Learning High-Level Visual Representations via Synthesized fMRI

Matt Gorbett
Matt.Gorbett@colostate.edu

Chaitanya Roygaga
Chaitanya.Roygaga@colostate.edu

Nathaniel Blanchard
Nathaniel.Blanchard@colostate.edu

Abstract

Visual networks that correspond to biological neural data are more robust to image degradation, and generalize better in multi-task contexts. The recent release of BOLD5000, the largest public fMRI dataset, has the potential to enable wide-spread usage of these findings. While the dataset consists of four test subjects' fMRI recordings in response to 4,916 unique images, these numbers are still a far cry from what's required to train deep learning models, and the cost of the collection underscores just how difficult it would be to collect fMRI for a full dataset. We propose an alternative: by training a convolutional neural network (CNN) to learn BOLD5000 fMRI in response to image stimuli, we can generate new, brain-like data on unseen images. We use three novel evaluation tools to validate our generated fMRI: 1) manifestation of traditional neural separability (i.e., coarse class labels are separable), 2) exhibition of appropriate behavior based on visual attention in complex scenes (i.e., if a person and an object are both in a scene, the synthetic fMRI corresponds with the gradient-weighted activation map), and 3) whether the unsupervised network exhibits high behavioral similarity with human fMRI data (human-model similarity). Practically, this method provides computer vision researchers with the means to generate large-scale, neural-like data and evaluate it with robust, biologically-focused techniques. All code and data will be released upon publication.

1. Introduction

There is a long history of comparing neural networks to biological brains, such as those of primates [3], mice [17], and humans [1]. Arguably, such comparisons have often seemed like novelty findings; latent evidence supporting Rosenblatt's [23] statement that a neural network is "capable of perceiving, recognizing and identifying its surroundings," the same way as biological organisms. However, recent research has shown there are benefits to having

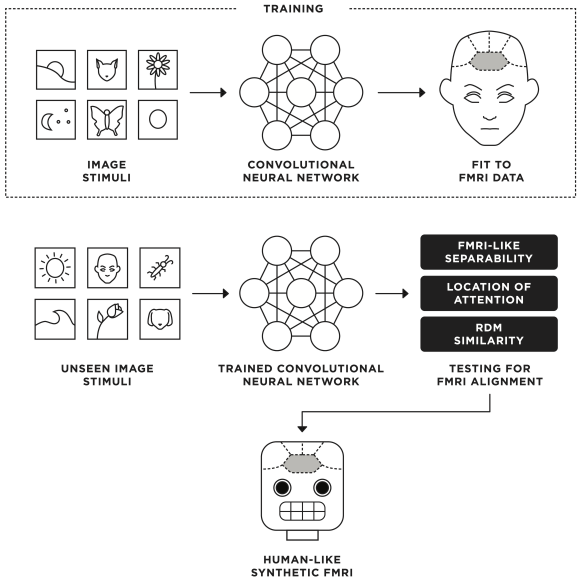

Figure 1: A major limitation of utilizing fMRI recordings to regularize deep learning models, which increases generalization and robustness, is the difficulty of obtaining neural data. In this work, we explore the possibility of training a neural network to generate human-like synthetic fMRI, which can be used on unseen data. We evaluate our network for biological consistency in three ways: 1) manifestation of neural-like separability on unseen data, 2) attention specific behavior in complex scenes, and 3) traditional behavioral similarity when compared with a fMRI representational dissimilarity matrix (RDM).

neural networks with biological consistency, such as improved robustness in the face of perturbed images [17, 5], and increased generalization in multitask contexts [1]. Further, while networks have traditionally been evaluated post-training, new techniques have shown that biological data can be used as a regularizer during the training process [17,

5].

A pressing issue limiting widespread usage of these practices is access to neurological data — many datasets are unreleased, or constrained in size and diversity. This issue has been somewhat alleviated by the release of BOLD5000 [4], the largest set of visual fMRI to date, which aims to enable researchers to explore statistical learning techniques on larger amounts of neural data. The study recorded fMRI responses from four human subjects in response to image stimuli across 4,916 unique examples. Yet still, while BOLD5000 is a notable addition to slow-event related fMRI data, Chang *et al*. [4] note in their conclusion that the dataset is small "compared to either human visual experience across one's lifespan or the millions of images used to train modern artificial vision systems."

Our novel method for creating synthesized fMRI intends to bridge the gap between the relatively small amount of actual fMRI data and the amount needed to create effective biologically-inspired neural networks. We use findings from the BOLD5000 research that applied the properties of region-specific fMRI to classify images into high-level categories. Namely, fMRI data from the parahippocampal place area (PPA), a brain region typically associated with scene processing [8], was able to distinguish images of objects from images of scenes (as shown in Figure 2), as well as images of animals from other objects.

We show that by fitting BOLD5000 images to their corresponding fMRI data, the brain naturally learns high-level classifications on new datasets. The method takes advantage of the known classification properties of the PPA region of the human brain in order to separate new images into distinct, high-level categories. Our model first maps an image to an fMRI vector with a CNN during train-time. As a testing-measure, we separate the predicted BOLD5000 fMRI into high-level categories, such as "animal" or "scene", via SVMs, and test the fitted SVM on unseen data including CIFAR10, Tiny ImageNet, and SUN images.

Additionally, we propose a novel validation measure where we assign course class labels based on network attention. Since BOLD5000 is composed of complex scenes, synthetic fMRI should correspond to the network's attention in the scene, and not necessarily the original label of the image. For example, consider Image B in Figure 2; this image depicts a woman applying lipstick, but the original dataset label is "lipstick". In cases where the network attention is focused on the lipstick, we relabel the image as "object," but in cases where the focus is on other parts of the image, we relabel the image as "not object." By hand-labeling 4,840 images, and reevaluating the SVM's predictions, we found that it improves results. Our results indicate that learned fMRI qualitatively maps images in a biologically consistent way.

Finally, as noted above, there is a long history of directly comparing neural network behavior to biological neural behavior in a computer vision context. Specifically, this is done by converting neural activation features into representational dissimilarity matrices (RDMs) and correlating biological and neural network RDMs [15]. We expect that a network producing plausible synthetic fMRI would exhibit high behavioral similarity with these traditional evaluations. Similarity ranges depend on the neural data used; we evaluated our network with human model similarity (HMS) [1] since the neural data was publicly available through the RDM toolbox [20] and the code for evaluation was publicly released.

The quantitative and qualitative validation measures in our experiments show real-world applicability in classification tasks, assuring that the neural activations obtained from CNN training exist in the same subspace as images of similar categories. By learning to simulate these activations using a multitude of training techniques, we can guide CNNs towards more biologically-grounded knowledge. These results suggest we are on the right path to understanding the multistage, non-linear nature of biological vision systems via deep learning models [2].

2. Related Work

BOLD5000 The BOLD5000 study recorded fMRI information from four human participants across 5,254 images (4,916 distinct) selected from ImageNet, SUN, and COCO datasets [4]. The study aimed to close the gap between computer vision and neuroimaging and promote statistical learning by addressing three major limitations in current research: (1) Dataset size. The success of modern computer vision is largely attributed to vast amounts of annotated data. (2) Diversity. Neurological studies often use only a small subset of possible images, with a small amount of categories and images per category; in contrast, computer vision datasets contain a diverse set of natural images across thousands of categories. (3) Image overlap. Neural studies often contain centered objects against white backgrounds, while computer vision datasets contain complex backgrounds such as natural scenes.

During the study, subjects participated in multiple image runs, where the subject was shown 37 images in succession. The study implemented a slow-event procedure, which consisted of showing 1 second of the stimuli image, followed by 9 seconds of a black screen with a cross, allowing precise matching of stimuli to the neural behavior. On top of gathering fMRI, half of the sessions included functional localizer runs in order to isolate specific regions of interest (ROIs). Areas associated with scene encoding, such as the PPA and the retrosplenial complex (RSC), were defined by comparing functional localizer runs with scene images against runs with scrambled and object-specific images. In this work,

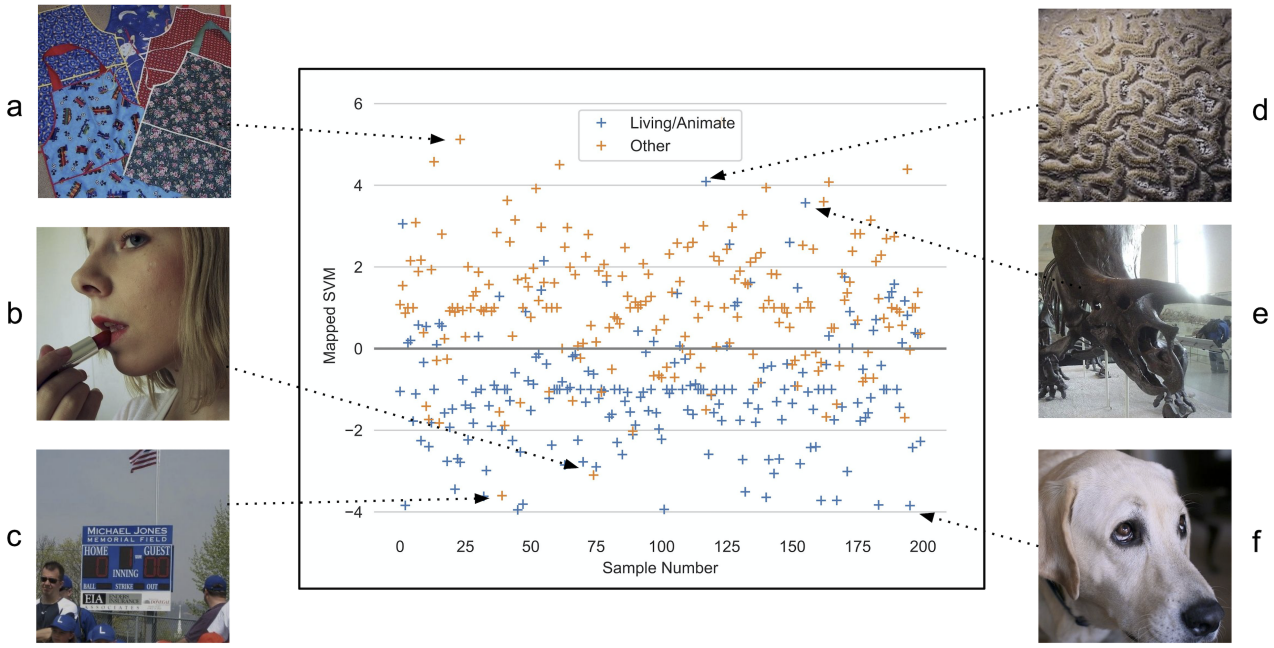

Figure 2: The above graph maps the SVM hyperplane separating images labeled 'living/animate' versus other ImageNet objects in fMRI ROI titled "RHPPA" in the first test subject in the BOLD5000 dataset. Image A shows an object classified correctly with high confidence. B/C show two images labeled as objects that are incorrectly classified as 'living/animate' with high confidence. Both pictures contain human faces, indicating neural responses were influenced by the human features of the image. D/E show images labeled 'living/animate' that were predicted as objects with high confidence. Image D is labeled 'brain coral', and image E is label 'triceratops'. Images F shows images predicted as 'living/animate' with high confidence. The figure highlights the intuitive classifications of well-labeled images as well as the ambiguous nature of incorrectly classified images. This analysis indicates noise in the fMRI data, while also confirming a general trend towards discriminability of distinct image classes and datasets.

we concentrate on the PPA ROI data, with additional ROI experiments included in the appendix.

Other Work Around the same time as Rosenblatt's perceptron [23], Hubel and Weisel discovered that neurons in the primary visual cortex (V1) behave selectively in response to image stimuli [12]. Recently, convolutional neural networks have displayed benchmark performance in predicting neural response patterns in primates[2] compared to long-established neurological models such as HMAX [22], GIST [21], and Gabor Filter Bank methods [14, 6]. These works highlight the inter-disciplinary work being done between the fields of neurology and computer vision.

While several computer vision works have found benefits when training models on biological data (as described in the intro), none have yet used the classification properties of human PPA to train images. PPA was first shown to encode spatial layout information by Epstein and Kanwisher [8], with some research suggesting that the PPA instead encodes context, defined as "information about objects which 'typically occur in the environment around us'" [9]. Evidence from BOLD5000 indicates that fMRI activations are able to separate not only scenes from objects, but also 'living/animate' objects from other objects.

Synthetic fMRI data has been generated mainly with two approaches: data driven synthesis and statistical modeling-based simulation. Data driven generative models are also explored in artificial fMRI synthesis, such as GAN and VAE networks [29, 25]. Statistically simulated fMRI is useful in experimental design evaluation and improving reproducibility [7, 11, 27], but only approximates limited properties of fMRI data.

3. Extended BOLD5000 Analysis

We expand on the work of BOLD5000 researchers by looking at the separability of BOLD5000 PPA data, which we use as our training set. We look at the separability of images labeled "living/animate" versus other objects, as well as "living/animate" versus images of scenes in Table 1. We

Data Type	Classification	SVM	ROC	PR
Original	Living-Other	Linear	.853	.851
		RBF	.805	.798
	Object-Scene	Linear	.720	.701
		RBF	.603	.605
Normalized	Living-Other	Linear	.882	.885
		RBF	.980	.980
	Object-Scene	Linear	.737	.723
		RBF	.986	.986

Table 1: Classification scores of BOLD5000 PPA data trained on two SVMs for subject one. These scores reflect the separability of the training set used in our experiments in the subsequent sections.

define 'living/animate' images by utilizing the WordNet hierarchy [19] to retrieve super-category labels for each image in the dataset. Our base class belongs to the WordNet ancestor "Living thing, animate thing", and we consider all other images in the BOLD5000 ImageNet repository as the second class. For our "Object-Scene" experiments we use all BOLD5000 ImageNet labels (excluding 28 images of scenes in the dataset), and classify against BOLD5000 SUN images. We remove 13 classes from the SUN data that might contain objects similar to those in the BOLD5000 ImageNet data (e.g. car factory, horse racetrack).

Table 1 shows the binary classification scores of BOLD5000 PPA data in the first test subject using linear and RBF SVMs. We evaluate binary classification using the area under the Receiver Operating Characteristic curve (ROC) and area under the Precision-Recall curve (PR). The table also shows that by normalizing each dimension of the target vector with respect to the entire set, we are able to discriminate the two classes of images with much higher confidence. As a result, all subsequent experiments use normalized fMRI ROI.

When evaluating the fMRI we also question whether the class separability made sense on an image-to-image basis. That is, do images of distinctive artificial objects align with images of other objects, and vice-versa? Alternatively, what are the properties of the images where fMRI data was classified incorrectly? Some examples of our analysis are shown in Figure 2. Results show that images of artificial objects sometimes contain humans, which may give the human viewer alternative activation patterns. Further, "living/animate" images classified as objects contained ambiguity: images containing subjects such as insects and brain coral were labeled as objects with high confidence.

4. Method

4.1. Classification

In our experiments we map an image x onto an fMRI Region of Interest (ROI) using a convolutional neural network architecture, which we denote $f(x)$. For experiments presented in this paper, we use the ROI entitled "RHPPA" in BOLD5000 for three subjects. Our normalized target vector, which we denote t, is size 200 for subject one, 198 for subject two, and 161 for subject three.

Our CNN model has a simple architecture consisting of 9 layers: 4 convolutional layers, 3 max-pooling layers, an average pooling layer, and a final linear layer. Further, ReLU activations are applied for non-linearity. We found that the simple architecture performed better with limited data. The final linear layer of $f(x)$ is a vector of size t. We train the model to minimize the L2 objective:

$$\frac{1}{|t|}\sum_{i=1}^{|t|}|f(x)_i - t_i|^2 \quad (1)$$

During test time, we employ a linear SVM and an RBF-SVM for binary classification. In section 5.1, we show results for the linear SVM, though the RBF-SVM reported very similar results.

SVMs are trained on BOLD5000 PPA data in section 5.1 (denoted in Table 2 as "real fMRI"), as well as the outputs of the BOLD5000 training images from the convolutional neural network (denoted "synthetic fMRI"). All SVM models are trained with balanced classes. Formally, we call the SVM model $h(y)$, where $y = PPA$ for results labeled "real fMRI," and $y = f(x)$ for results labeled "synthetic fMRI."

Inter-Object Classification For our first two experiments we consider a base class consisting of living things. We define images of living things in BOLD5000 by utilizing the WordNet hierarchy as described in Section 1. In all, we use 760 in-domain images and 1156 out-of-domain images for a total of 1916 examples. During training and test time, we perform random horizontal and vertical flips on the BOLD5000 images. Our BOLD5000 images are transformed to a pixel height and width of 32.

For our first test set we use CIFAR-10 [16], consisting of six animal classes and four vehicle classes. Our base class consists of animal classes, while our second class consists of the four remaining vehicle classes. We balance these sets so that we have 4,000 animals and 4,000 vehicles, which are passed to the SVM for classification during test time. Our second test set is Tiny ImageNet [13], which consists of 200 classes of 64x64 colored images taken from the ImageNet repository. Bounding box information was used for cropping each image, giving objects more prominence in the photo. Tiny ImageNet contains a WordNet hierarchy, giving us the ability separate the classes into "Living thing,

animate thing", and all other object classes. We use the validation set of Tiny ImageNet during test time, and select 2700 non-ambiguous images of animals and objects; for example, we remove images of objects that contain humans in them.

Living/Animate-Scene Classification For our next set of experiments we consider fMRI data taken from both the ImageNet and SUN datasets. While the ImageNet database consists of specific objects within pictures, the SUN database [28] classifies images based on the scene (e.g. tennis court, hospital, office, etc.). BOLD5000 uses 250 categories of scenes, attaining four images per category. Our final training set consists of 760 ImageNet images of Living/Animmate objects and 948 SUN images with their associated fMRI data. We remove 7 ambiguous SUN classes from the BOLD5000 data, such as "horse racetrack" and "aquarium", since the images may contain animals.

The base class for our test set consists of CIFAR-10 animals and TinyImageNet animals as described in the above section. We consider only living/animate images to classify against SUN images during test time. Our SUN test set consists of 1,946 images taken directly from the SUN dataset. Our SUN test images are distinct from the BOLD5000 SUN images because BOLD5000 researchers used Google Images to attain their SUN images. We use 141 categories from the SUN dataset that also belong to the BOLD5000 SUN categories, removing pictures containing people and object-specific features.

Standard Training For a baseline comparison, we train a CNN with standard classification techniques on our BOLD5000 training sets. To do this, we add a layer to the end of the CNN for binary classification and fit the model with cross-entropy loss and softmax. We train our SVM on the binary classification target output from the CNN vector for BOLD5000 images, i.e. $[1, 0]$ for "animals" and $[0, 1]$ for "other objects" in the case of experiment one. Then, we retain the binary outputs of CIFAR-10, Tiny ImageNet, and SUN test examples from our CNN and pass them into the SVM model. All test sets are normalized similarly to the training set, with images resized to 32 x 32 pixels. Results are included in the final two columns of Table 2.

While this model serves as a baseline to compare against our method of fitting images to fMRI via a CNN, it should be noted that during train-time BOLD5000 images were given their correct label from the WordNet hierarchy and SUN datasets. This results in direct labeling of images during training for our baseline experiment, whereas our fMRI-trained CNN is fit in an unsupervised manner that includes noisy activation patterns for ambiguous images.

4.2. Network Attention

Motivated by the visual explanations of CNNs, in this section we analyze the heatmap activations learned by our CNN mapped to fMRI data. Our approach is a modified version of the popular Grad-Cam algorithm [24], used to analyze the attention of CNNs. Similar work was done for diabetic retinopathy detection [26], though the use case was different.

Algorithm We denoise BOLD5000 PPA data by applying PCA to the target vector t prior to training such that $|t|$=1 for each example. (We performed experiments on the full vector fMRI data, rather than PCA-applied features, however, we were unable to find conclusive results in the analysis.) We fit our CNN to this single output vector and capture the heatmaps of images after each training epoch. We use the following steps to generate a heatmap for each image:

1. We feed a BOLD5000 image x to a CNN and capture the activations at the last convolutional layer, i.e. A_x^k, where the activation layer contains k feature maps.

2. We take the final regression output, $f(x)$, and backpropagate it to layer A_x to obtain the gradients $\frac{\partial y}{\partial A_x}$. Our resulting vector G contains gradients for k feature maps along with height and width channels i and j.

3. We take the mean over the height and width channels of G, giving us the gradient weights for each feature map, i.e.
$$\bar{w_k} = \frac{1}{IJ} \sum_i^I \sum_j^J G_{i,j,k} \tag{2}$$

4. Finally, we multiply the result $\bar{w_k}$ by A_x^k, and take the ReLU:
$$ReLU\left(\sum_k \bar{w_k} A_x^k\right) \tag{3}$$

The result is a vector of size (i, j) containing values which we can then map to the original image. In our model, the result vector is size 16 x 16, which we resize onto the original image to obtain a heatmap of relevant and irrelevant activations with respect to image x.

From the method described above, we generate heatmaps of BOLD5000 images during train time to analyze the attention of the CNN over a large number of images. We look at 4,840 images of animals and objects over 22 epochs selected from a model trained for 500 total epochs. During the generation of the heatmap, we also collect the static SVM prediction on the BOLD5000 fMRI as well as the prediction for the generated CNN output, which we label PL. For each of these metrics, a positive number indicates an object, while a negative score indicates an animal. We denote PL values by giving them a pseudo-confidence equal to their distance from the SVM hyperplane.

Our analysis deals with an important second component: the Heatmap Hand Label (HHL). We hand label each of our

		Animal-Object Classification						Animal-Scene Classification (SUN)					
		Real fMRI		Syn. fMRI		Baseline		Real fMRI		Syn. fMRI		Baseline	
Sub.	**Dataset**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**
1	C-10	.897	.889	.899	.895	.894	.897	.862	.861	.879	.873	.891	.894
	TinyIm.	.781	.778	.805	.813	.794	.780	.844	.833	.850	.837	.857	.861
2	C-10	.843	.827	.840	.815	.889	.843	.841	.842	.852	.847	.859	.807
	TinyIm.	.754	.763	.761	.750	.838	.844	.832	.825	.839	.831	.780	.798
3	C-10	.844	.849	.857	.854	.845	.788	.810	.812	.831	.828	.866	.866
	TinyIm.	.715	.705	.756	.743	.842	.847	.794	.785	.791	.789	.798	.810

Table 2: Classification scores for synthetically-generated fMRI on our test sets. On the left, we classify 'living/animate' versus 'objects' with images from the 'Dataset' column. On the right, we take 'living/animate' images from the 'Dataset' column, and classify them against the SUN dataset. While baseline classification performs better in 7 out of 12 experiments, our fMRI experiments contain noisy data where image labels are sometimes mapped to opposite classification values, as shown in Figure 2. Despite this, classification scores are competitive with the labeled baseline model. The top figure on the right shows shows AUROC curves for CIFAR-10 experiments classifying 'animals' versus 'vehicles' for each of the three test subjects, and the bottom figure shows TinyImageNet-SUN classifications for each of the test subjects.

4,840 images in the observation set, which can be either an A (Animal) or NA (Not Animal) in the case of animal images, and O (Object) or NO (Not Object) in the case of object images. We base our labels on the attention of the heat maps in the image. To support these hand labels, Heatmap Hand Label confidences are also provided, which can be 0.25 (not confident), 0.5 (could be either an A or an O), 0.75 (confident enough) and 1 (very confident).

A few conditions were set before hand labeling the images. In order for the hand label to be labeled correctly, either the maximum area of the heatmap should be on the animal or object, or the heatmap on the animal or object should be brighter than the other regions. If the heatmap condition is not satisfied, then the "Not" labels are used. For example, if the conditions are not met for an image of an animal, it is assigned the 'NA' (Not Animal) label.

One judge coded the full set of 4,840 images. A random subset of 200 images from the full dataset was coded by an additional judge, following the same set of established rules. Overall, on the evaluation subset, coders agreed with a Kappa of 0.93. Kappa quantifies inter-rater reliability accounting for chance, with a score of 1 indicating perfect agreement, 0 indicating no agreement, and -1 indicating perfect disagreement.

4.3. Human-Model Similarity

Human-model similarity (HMS) was proposed by [1] and we follow their methods. Briefly, neural network activation behavior was collected in response to 92 visual stimuli. The behavior of the network is distilled into a Representational Dissimilarity Matrix (RDM) by correlating the activations for each set of stimuli pairs. Previously, [15] collected human fMRI data for the same stimuli, which had already been distilled into an RDM for public release [20]. The neural network RDM and the human RDM were then correlated, resulting in the final HMS metric. On average, HMS tends to be relatively low due to the noise in fMRI — [1] trained 95 models with random hyperparameters and found that the average HMS was 0.11 (SD = 0.05), with the 10 best models having an average HMS of 0.18 (SD = 0.01). Despite this limited range, they founf that models with higher HMS scores have statistically significant higher task performance.

5. Results

5.1. Classification Results

By training our CNN to have PPA-like output activations, we show that the model is able to generalize to new images. Table 2 shows results for our linear SVM models trained on real fMRI, synthetic fMRI, and the baseline model.

We have the most success in separating CIFAR-10 animals from CIFAR-10 vehicles. Surprisingly, this result shows discriminability across an entirely novel dataset (i.e. one not based on the WordNet hierarchy or scene representations). We see the generalization of these results in Figure 3, showing high scores across all test subjects.

Both CIFAR-10 and TinyImageNet animal images were able to be classified against SUN images with high accuracy, while the model didn't perform as well when discriminating between Tiny ImageNet specific objects. Results in general were competitive with baseline, showing that despite the noisy activation patterns exhibited in BOLD5000 fMRI data, there is strong potential for machine learning models to learn high-level image classifications via biologi-

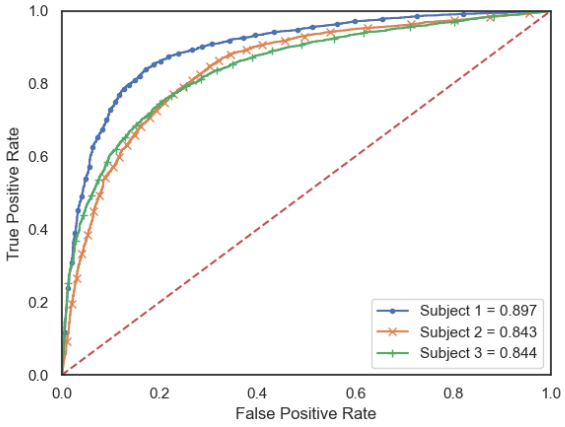

Figure 3: AUROC curve for CIFAR10 Animal-Object classification for each test subject. These results highlight the generalizable qualities of this method across multiple subjects.

	PL-C (0-0.15)	**PL-C (0.15-0.3)**	**PL-C (0.3-0.6)**	**PL-C (0.6+)**
No HHL-C	752/1242 (60.54%)	792/1232 (64.29%)	893/1219 (73.26%)	955/1147 (83.26%)
HHL-C $>$ 0.25	728/1212 (60.07%)	781/1220 (64.02%)	866/1190 (72.77%)	927/1117 (82.99%)
HHL-C $>$ 0.50	571/975 (58.56%)	640/1021 (62.68%)	681/944 (72.14%)	699/799 (87.48%)
HHL-C $>$ 0.75	371/634 (58.52%)	488/779 (62.64%)	464/633 (73.30%)	445/505 (88.12%)

Table 3: Accuracy of Predicted label [PL] vs Heatmap Hand label [HHL]. Results are calculated over an observation set of 4840 images. Each column has approximately 25^{th} percentile data (according to the PL-C range provided at the top of every column) so that the results can be shown over an even data distribution. PL-C is the Predicted Label Confidence, while HHL-C is the Heatmap Hand Label Confidence.

cal data. Finally, we did an overall comparison between the SVM prediction of the real fMRI versus the prediction of the synthetic fMRI. We compare the images used in Section 4.2 for our analysis. We observed that for 85% of these images, there was agreement between the real fMRI and synthetic fMRI in terms of the label assignment. Of the 15% of instances where the synthetic fMRI and real fMRI did disagree, 77% of the cases involved the real fMRI classifying images as "animals" while the synthetic fMRI classified the image as "object".

5.2. Network Attention Analysis

We train a CNN on BOLD5000 ImageNet images and look at the resulting heatmaps *during* training to validate progress. As a first step, we train an SVM on real fMRI data to predict 'living/animate' versus 'other' object. After each training epoch, we take the heatmap of an image as well as its synthesized fMRI output value. We then pass the synthesized output from the CNN into the SVM model and obtain its distance from the hyperplane. Naturally, as training of the CNN progresses, the SVM gets better at predicting the class of the synthesized outputs, as shown in Figure 4.

We perform a thorough analysis of this phenomena in Table 3, basing our analysis on both the predicted SVM label (PL) and the Heatmap Hand Label (HHL). For each image in the observation set, we assign value 1 if the PL matches the HHL, otherwise we assign 0.

Analysis of these trends show several things. First, by looking at the trend with respect to the HHL confidence, we can see that with increasing cut-off of the HHL, the overall accuracy of the model does not seem to vary by much. This goes to show that the confidence with which the hand labels are assigned does not contribute towards the overall performance of the model. We hypothesize this could be the result of two things: 1) a limited training set lacking thousands of objects per class or 2) noisy fMRI data with ambiguous classifications. Next, we look at the trend with respect to the PL confidence. We can see a direct correlation between increasing PL and accuracy. By combining this factor with HHL confidence, we see a general correlation between high PL confidence and high HHL confidence. While no HHL confidence has an accuracy of 83.26% with highly confidence PL, high HHL images have an accuracy of 88.12% with high PL images.

We also observed that as the training progressed, the Heatmap Hand Label accuracy increased consistently, showing that the trained model is able to learn where it should focus its attention within the image in order to correctly classify the images as objects or animals.

As a final validation measure, we generated heatmaps for the same images on randomized fMRI data, where our output feature was sampled from a random normal distribution. Results were around 50% for both HHL and PL .

5.3. Human Model Similarity

We calculated HMS in a multitude of different ways, comparing individual layers and the collective set of all layers. Overall, HMS varied early in training and then largely stabilized, matching the trend established in [1], although after stability HMS trended slightly downward. For the collective set of layers, the maximum HMS was 0.18, similar to top models found by [1], and comparable to pretrained wide resnet, which has a HMS of 0.20. The second convolutional layer had the highest HMS, 0.21, but each layer had competitive HMS, never dropping below 0.14 (around epoch 84).

One point of note was that the synthetic fMRI itself had

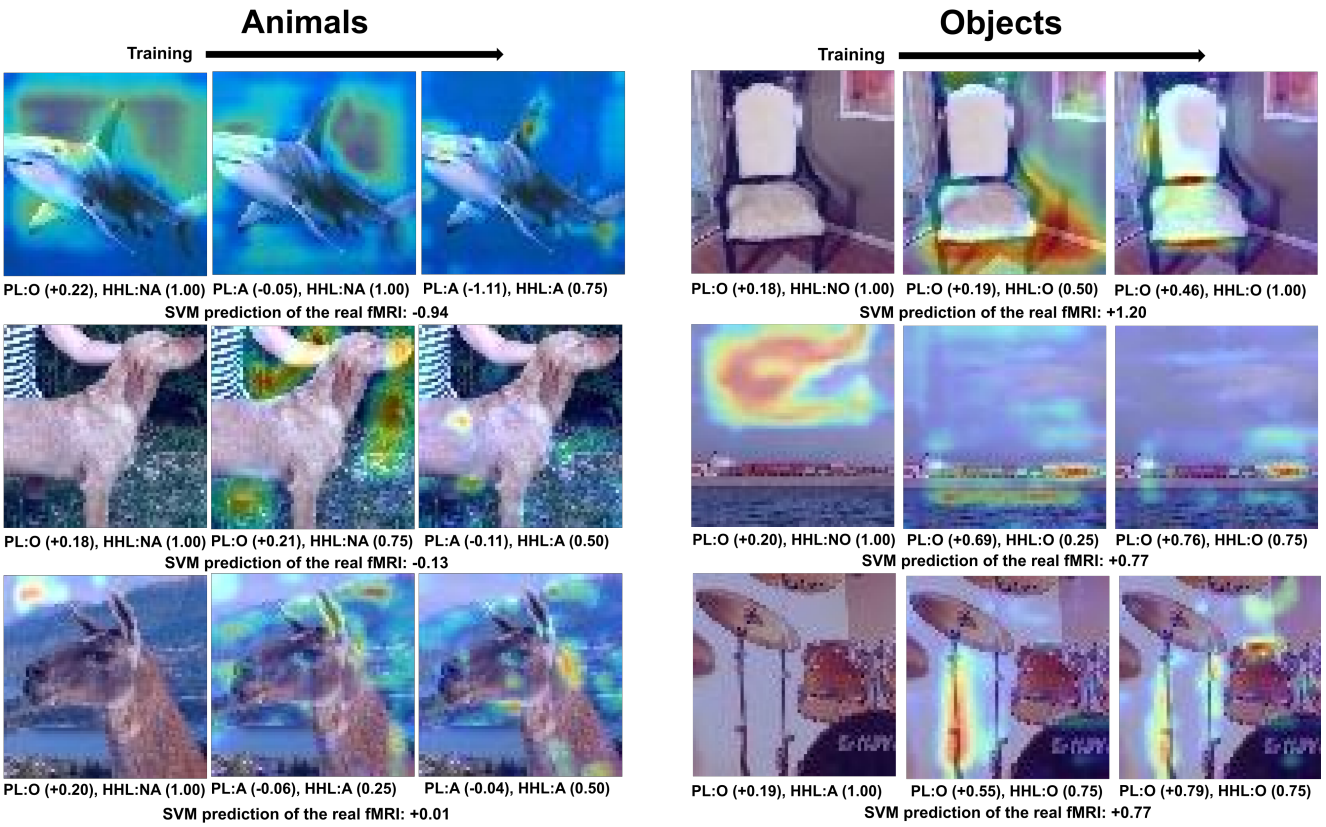

Figure 4: Heatmaps rendered over BOLD5000 images as training progresses. Denoted with each image is the Predicted Label (PL), which is the prediction made by the SVM on the synthesized output of our CNN. The number indicates distance from SVM hyperplane. We also include our hand-label confidence (HHL). Finally, with each set of images, we include the SVMs prediction on the real fMRI data. Each series of images start with an example that was predicted incorrectly, and as training progressed, improved in the isolation of animate and objects. Heatmap analysis shows that, despite noise, the synthesized fMRI was able to match its target in each image.

low HMS. The stimuli are explicitly selected to cover a wide range, but since we are training synthetic fMRI in a relatively limited context we would only expect our model to generate realistic fMRI on the domain it was trained for. This seems to be confirmed by the low HMS, indicating, unsurprisingly, that this technique would *not* generate accurate fMRI outside of the training domain. Further, this indicates that at least some fMRI will be required and that future collections are needed. For example, one set of stimuli used for HMS is faces — faces are not explicitly included as part of the BOLD5000 training set and we would not expect to, or want to, try to generate realistic fMRI for faces.

6. Conclusion and Discussion

While fMRI data presents meaningful information about the human brain, collecting this data is expensive and time-intensive, and the resulting data is noisy [18]. We present novel approaches for validating synthetic fMRI by categorizing the synthetic outputs via an independent classifier. By learning the expected data space of higher level image representations in fMRI regions of interest, we can understand what this synthetically-generated data should look like in humans.

In the area of computer vision, training with biological data is linked to greater model robustness [17], but this is just one application; we can easily see such data proving useful for other research, such as out-of-domain recognition, where the goal is detecting new examples outside of a trained models purview [10]. By learning to recognize the data manifold of higher-level classes as they may look within the brain, an fMRI model can conceivably be trained to filter out examples of classes outside of a model's domain.

While the work presented in this paper can guide classification tasks in models trained with biological data, important questions remain. First, as we refine our modeling of brain data, finer-grained classification can help in building

more robust models. Secondly, understanding and defining the fundamental properties of images that guide brain activations remains an open question in neurology. By disentangling these properties in images as well as fMRI, we can reduce the noise present in our models to gain a clearer picture of biological intelligence.

References

[1] Nathaniel Blanchard, Jeffery Kinnison, Brandon Richard-Webster, Pouya Bashivan, and Walter J Scheirer. A neurobiological evaluation metric for neural network model search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5399–5408, 2019.

[2] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[3] 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. 10(12):e1003963, 2014. Publisher: Public Library of Science.

[4] 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. 6(1):1–18, 2019. Number: 1 Publisher: Nature Publishing Group.

[5] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. page 2020.06.16.154542, 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[6] John G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[7] Cameron T. Ellis, Christopher Baldassano, Anna C. Schapiro, Ming Bo Cai, and Jonathan D. Cohen. Facilitating open-science with realistic fMRI simulation: validation and application. *PeerJ*, 8:e8564, Feb. 2020.

[8] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[9] Russell Epstein, Mary Smith, and Emily Ward. What is the function of the parahippocampal place area? testing the context hypothesis. 9(8):963–963, 2009. Publisher: The Association for Research in Vision and Ophthalmology.

[10] Matt Gorbett and Nathaniel Blanchard. Utilizing network properties to detect erroneous inputs. *arXiv:2002.12520 [cs]*, 2020.

[11] Jason E. Hill, Xiangyu Liu, Brian Nutter, and Sunanda Mitra. A task-related and resting state realistic fMRI simulator for fMRI data validation. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101332N. International Society for Optics and Photonics, Feb. 2017.

[12] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[13] Stanford Johnson. Tiny imagenet visual recognition challenge. pages 211–252, 2015.

[14] Judson P Jones and Larry A Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258, 1987.

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

[16] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. pages 211–252, 2009.

[17] 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 32*, pages 9529–9539. Curran Associates, Inc., 2019.

[18] Martin A. Lindquist. The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464, 2008. Publisher: Institute of Mathematical Statistics.

[19] George A. Miller. WordNet: a lexical database for english, 1995.

[20] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. 10(4):e1003553. Publisher: Public Library of Science.

[21] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[22] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. Number: 11 Publisher: Nature Publishing Group.

[23] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[25] Rufin VanRullen and Leila Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, 2(1):1–10, May 2019. Number: 1 Publisher: Nature Publishing Group.

[26] Zhiguang Wang and Jianbo Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. *arXiv:1703.10757 [cs]*, 2019.

[27] Marijke Welvaert and Yves Rosseel. A Review of fMRI Simulation Studies. *PLoS ONE*, 9(7), July 2014.

[28] 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. IEEE, 2010.

[29] Peiye Zhuang, Alexander G. Schwing, and Sanmi Koyejo. FMRI data augmentation via synthesis. *arXiv:1907.06134 [cs, eess]*, pages 1783–1787, July 2019. arXiv: 1907.06134.