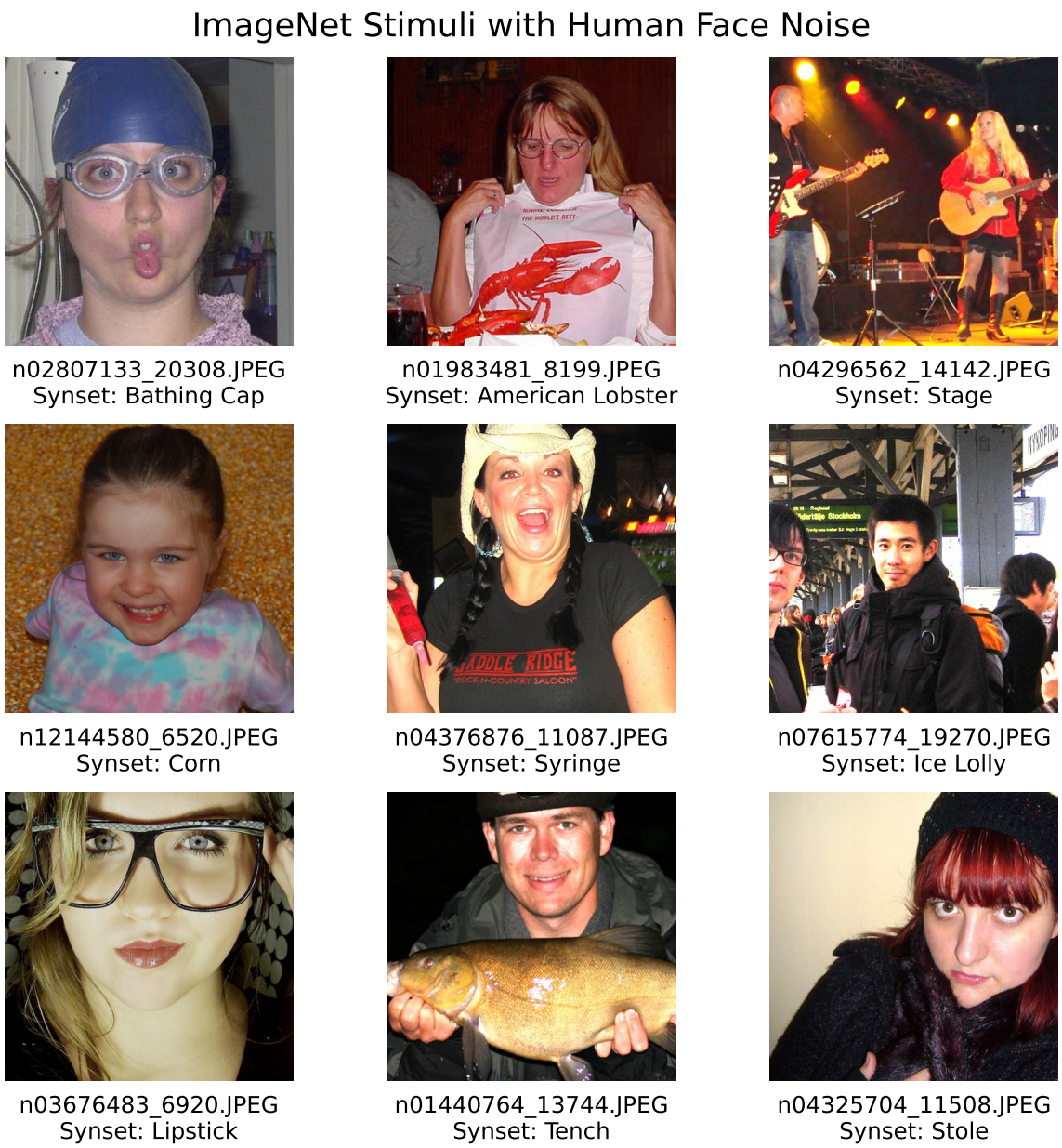

Figure 3.3: Nine BOLD5000 stimuli and their respective ImageNet labels. Each image is shown as it was presented during the original study, i.e. with cropping and resizing applied. These examples demonstrate how many of the stimuli images chosen for the BOLD 5000 prominently contain human faces alongside, or even instead of, the entity described in the image label.

3.1.3 visfAtlas Cortex Atlas

Similar to vcAtlas, visfAltas is an atlas of the human cortex, but instead of focusing on cytoarchitectonic regions of the brain, visfAtlas is focused on functional ROIs of the early-visual and category-selective regions of the cortex [48]. Early visual areas of the brain were mapped using retinotopy. Category-selective regions for characters, bodies, places, faces, and objects were mapped using functional localization. Additionally, a motion-selective region, hMT+, was defined using a separate localizer.

Combined, vcAtlas and visfAtlas give both structural and functional ROI mappings of the human visual cortex enabling future vision research.

3.2 Preprocessing

All betas were provided in NIfTI format, divided by subject and session. The image coordinate transforms, provided within the file headers, did not correspond to the transforms used for brainmasks, ROI masks, and T1w anatomical images from the original BOLD5000 release. This transform information is required for several other processing steps, including the re-application of the functional ROI masks from the original release of the BOLD5000 and the application of the two new ROI atlases, vcAtlas, and visfAtlas, to the four participant brains. We solved this issue by intuiting that the provided NIfTI files were derived from the same fMRIPrep derivatives as the original BOLD5000, thus allowing us to utilize the same alignments and brainmasks. The affine transforms from the original BOLD5000 brainmasks were applied to the GLMSingle beta files and results were visually checked against both the original brainmasks and the T1w anatomical scans of the participants to confirm good alignment. The generation of a global brainmask intersection was also required for each of the four subjects across all sessions. RSA analysis calculates distance metrics for each pair of input stimuli and therefore requires that the input vectors for each of the stimuli have the same number of dimensions (in the case of fMRI, each dimension is a voxel). The BOLD5000 is somewhat unique in that it is largely made up of single presentations of each stimulus, and the order of the stimuli is randomized across multiple sessions for each participant. This

poses a challenge because even very minor positional changes between sessions can lead to the introduction of invalid voxel values, especially around the very edge or pial surface of the brain.

The fMRIPrep pipeline uses several advanced tools to correct for any changes [49], however, it was found that the participant brain masks provided in the original BOLD5000 release still resulted in invalid voxels being included for some trials. To address this issue, a global mask was calculated for each participant using the intersection of the valid voxels for each input across all sessions. These global participant brain masks were applied to every ROI to ensure that no invalid voxel data was entered into the RSA calculations.

3.2.1 FreeSurfer

FreeSurfer is an incredibly powerful suite of tools originally developed to reconstruct cortical surface models from T1w anatomical scans. A further goal of this original development in reconstructing the cortical surface is finding alignments between subject brains based on cortical folding patterns [50]. It is this alignment functionality that makes the FreeSurfer a vital component of the fMRIPrep pipeline used in the original BOLD5000 release.

As follow-on researchers, we leverage these FreeSurfer derivatives to extract additional information from the dataset. We use FreeSurfer to parcellate a reconstructed cortical surface based on its folding patterns using specially crafted atlases. We used this functionality to identify and extract additional areas relevant to vision based on structural connectivity or functional response to images using vcAtlas and visfAtlas respectively. Our analysis is concerned with comparing the BOLD activations of voxels in volumetric space. Thus, several steps were required to convert these surface atlases into volumetric ROI masks.

First, the labels from the atlases were resampled from the standard fsaverage surface to each of the subjects' cortical surfaces. This is accomplished using the mri_surf2surf command. With the labels for each atlas and ROI now resampled onto the subjects' cortical surfaces, the labels were used to define a volumetric ROI as by the volume of gray matter that makes up the cortex beneath the cortical surface label. This is accomplished with the mri_label2vol command with projection

fraction set to include 100% of the volume between the pial and white matter surfaces. The output of this function is a series of volumetric ROI masks in NIfTI format, similar to the ROI masks from the original BOLD5000. All ROI masks generated using FreeSurfer also had the global mask for each participant applied to them to ensure that only valid voxels would be extracted for a given ROI.

3.3 Representational Similarity Analysis

After preprocessing and utilizing FreeSurfer to identify ROIs, we create RDMs from the neural data. We construct RDMs using the established methodology [6, 18]. Here, we briefly summarize the process:

RDM construction. Given a single feature f and a single stimulus s, $v = f(s)$, where v is the value of feature f in response to s. Likewise, the vector

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}^T = \begin{bmatrix} f_1(s) \\ f_2(s) \\ \vdots \\ f_n(s) \end{bmatrix}^T \tag{3.1}$$

can represent the feature values of a collection of n features, $f_1,\ f_2,\ ...,\ f_n$, in response to s. If one expands the representation of s to a set of m stimuli $S = s_1,\ s_2, ...,\ s_m$, the natural extension of $\vec{v}$ is the set of feature value collections $V = \vec{v}_1,\ \vec{v}_2, ...,\ \vec{v}_m$, in which $s_i \in S$ is paired with $\vec{v}_i \in V$ for each $i = 1, 2, ..., m$. The last step before constructing an RDM is to define the dissimilarity score between any two $\vec{v}_i \in V$ and $\vec{v}_j \in V$. We use the symmetric function

$$\psi(\vec{v}_i, \vec{v}_j) := 1 - \frac{(\vec{v}_i - \bar{v}_i) \cdot (\vec{v}_j - \bar{v}_j)}{\|\vec{v}_i - \bar{v}_i\|_2 \|\vec{v}_j - \bar{v}_j\|_2} \tag{3.2}$$

where $\bar{v}$ is the mean of the features in $\vec{v}$. An RDM R may then be constructed from S, V, and ψ as:

$$R = \begin{bmatrix} \psi(\vec{v}_1, \vec{v}_2) & \psi(\vec{v}_1, \vec{v}_3) & \dots & \psi(\vec{v}_1, \vec{v}_m) \\ & \psi(\vec{v}_2, \vec{v}_3) & \dots & \psi(\vec{v}_2, \vec{v}_m) \\ & & \ddots & \vdots \\ & & & \psi(\vec{v}_{m-1}, \vec{v}_m) \\ & & & \end{bmatrix} \tag{3.3}$$

3.3.1 Biological Similarity Metric

The methodology for comparing a network to a biologically constructed RDM is simple: After constructing an RDM R_1 for the network following the procedure outlined in 3.3 using the same stimuli set S, one can compute the similarity to the biological RDM R_2 with the function

$$biologicalSimilarity = \rho(\hat{R}_1, \hat{R}_2) \tag{3.4}$$

where $\hat{R}$ is the flattened RDM and ρ corresponds with a similarity metric e.g., Pearson's correlation. Note, many works suggest estimating the RDM during training by only considering a subset of the stimuli [19].

3.3.2 The `rsatoolbox` Package

`rsatoolbox` is a Python package for representational similarity analysis developed by Nili et al. from the laboratory of Nikolaus Kriegeskorte, one of the pioneers of RSA use in fMRI research [51]. Originally developed for Matlab, `rsatoolbox` is under active development and can be used for the generation and comparison of RDMs, the creation and evaluation of multiple types of models with various statistical tools, and visualization tools. All fMRI RDMs, RDM

comparisons, and models were performed with `rsatoolbox`. (Initial RDM generation for the ANNs was generated using functionality built into the Net2Brain tool as detailed in 3.4.)

3.3.3 Categorical Model Analysis

While the end goal of our RSA analysis is to compare the biological data from the BOLD5000 fMRI trials to ANNs, RSA also allows us to leverage other types of dissimilarity models such as the supercategories within ImageNet as described in Section 3.1.1. First, categorical RDMs are generated for each supercategory as illustrated in Figure 4.1. These consist of an RDM where all images of the same category are assigned the minimum distance/dissimilarity for a given metric and all images from other categories are assigned the maximum distance/dissimilarity for a given metric.

Using rastoolbox's Model Weighted functionality, these individual category RDMs are combined as a weighted sum and are linearly fit to the Mean Subject RDMs for the vcAtlas ROIs. The model weights are then used to predict the final categorical model shown in Figure 4.2. This categorical model is a representation of the relative similarities of each of the supercategories as perceived by the human brain. Categorical models such as this can act as a reference point for later RSA analysis because they rely on additional structural information that is embedded into the ImageNet image labels.

As a final analysis step, each of the vcAtlas ROIs from the mean subject is compared back against the predicted categorical model to determine which ROI or ROIs best represent the supercategorical structure of the data. Figure 4.3 shows the correlation of each of the ROIs to the categorical model. In the case of the BOLD5000 data, the Left-Hand Fusiform Gyrus 3 (LHFG3) is the best exemplar of the categorical model. Going forward, LHFG3 is selected going forward as the best single ROI with which to compare the ANNs.

3.4 Net2Brain

Here, we link our preprocessed data and subsequent evaluations to Net2Brain, a toolbox for researching the internal geometric representations of artificial deep neural networks, particularly convolution neural networks, using RSA. One of the strengths of Net2Brain is the very extensive set of over 600 models that it is preconfigured to pull down, extract activations from, and calculate RDMs for. Net2Brain can pull models not only from the official PyTorch model zoo but also from timm, the Pytorch Image Models library created by Ross Wightman. All of the aforementioned 600+ models available to Net2Brain come pre-trained and are fully ready for activation extraction. All of the stimuli from the BOLD5000 are made available to Net2Brain and once it pulls down the pre-trained model in question, it presents each of the BOLD5000 images to the model as input and performs a forward pass. The model activations from each of the model's convolutional layers are then extracted and stored on disk. Once all of the activations have been extracted, RDMs for each of the convolutional layers are calculated. As of the time of writing, the toolbox enables creating RDMs using Pearson's correlation, and there are plans to add various other distance metrics.

3.4.1 Model Selection

Of the over 600 models available, four were chosen based on a couple of criteria. First, due to the limitations in the architecture of both Net2Brain and `rsatoolbox`, the calculation of RDMs required substantial amounts of memory given the number of unique stimuli in the BOLD5000. There was, therefore a relative size limit to the number of output activations in a model given the memory limits of available hardware. The second criterion was to achieve a representative sampling of ANN model architectures that are designed for image classification tasks and pre-trained on the ImageNet dataset over time. The four models chosen were: AlexNet [52], progenitor of all subsequent deep convolutional neural networks, ResNet50 [53], which introduced skip connections to neural network architectures, MobileNetv2 [54], which was specifically designed to perform well even on restricted hardware such as mobile devices, and finally, EfficientNet [30],

which expands on the same architectural concepts present in MobilNet with efficient network scaling.

3.5 Fiedler Vector Partitioning

In this section, we detail how we employ Fiedler partitioning, a graph-based technique, on the processed data. Fiedler partitioning aims to partition a graph into two distinct groups by utilizing the Fiedler vector, which corresponds to the second smallest eigenvector of the graph Laplacian matrix [55, 56].

3.5.1 Experiments

We analyzed individual RDMs for three BOLD5000 participants (CSI1-3), and a mean RDM (averaged subject data) for fMRI data specific to the $LHFG3$. Each RDM is composed of the following supercategories: vertebrate, invertebrate, natural object, artifact, and place. From these super categories, we first extracted subsets of two classes from each RDM before combining all five super categories into two overarching classes: animate and inanimate. Here, the animate class included the vertebrate and invertebrate categories, while the inanimate class encompassed the natural_object, artifact, and place categories. We then applied Fiedler partitioning to these RDMs and recorded the classification accuracy for each class in a pair. The pseudo-code for finding the Fiedler partitioning accuracy for an RDM is detailed in Algorithm 1.

Algorithm 1 Fiedler Partitioning Classifier

Require: Representational Dissimilarity Matrix R
Ensure: Classification Accuracy
1) Get a subset R_i of R with two categories.
2) Compute Adjacency Matrix $A = 1 - R_i$.
3) Compute Degree matrix from A.
4) Compute Laplacian matrix: $L = D - A$.
5) Get the second smallest eigenvector e_2 for L.
6) Compute Fiedler partitioning: $P_1 = \{i \in N : e_2(i) < 0\}$ and $P_2 = \{i \in N : e_2(i) > 0\}$.
7) Compute Accuracy = $(|P_1| + |P_2|)/len(e_2)$

Chapter 4

Results

4.1 Representational Similarity Analysis

4.1.1 Categorical Model Analysis

Figure 4.4 shows the RDM for LHFG3 from the mean subject with the images sorted by supercategory. Comparing LHFG3 Figure Figure 4.4 to the categorical model Figure 4.2, it is clear how the supercategory representations cluster together. This can be explored further through the comparison of RDM correlations.

Figure 4.5 shows the RDMs of the layer with the highest correlation to the categorical model for each of the four ANNs investigated. When visually comparing the categorical model, Figure 4.2, the mean subject fMRI response, Figure 4.4 and the ANN responses, Figure 4.5, a correspondence between the representation of the supercategories is evident.

4.1.2 RDM Comparison

Direct comparison of RDMs can be accomplished through several different similarity measures. Here, we report Pearson correlation, an established standard for use in RSA [6]. Table 4.1 presents the Pearson correlation between the categorical model and each of the four ANNs under test.

Table 4.1: Comparison of Mean Subject LHFG3 RDM to Categorical Model and ANN RDMs

Model	Pearson Correlation ± SEM	p (against 0)
Categorical	0.165 ± 0.009	< 0.001
AlexNet	0.054 ± 0.006	< 0.001
MobileNet v2	0.023 ± 0.003	< 0.001
ResNet50	0.031 ± 0.004	< 0.001
EfficientNet b0	0.015 ± 0.002	< 0.001

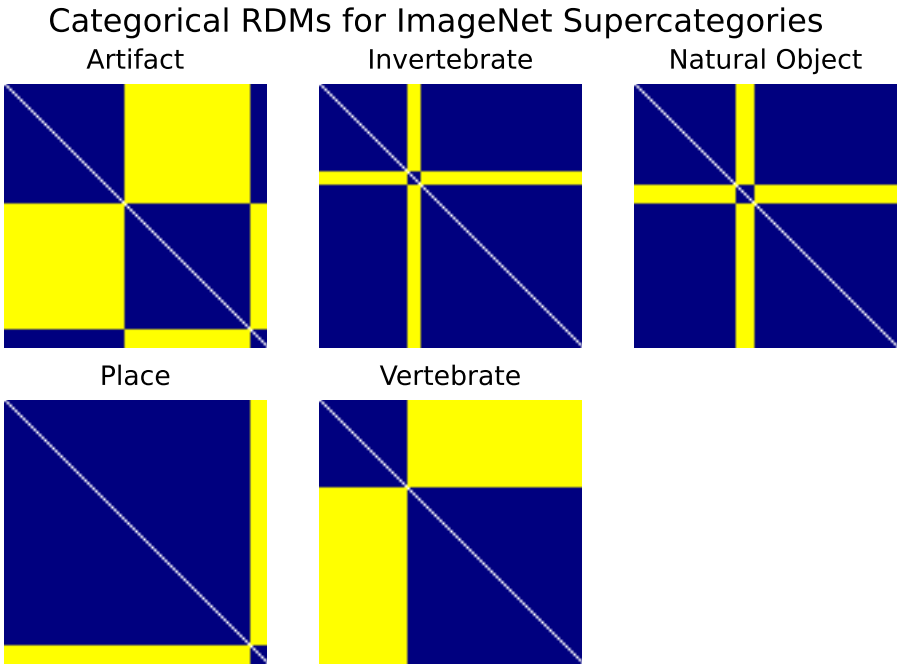

Figure 4.1: Categorical RDMS for each ImageNet supercategory. Categorical RDMs consist of an RDM where all images of the same category are assigned the minimum distance/dissimilarity for a given metric (i.e., for the 1-r distance metric, 0), and all images from other categories are assigned the maximum distance/dissimilarity for a given metric (i.e. 1).

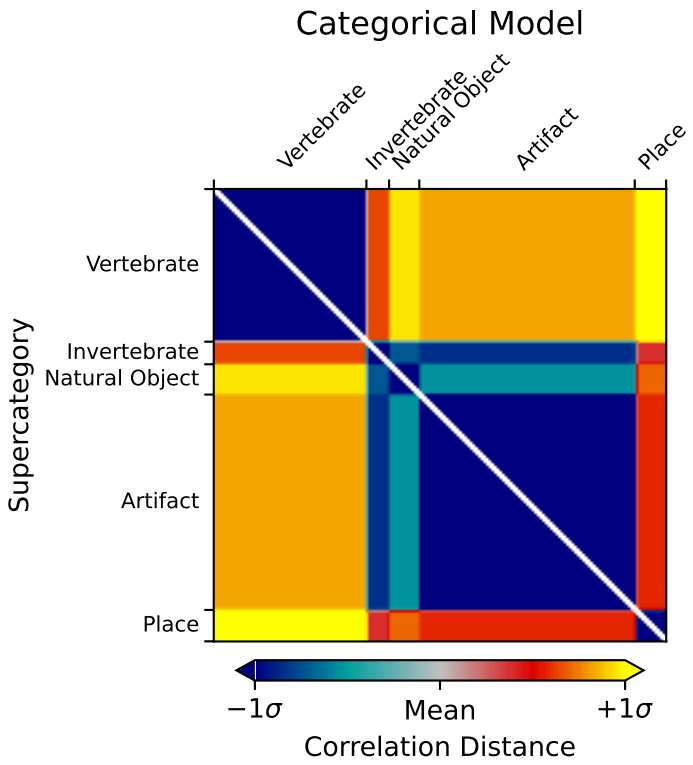

Figure 4.2: Predicted output RDM of the weighted categorical model.

An unexpected result of this analysis is the inverse relationship between model age and its biological similarity. AlexNet [52], the model arguably kicked off the deep convolutional neural network revolution in machine vision, has the highest biological similarity of the models tested, and EfficientNet [57], the most modern and highest performing classification model, has bar far the lowest biological similarity.

4.1.3 Comparing fMRI ROIs to Individual ANN Layers

In Figure 4.6, we break down our evaluation layer-by-layer to provide fine-grained details on which components of the trained network best exhibit biological similarity.

4.2 Fiedler Vector Partitioning

One of the goals in reprocessing the BOLD5000 dataset using the vcAtlas and visfAtlas maps was to enable future research into comparing how various components of an ANN, such as individ-

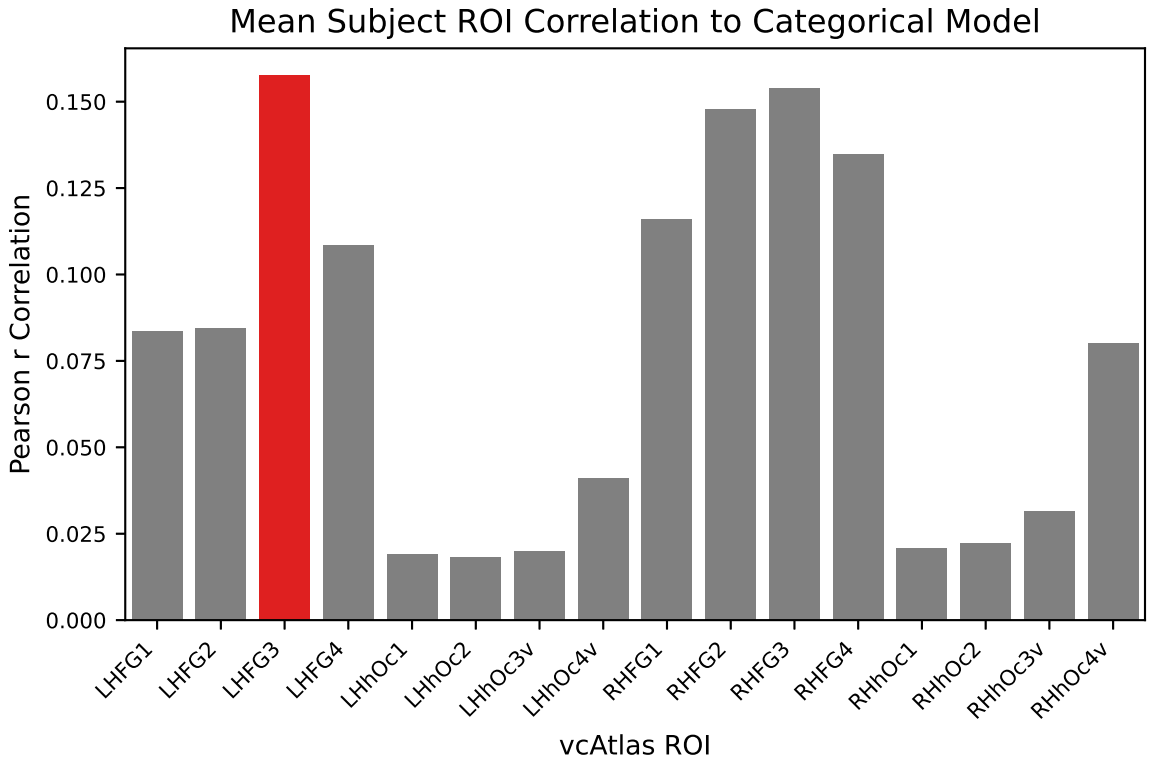

Figure 4.3: An exemplar ROI is chosen from the available vcAtlas ROIs by comparing its Pearson correlation to the categorical model (Figure 4.2). The Left-Hand Fusiform Gyrus 3 (LHFG3) (highlighted in red), was found to have the highest correlation with the categorical model.

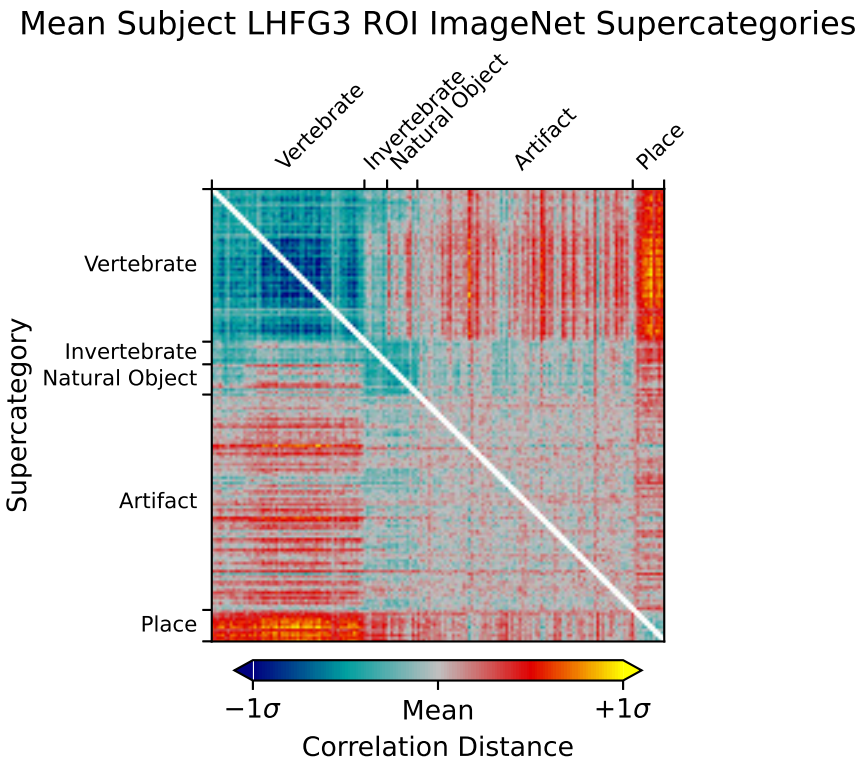

Figure 4.4: RDM of the Left-Hand Fusiform Gyrus 3 (LHFG3) ROI calculated from the mean subject using the correlation distance metric. Image inputs are sorted by their ImageNet supercategory. The clustering of similar images within supercategories is visible, as is the dissimilarity between supercategories.

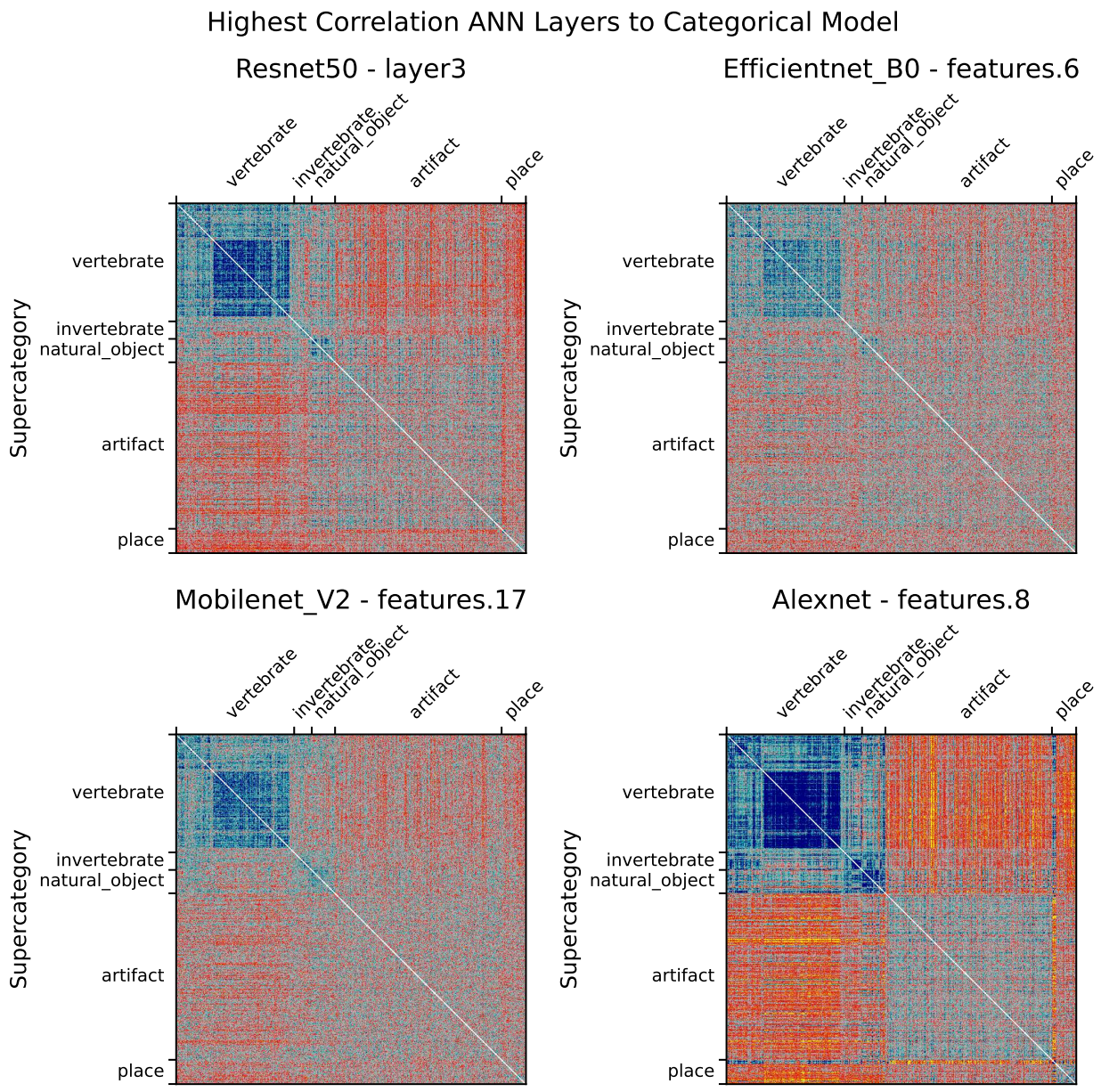

Figure 4.5: RDMs from each of the four ANNs ordered by ImageNet supercategory. Each RDM is taken from the ANN layer with the highest correlation to the categorical model.

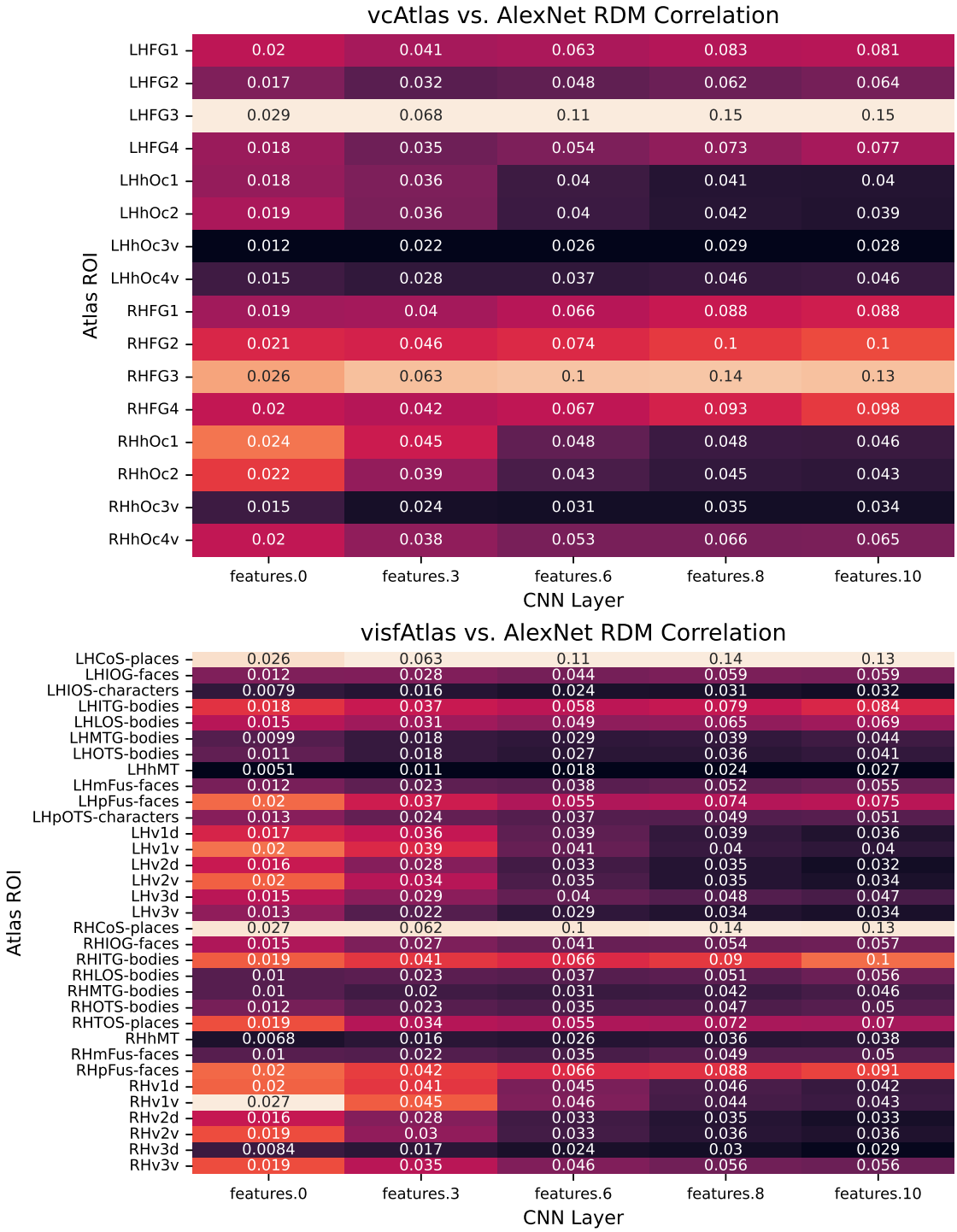

Figure 4.6: Pairwise analysis of each of the layers of AlexNet, the ANN found to have the highest biological similarity, to the two new ROI brain maps used in this research: vcAtlas and visfAtlas. In the vcAtlas comparison it can be seen that while the LHFG3 ROI does dominate the comparison, there is a correlation between the first two layers of AlexNet with the early visual cortex in Oc1 and Oc2.

ual convolutional layers, can be compared to specialized structures in biological representations. For example, the theoretical concept behind the ventral visual stream in the human brain is that visual information flows from the early visual cortex at the back of the brain forward into the Fusiform Gyrus. Along the way, the visual stimuli are decoded in increasingly higher-order representations. Our findings give credence to the observation that deep convolutional neural networks mimic some of what occurs with this process.

The human brain also has several very specialized areas for certain tasks such as facial recognition in the Fusiform Face Area (FFA) [46], one of the ROIs included in the visfAtlas. The goal is to provide the data so that these specialized areas of the brain can be used to analyze and train equivalent specialized components of ANNs.

Figure 4.7 displays the Fiedler partitioning accuracies for the various ANNs from our experiments, and Figure 4.8 shows the partition accuracy for the biological data. All accuracies illustrate the separability of class pairs — the results indicate that the human subjects consistently achieved higher classification accuracy when discriminating between the vertebrate class and the invertebrate, natural object, and place categories. This shows that the feature embeddings in the LHFG3 are well clustered for those categories.

Overall, our findings indicate that the representational dissimilarity matrix effectively clusters the data. Similar to our findings with the RDM comparisons, a surprising trend emerges with the ANNs. AlexNet, the oldest of the ANNs, produces a far higher Fiedler partitioning accuracy than the newer models. EfficientNet B0, in particular, does not produce results significantly above noise for most of the supercategory pairings. This is an interesting result that will be explored in more depth in Section 5.

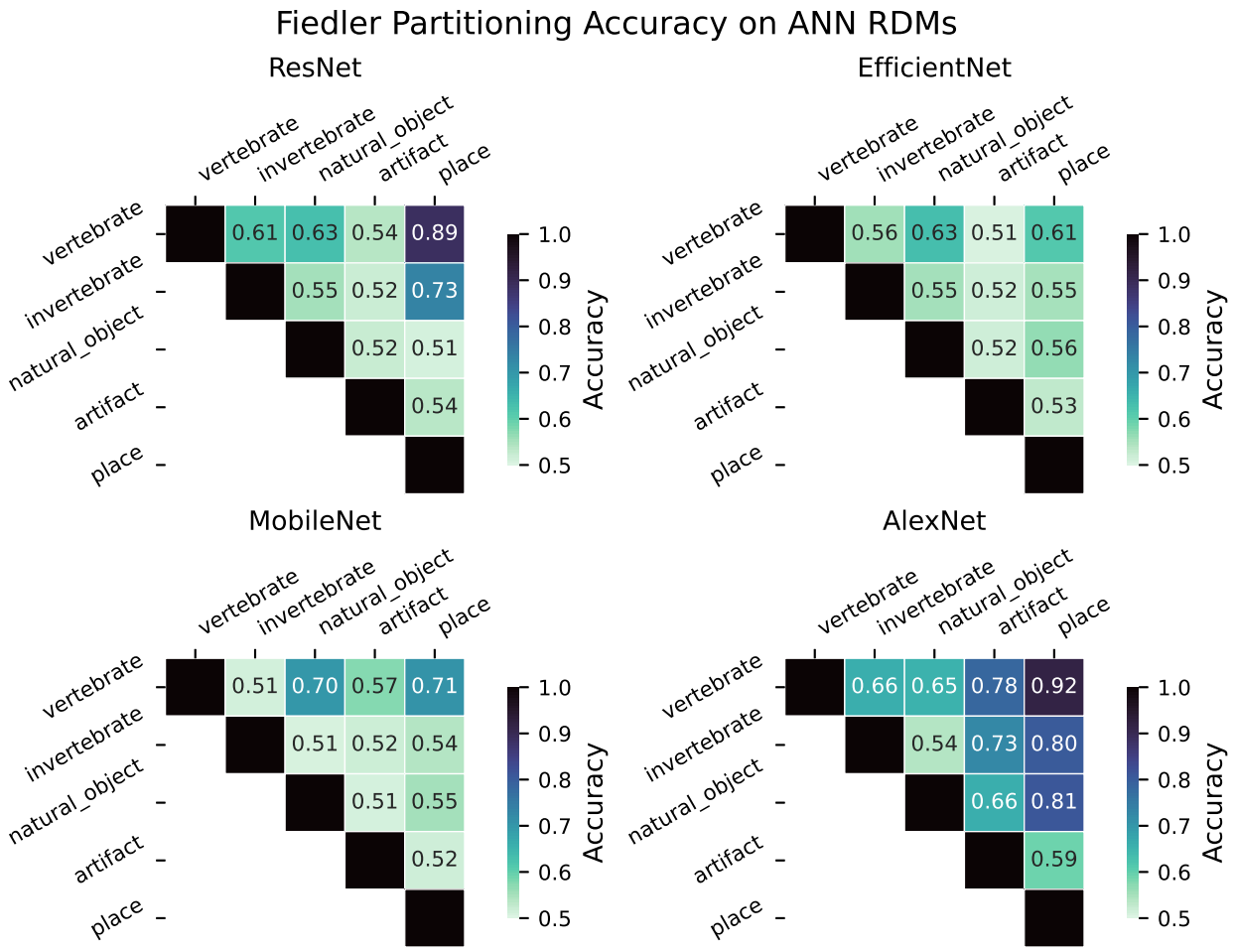

Figure 4.7: Fiedler partitioning accuracy for each of the four ANNs. AlexNet showed the highest accuracy using Fiedler partitioning on convolutional layer activations, followed by ResNet, then MobileNet, and finally EfficientNet. EfficientNet did not show accuracy above noise for the majority of supercategory comparisons.

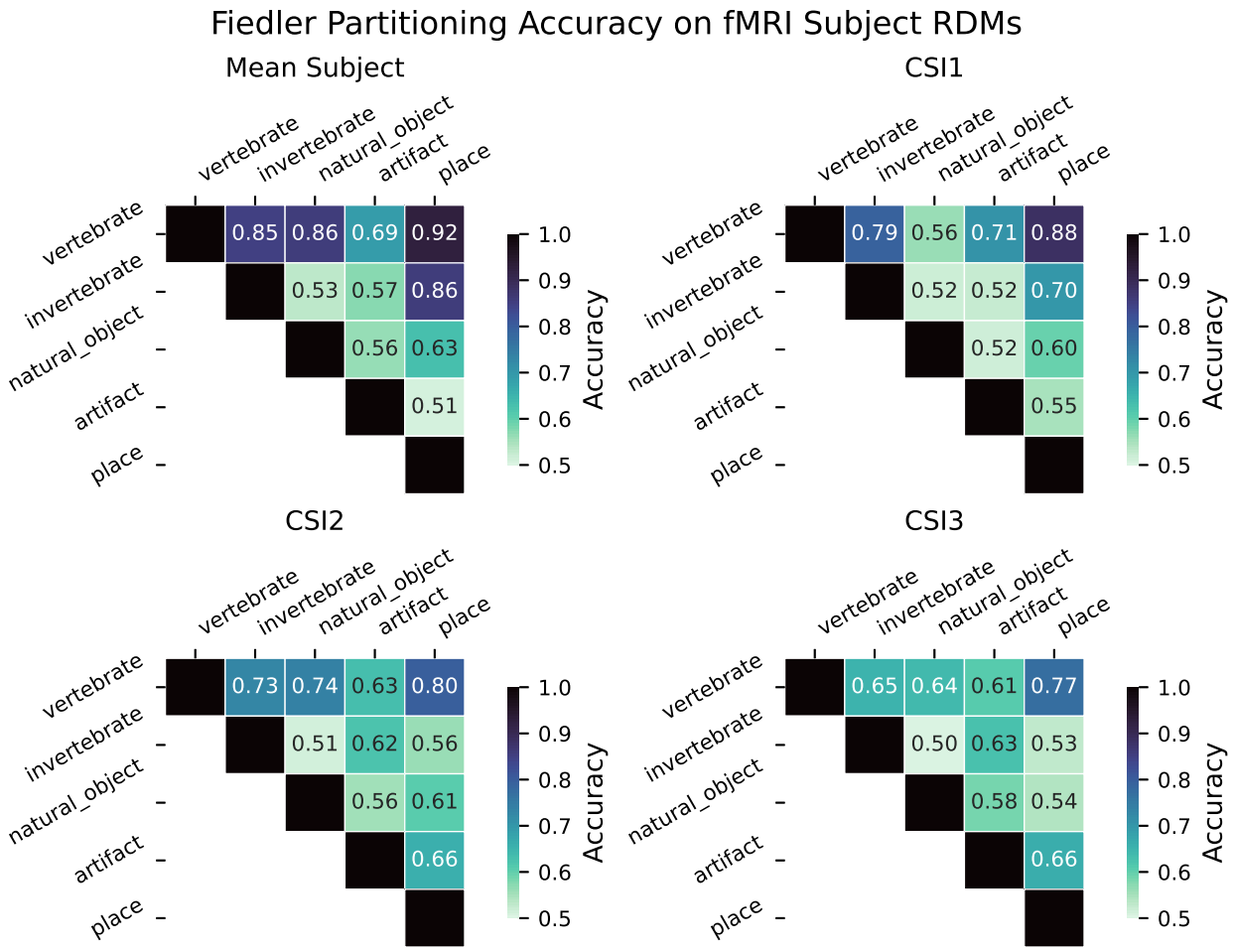

Figure 4.8: Fiedler partitioning accuracy for LHFG3 for CSI1-3 BOLD5000 subjects across the five super categories.

Chapter 5

Discussion

ANNs have long suffered from decreased performance as a result of their sensitivity to random noise and adversarial attacks. Recent works have shown that fine-tuning a network representation to align with a biological standard fortifies networks against both noise and adversarial corruptions of images [18, 19]. However, exploration of these ideas has been limited by the unavailability of public datasets: prior works have relied on private datasets [19] or datasets with a limited number of stimuli [18]. The BOLD5000 dataset has always been a promising resource for investigating just this, but the data was not intended for use by researchers without a strong neuroscience background to explore. Here, we eliminate this barrier — our curation and investigations of the BOLD5000 data will now enable the broader community to explore the viability of biological representation in networks.

An important and surprising result of our analysis is that recent, more advanced, neural networks, such as EfficientNet [57], have lower neuro-similarity than the much older and simpler AlexNet, despite also performing much better on the standard ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The discovery that ANNs are diverging from their biological inspiration is not, in and of itself, surprising, but it does emphasize the fundamental question of whether or not neuro-similarity is an asset, a hindrance, or simply a non-factor. Are these newer models performing better on an, admittedly artificial, metric because of their neuro-dissimilarity or in spite of it? Humans are not susceptible to the same adversarial attacks that ANNs have been shown to be susceptible to, so this divergence in the geometry of ANN embedding spaces from their human counterparts may open up new avenues of attack.

To put a finer point on it, are more advanced models achieving higher accuracy by focusing on minutia instead of the complete composition, e.g., are the features being extracted from an image of one of ImageNet's many dog breed classes focused on things like fur texture and color as a way to correctly classify the breed, the source of the image classification, or on the fact that the image

depicts a four-legged creature with two eyes and other mammalian features? Having a fragmented embedding space that emphasizes minutia is likely to make a model more susceptible to adversarial attacks. To use the example above, a model that has been overfit to the point where it only focuses on fur pattern features to identify something as a dog, could be tricked into misidentifying a common artifact such as a box, by covering it with a fake fur texture or image.

We expect images containing similar features to elicit activations that are closer together within the embedding space while dissimilar activations should exist further apart — [21] investigated the presence of this phenomenon, finding certain neuron groups in CLIP activated or deactivated in response to similar concepts. Fiedler partitioning of an RDM should be able to exploit this clustering of like embeddings to get us in the ballpark of a reasonable classification by selecting an appropriate class category regardless of whether or not there is a strong correlation between the ANN and biological benchmark. By demonstrating that this works well for a model like AlexNet, but not for a model like EfficientNet, we imply that these more advanced models are not creating the expected clusters within their embedding space. This leads to the question of how these new ANNs are actually structuring their feature space or whether they are extracting a similar set of features at all. Our work shows that ANNs trained for classification performance are evolving internal embedding space geometries ever more dissimilar from the human vision system and that these embedding spaces lack a geometry that clusters like image subjects together. We can either conclude that state-of-the-art ANNs are creating a novel way to learn and store image feature representations, or we must conclude that embedding spaces are becoming ever more disjoint because of the singular push to maximize classification accuracy.

Since learned representations like [21] do seem to demonstrate this phenomenon with CLIP embeddings, and since CLIP embeddings match or surpass the performance of the models we evaluate [24], it seems likely that the biological ideal does correspond with robustness. However, a full investigation of the viability of the biological benchmark is beyond the scope of this work — and likely beyond the scope of any singular work. Instead, a wealth of future research is needed to tease out the intricacies of what kinds of representations correspond with robustness. The most

impactful outcome of this work is the facilitation of these future research projects via a shared, publicly available dataset that allows researchers and practitioners to scrutinize the evidence for a biologically grounded representation, and investigate alternatives.

Finally, the curation of this data also facilitates additional uses of the data: modeling neural processes and creating new biologically consistent architectures. Neural networks are the premier means for modeling neural data. However, it has also been shown that current architectures have largely plateaued [15] and that all networks are equally predictive of the human inferior temporal cortex. This is problematic because these models still fail to predict certain properties of visual processing [15]. Our data could facilitate the creation of neural network designs that are biologically grounded. Previously, work has shown that networks deliberately modeled on neural phenomena exhibit higher biological consistency than traditional CNNs [22], which corresponds with higher performance. However, even this work would vastly benefit from expanding methods for comparing with biological benchmarks via novel techniques like extending RDMs into Laplacian matrices [20].

Chapter 6

Conclusion

Here, we establish a new biological benchmark for embedded representations. Our experiments on our benchmark establish the viability of utilizing this data to enhance the robustness of learned representations to inputs like adversarial attacks. Specifically, our experiments with Fiedler partitioning showcase how biologically grounded representations facilitate interwoven separability and clustering of data. As part of this work, we release our curated data and a framework to facilitate further investigation.

Bibliography

[1] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[2] Ha Hong, Daniel L. K. Yamins, Najib J. Majaj, and James J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, April 2016.

[3] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, 6(1):32672, September 2016.

[4] Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016.

[5] Daniel L Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[6] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[7] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915, November 2014.

[8] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in

higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014.

[9] Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, 10(12):e1003963, December 2014.

[10] Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015.

[11] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, June 2016.

[12] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization. *Scientific Reports*, 8(1):3752, February 2018.

[13] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? Preprint, Neuroscience, September 2018.

[14] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, January 2021.

[15] Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, September 2021.

[16] Kshitij Dwivedi, Michael F. Bonner, Radoslaw Martin Cichy, and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, 17(8):e1009267, August 2021.

[17] Nicholas J. Sexton and Bradley C. Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, 8(28):eabm2219, July 2022.

[18] Nathaniel Blanchard, Jeffery Kinnison, Brandon RichardWebster, Pouya Bashivan, and Walter J. Scheirer. A Neurobiological Evaluation Metric for Neural Network Model Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5404–5413, 2019.

[19] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] Huma Jamil, Yajing Liu, Turgay Caglar, Christina Cole, Nathaniel Blanchard, Christopher Peterson, and Michael Kirby. Hamming Similarity and Graph Laplacians for Class Partitioning and Adversarial Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 590–599, 2023.

[21] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021.

[22] Nathaniel T. Blanchard. *Quantifying Internal Representation for Use in Model Search*. PhD thesis, University Of Notre Dame, March 2019.

[23] Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

[25] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022.

[26] Pouya Bashivan, Mark Tensen, and James J. DiCarlo. Teacher Guided Architecture Search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5320–5329, 2019.

[27] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural Architecture Search: A Survey. 2018.

[28] Chi-Hung Hsu, Shu-Huan Chang, Jhao-Hong Liang, Hsin-Ping Chou, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. MONAS: Multi-Objective Neural Architecture Search using Reinforcement Learning. 2018.

[29] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[30] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient Neural Architecture Search via Parameters Sharing. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4095–4104. PMLR, July 2018.

[31] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[32] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing.

[35] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, June 2010.

[36] Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. GLMsingle: A toolbox for improving single-trial fMRI response estimates, February 2022.

[37] Sarah Aliko, Jiawen Huang, Florin Gheorghiu, Stefanie Meliss, and Jeremy I. Skipper. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, 7(1):347, October 2020.

[38] Brett D. Roads and Bradley C. Love. Enriching ImageNet with Human Similarity Judgments and Psychological Embeddings. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3546–3556, Nashville, TN, USA, June 2021. IEEE.

[39] Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, February 2021.

[40] Linda B. Smith and Lauren K. Slone. A Developmental Approach to Machine Learning? *Frontiers in Psychology*, 8, 2017.

[41] Brandon RichardWebster, Samuel E. Anthony, and Walter J. Scheirer. PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, September 2019.

[42] Leslie G. Ungerleider and Mortimer Mishkin. Two Cortical Visual Systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT press, Cambridge, Mass, 1982.

[43] Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992.

[44] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13073–13087. Curran Associates, Inc., 2020.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-

Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.

[46] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11):4302–4311, June 1997.

[47] Mona Rosenke, Kevin S. Weiner, Michael A. Barnett, Karl Zilles, Katrin Amunts, Rainer Goebel, and Kalanit Grill-Spector. A cross-validated cytoarchitectonic atlas of the human ventral visual stream. *NeuroImage*, 170:257–270, April 2018.

[48] Mona Rosenke, Rick van Hoof, Job van den Hurk, Kalanit Grill-Spector, and Rainer Goebel. A Probabilistic Functional Atlas of Human Occipito-Temporal Visual Cortex. *Cerebral Cortex*, 31(1):603–619, January 2021.

[49] Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, January 2019.

[50] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012.

[51] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, 10(4):e1003553, April 2014.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[54] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, Salt Lake City, UT, June 2018. IEEE.

[55] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[56] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633, 1975.

[57] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019.