Learning Perceptual and Quasi-Perceptual Representations from fMRI

Abdul Wahab⋆**, Muhammad Kamran Janjua**⋆
University of Alberta, Canada
{wahab1, mjanjua}@ualberta.ca

Abstract

Human visual system depends on imagery representations in addition to the visual input to build a concrete holistic picture of the object of interest. These representations are correlated and shared in the human brain and help provide a crisp understanding of the environment. In this work, we study if a similar correlation exists in Artificial Neural Networks (ANNs) and if the correlation is strong enough to use the corresponding fMRI data of both perceptual and quasi-perceptual experiences interchangeably for downstream tasks such as object category prediction. The codebase of this entire project can be accessed here: https://github.com/mirzaabdulwahab1612/interpreting-fMRI-signals.

1 Introduction

A human brain is capable of producing novel ideas, objects, sounds, etc., without any explicit input from the senses. This phenomena is called imagination, and it is rather intuitive to think that the imagined representations we have inside the brain are similar to what we actually experience through our senses. Mental imagery is a form of a perceptual experience that a subject undergoes in the absence of any visual stimulus. Studies have concluded that this form of mental imagery is 'quasi-perceptual experience' which functions as some sort of mental representations [15].

In 1910, Perky performed the famous 'Banana Experiment' which was the first of its kind to explore the link between imagery and perception [13]. Without knowing that experimenters were projecting a visual stimulus with increasing intensity onto the screen, the subjects mistook this visual projection as a product of their imagination and actually modified the details related to the stimulus such as veins in a leaf, title of a book in accordance with the projection. The projected stimulus modified the subjects' experience. They reported that banana is vertical and not horizontal as they had previously imagined as the stimulus projection intensified. Perky concluded that since participants reported such vivid imagination pertaining to the projections that there was some link between mental imagery and visual perception. This is known as the Perky effect which states that 'the tendency of an imagined stimulus to interfere with seeing an actual target stimulus when the imagined form is close to that of the target' [13]. The Perky effect then went on to garner much interest from scholars throughout the 20th century.

However, if this mental imagery is in fact similar to visual perception has had psychologists divided for decades. This divide allowed for and encouraged a shift towards empirical studies and a functional or operational understanding of the concept of imaginary representations since that narrowed the scope of understanding the existence of such a phenomenon and its role without having to explicitly define any structure or manner. With the advent of imaging apparatus such as functional Magnetic Resonance Imaging (fMRI) which is the measurement of blood flow changes in accordance to the brain activity, scholars started to functionally analyze how mental imagery and visual perception correlated with each other, if at all, by qualitative and quantitative analysis of fMRI response obtained.

⋆ indicates equal contribution.

In [4], fMRI data is used to find the degree of shared neural processing in visual mental imagery and visual perception. It was observed that the visual mental imagery and visual perception draw on most of the same neural machinery. The authors concluded that cognitive control processes function comparably in both imagery and perception, whereas some sensory processes may be engaged differently by visual imagery and perception.

The ability of the human brain to imagine visual scenes, and their close similarity to perceptual visual scenes suggests a correlation in the neural mechanism of the two. In [1], the authors studied this correlation by applying multivariate pattern classification on a combination of fMRI data. They conclude that the category representations are shared in the high-level ventral cortex, while perception and imagery share representations of object location in low-level visual cortex. These results indicate and hint towards the existence of some correlation between imagery and perception representations. Using fMRI and harnessing the power of deep neural networks while using such pattern recognition algorithms capable of tuning and attending to entire spatial response map has allowed for a varied and advanced understanding [2, 12] of what goes on in the human brain.

In this study we want to explore if this correlation is strong enough to use fMRI data corresponding to visual mental imagery and visual perception inter-changeably for downstream tasks like object category prediction. This would mean that if a cognitive model trained on either imagery data or perception data, for some downstream task would do well taking the other representation as input as well. Since there exists a correlation in the imagery and perception and we know from empirical research that these representations are somewhat similar and both perception and imagery trigger same neural machinery, we quantify if a artificial neural network (ANN) is able to learn a representation that can work for both since they are similar in the brain [4]. Eventually, we perform post-training adhoc analysis on the learned representations, going beyond the signal averaging mechanisms by actually classifying the objects given imagery and perceptual representations as input i.e. the previous approaches used averaged representations for predictions while we make predictions based on individual data point.

We conclude that there exists some correlation between perception and imagery representations, and that a shared representation can be learned by training these two categories of data together which encourages the model to learn a shared representation. Quantitatively, we report above chance accuracy results on each perception and imagery held-out sets given an ANN trained on combination of both perception and imagery. Furthermore, we also conclude that pre-processing in addition to averaging the data across volumes harms the imagery data.

2 Related Work

In this section we study the previous and related works under different sub-categories. We first discuss the neurological correlation of visual and imagery representations, and the neural mechanism inside the brain under shared cortical representations. We then discuss different fMRI decoding approaches in practice for perception and imagery stimulus.

2.1 Shared Cortical Representations

In [1], the authors studied the correlation between imagery and perception by applying multivariate pattern classification on a combination of fMRI data. They studied whether imagery (brain representations when an object is imagined) and perception (brain representations when an object is seen) share the object category and object location representations in category-selective regions. The authors perform this analysis by training a Support Vector Classifier (SVC) that discriminates between the activation parameters (input) of different object categories. The SVC is trained using the activation parameters evoked by imagery of different object categories, and is tested on activation parameters evoked by perception of similar object categories. The main intuition behind this setting is that if imagery and perception share representations, i.e. neural mechanism inside the brain, then they should evoke similar fMRI response. Thus training on activation patterns evoked during imagery and then testing on activation patterns evoked during perception would amount to testing whether imagery and perception share representations or not. Although this approach can show to some extent that the perception and imagery share the same representations, however using the discriminating accuracy of the SVC alone does not explain how, or to what extent, and in which parts of the brain are these representations shared. Further analysis performed by the authors for each category pair in

each category-selective region, reveals that the category representations are shared in the high-level ventral cortex, and not in the low-level visual cortex. In contrast, perception and imagery share representations of object location in low-level visual cortex.

Similar work to determine to what extent item-specific information about complex natural scenes are represented, and in which category-selective regions of the brain, during perception and imagery is done in [9]. The authors performed a multi-voxel pattern analysis using a SVC to determine if item-specific information was represented in different scene-selective regions of the brain during perception and imagery. The authors also performed analysis to determine to what extent the imagery representations in the scene-selective regions is a re-instantiation of the previous perception representations in those regions.

We believe that this work provides a good explanation and argument on whether and how imagery and perception share representations in the brain. We would like to build on this idea that imagery and perception do share representations, and would like to extend this idea to ANNs. In this work we first analyze whether the correlation in imagery and perception is strong enough to use the two interchangeably for down stream tasks like object category prediction. We go one step further from the previous work and learn a shared representation that encompasses both perception and imagery representations.

2.2 Decoding fMRI

The ability of machine learning and specifically deep learning, to learn powerful representations for a given signal, has allowed decoding of the brain signal (fMRI) for different tasks. This includes analysis of the fMRI for different activities including response to visual and sound stimuli, and more over to analyse and decode neural activities evoked when a person is imagining or dreaming. In [6], the authors propose a methodology for the object classification tasks of both visual and imagery data using fMRI input. Instead of using a standard decoding approach that predicts the object category labels, the authors have proposed a decoder that instead predicts a visual representation, similar to what a CNN architecture would learn in an image classification task. The authors argue that decoding the fMRI to specific labels constrains the ability of the decoder to only those classes observed during training. The authors are more interested in getting more general feature representations, retinotopically organized image level features [10], which can then be used for the task of classification. The authors have used a combination of visual features including features learned at different layers of a CNN architecture, HMAX(1-3), GIST, and SIFT + BoF, to serve as the target for the decoder. Finally, after the decoder is trained, the fMRI of seen or imagined object is used as input to the decoder which predicts the visual feature representation for the fMRI and then classifies the object category by calculating the similarity of the decoded visual features with the means of visual features for each category in the object category database.

In [7] a similar approach was extended to study whether brain activity evoked while dreaming can be decoded to visual features, representations in different layers of a CNN, which can then be used for the category prediction task. The authors show that the decoded visual features from the dream fMRI data positively correlated with the corresponding visual features of the dreamed category at mid and high-level CNN layers. The authors also show that using the dream decoded features, the object category can be predicted at above-chance levels by comparing them with the averaged features for each category in the database.

In [14] the authors test whether the correlation in perception and imagery is strong enough to use the two interchangeably for object category prediction. They use a SVC classifier and test their hypothesis using the following experiments: 1. Train on perception data (P), test on P. 2. Train on P, test on Imagery data (I). 3. Train on I, test on I. 4. Train on I, test on P. The authors performed the experiments using fMRI data corresponding to 4 classes.

In this work we use individual fMRI scans for the prediction of object category as opposed to [6], in which decoded features are averaged across classes before making predictions. We also train and test an ANN based classifier instead of using a similarity/distance based classifier. We extend the idea in [14] to 50 classes, and also explore a simple method to explicitly learn a shared representation in ANNs that corresponds to both imagery and perception.

3 Dataset

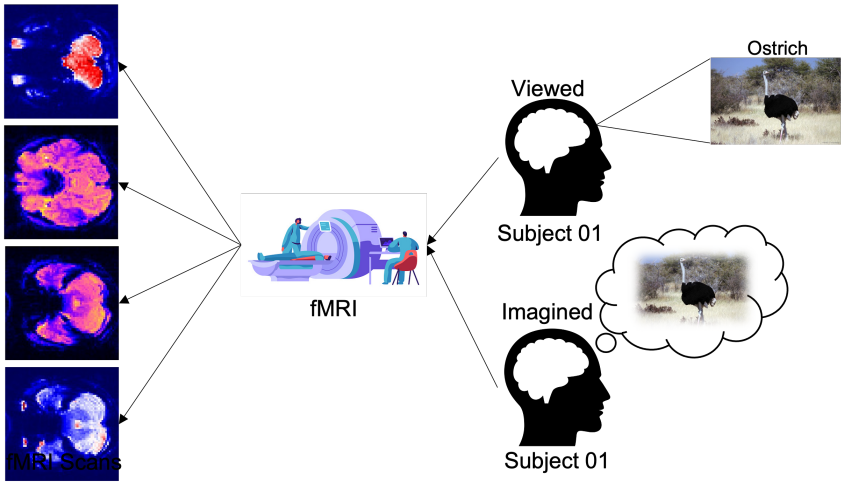


Figure 1: The data collection process. The subject first views an image (perception) of a visual stimuli and then imagines (imagery) the object and each time fMRI is recorded to obtain the brain scans.

The primary data we use in this project is the fMRI data of perceived and imagined object categories used by the authors in [6]. The fMRI measures the neural activity in the brain by detecting changes in the blood flow in different regions of the brain. The blood flow is coupled with the neural activity of the brain, thus when a specific region of the brain is in use, the blood flow of that region also increases [11]. The fMRI is widely used today to measure the neural activity of brain in humans and animals because of its non-invasive nature [8]. The object categories used for perception and imagery tasks are taken from the Fall, 2011 release of ImageNet data [3]. The object categories range from animate to inanimate objects i.e. has a wide variety of objects. For visualization purposes, some samples are given in Figure 2.

The fMRI dataset we used was collected from five healthy individuals (one female and four males), see the data collection process in Figure 1. The data collection experiment was performed by showing the individual subjects images of the selected categories from the ImageNet data. The brain representations evoked during this stimuli were collected using fMRI (perception fMRI). The subjects were then also told to imagine objects from the selected categories while the fMRI (imagery fMRI) was being recorded. The dataset was collected in multiple sessions for training, testing and for imagined object categories. The subjects were also required to do a one-back repetition detection task on the images, to keep the subjects focused on the visual stimuli. For the perception experiment the subjects were shown each image for a total of 9 seconds and the TR for fMRI acquisition was 3 seconds. This means that a total of 3 fMRI scans were acquired corresponding to one image. These 3 scans (volumes) were averaged to get a single fMRI scan corresponding to one image. In case of imagery experiment the subjects were asked to imagine an object for a total of 15 seconds and the TR was again 3 seconds. A total of 5 fMRI scans (volumes) were acquired corresponding to one imagined object. These 5 scans were also averaged to get one fMRI scan corresponding to one imagined object. The raw fMRI data collected during the imagery and perception experiments underwent a series of pre-processing steps, i.e., 1. Three-dimensional motion correction using SPM5. 2. Co-registered to the within-session high-resolution anatomical image. 3. The co-registered data were then re-interpolated. 4. Voxel amplitudes were normalized relative to the mean amplitude of the entire time course within each run. The authors of the paper [6] also performed separate experiments to identify different ROIs, and made masks corresponding to each ROI. The authors used the following data distribution in their experiments for object category prediction.

1. **Perception fMRI Training Data:** A total of 1200 images from 150 categories of ImageNet were shown to the 5 subjects.
2. **Perception fMRI Testing Data:** A total of 50 images, 1 for each of the 50 selected object categories were shown to the 5 subjects 35 times.

3. **Imagery fMRI Testing Data:** The 5 subjects were told to imagine the object category for the similar 50 selected categories used during testing for 10 times.

Figure 2: Some sample images from the ImageNet [3] dataset viewed and imagined by subjects in the fMRI dataset this work uses. The names of categories from top-left to bottom-right are as follows: *Ostrich, Snake, Dolphin, Old Telephone, Jacuzzi, Palm Tree*.

We only use Perception fMRI Testing Data and Imagery fMRI Testing Data in our experiments to keep the object categories consistent in perception and imagery data. The fMRI data is prone to a lot of noise as there are many other phenomenas the brain is attending to. The other issue with the fMRI data is the high dimensionality. For this reason we monitor specific ROIs for the project. This is done by applying the masks provided by the authors of [6] and using ROI specific data only. In this study we use VC ROI since it covers the entire visual cortext (V1 - V4 + HVC).

4 Methodology

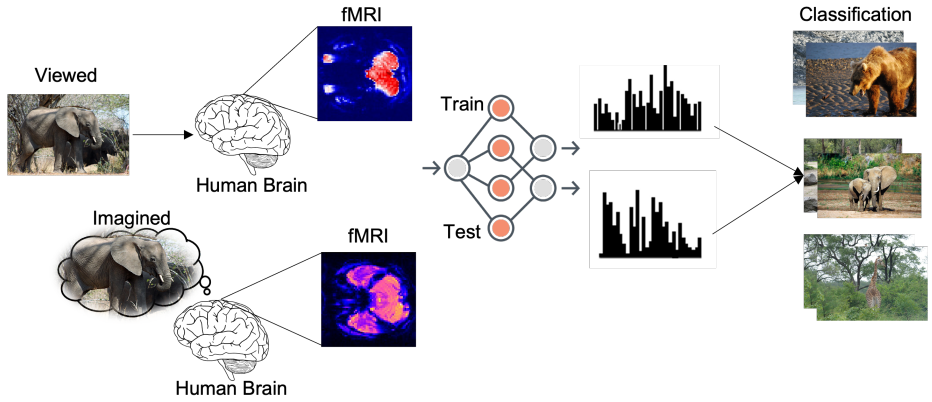


Figure 3: The one-picture idea to qualitatively illustrate the overall process of learning the representations. The viewed (perception) and imagined (imagery) fMRI is fed to the artificial neural network (ANN) to learn the representations and then a classification head learns to map these representations to classes for object classification.

From the literature we learn that quasi-perceptual (visual imagery) representations inside the brain are correlated with the perceptual representations, since visual imagination and perception stimulus evoke the same neural machinery inside the human brain. In this section we investigate whether this correlation is strong enough to use the two types of representations interchangeably for down stream decoding tasks like object category prediction. Also, can a shared representation be learned in

ANNs which corresponds to both imagery and perception. We would first discuss the formulation we developed, i.e., the inputs and outputs of the classification model. We will then discuss the details of the inputs and outputs. Lastly, we discuss the different decoding methodologies used for object category prediction using fMRI data.

Our methodology is to use the fMRI data corresponding to perceptual and imagery stimulus to predict the category of the object subjects had seen or imagined during the acquisition of fMRI. Thus the input of our predictive model is the fMRI data, and the output is the object category label. The input data is classified into two sets, i.e., 1) Perception fMRI, which corresponds to the fMRI data collected while the subjects were shown images corresponding to a specific category. 2) Imagery fMRI, which corresponds to the fMRI data collected while the subjects were asked to imagine a certain object.

To test our hypothesis, i.e., whether imagery and perception fMRI data can be used interchangeably for down stream tasks, object category prediction in this case, we follow a similar experimental setting as used in [14]. We however do another type of experiment in which we combine perception and imagery data while training and then test separately on perception and imagery held out data. We do this to test if we can explicitly make the model learn a shared representation that corresponds to both perception and imagery. We use three different types of classification models as follows: 1. Logistic regression model 2. 8-layer FC Network. 3. 2-layer MLP model. We tested these models on three types of datasets, 1. Pre-processed fMRI data [6]. 2. Raw fMRI data. 3. Decoded visual features data [6]. The summary of the experiments is given in Table 1. Initially we tested the three

Models $\rightarrow$	**LogReg | 8-Layer FC Net | 2-Layer MLP**	
Data Type $\downarrow$	***Train***	***Test***
Pre-processed fMRI Data	Perception	Perception
	Imagery	Imagery
	Perception	Imagery
	Perception + Imagery	Perception
	Perception + Imagery	Imagery
Raw fMRI Data	Perception	Perception
	Imagery	Imagery
	Perception	Imagery
	Perception + Imagery	Perception
	Perception + Imagery	Imagery
Decoded Visual Features	Imagery	Imagery

Table 1: Summary of train and test setting used for each of the models (Logistic Regression, 8-Layer FC Net, and 2-Layer MLP). The data types are also given. In this work, we used three data types, namely: pre-processed fMRI data, raw fMRI data and decoded imagery to visual features.

prediction models on the pre-processed fMRI data provided by the authors of [6]. One important thing to note is that this pre-processed fMRI data was averaged across the multiple volumes, i.e., 3 and 5 corresponding to perception and imagery data respectively. We also performed the experiments on the decoded visual features data using the approach proposed in [6], and on raw fMRI data without any pre-processing and averaging of volumes. The decoded visual features data is extracted by decoding the fMRI data to visual features, using sparse linear regression and the actual visual features corresponding to the image samples for that object category as target [6].

4.1 Experiment settings

This section discusses the train-test splits used for training and evaluating the models. As discussed above, we use three different types of datasets. There are a total of 1750 perception data samples, and 500 imagery data samples corresponding to 50 object categories for pre-processed fMRI case. In raw fMRI's case, there are a total of 5775 perception data samples and 2500 imagery data samples. In decoding case, we follow the approach outlined in [6] and only decode imagery to visual features and therefore the total data samples are 500 in this case. We split the data into training and held-out test set which is around 67% for training and around 33% for the test set. The complete summary of dataset split for each type of the data is given in Table 2.

Data Type ↓	***Train Setting***	***# of Samples***	***Test Setting***	***# of Samples***
Pre-processed fMRI	Perception	1172	Perception	578
	Imagery	335	Imagery	165
	Perception	1172	Imagery	165
	Perception + Imagery	1507	Perception	578
	Perception + Imagery	1507	Imagery	165
Raw fMRI	Perception	3869	Perception	1906
	Imagery	1675	Imagery	825
	Perception	3869	Imagery	825
	Perception + Imagery	5544	Perception	1906
	Perception + Imagery	5544	Imagery	825
Decoded Visual Features	Imagery	335	Imagery	165

Table 2: Summary of train and test splits for each of the three data types. Note that the test-set is a held-out test set.

4.2 Hyperparameters

In this section, we discuss the hyperparameters for the models we used. Our baseline model is Logistic Regression (LogReg) and we use l2 penalty and LBFGS solver. LBFGS stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno and is an optimization method which approximates second derivative matrix updates to optimize an objective.

In case of 8-layer FC Net and 2-layer MLP we initialize the weights following Kaiming He-normalization [5] which initializes the weights following a normal distribution where the standard deviation is $\sqrt{\frac{2}{n}}$ since it takes the non-linearity of the activation functions into account. We use ReLU activation in the hidden layers and alternate the layers with Dropout of 0.5 and BatchNormalization, since that is the standard in artificial neural network design. The last prediction layer has 50 neurons, one per class and softmax is applied at the end to convert the last layer's output to a 50 class probability distribution. The loss function is Sparse Categorical Crossentropy which is also a standard for multi-class classification problems. The optimizer used is Adam with a learning rate of $1e^{-4}$. We did not perform a grid search over the hyper-parameters in this study, and leave it for future work. The summary of all these parameters is given below.

Hyperparameters Summary for 2-layer MLP model & 8-layer FC Net model

1. Weights initialized with He-normalization
2. ReLU activation in the hidden layers
3. Dropout with 0.5 probability
4. Batch normalization
5. Final prediction layer with 50 nodes and Softmax activation
6. Loss function: Sparse categorical crossentropy
7. Optimizer: Adam
8. Learning rate: $1e^{-4}$

5 Experiments & Results

We conducted exhaustive experiments to test our hypothesis that if there exists a correlation in imagery and perception, and if ANNs are able to learn a representation that can work for both imagery and perception since they are similar in the brain [4]. The train and test settings are explained and summarized in Table 1. We conduct experiments for each of the data type and since there are three models, we get a total of 15 experiments where there are 5 settings and 3 models for pre-processed and raw fMRI data type and only 3 in case of decoded visual features. We conducted the experiments

on pre-processed and raw fMRI data for a total of 3 subjects while the experiment on decoded visual features was conducted for only 1 subject.

5.1 Pre-processed fMRI Data

Train ↓	***Test*** ↓	**LogReg**			**8-layer**			**MLP**		
		Test Acc (%)			Test Acc (%)			Test Acc (%)		
		S01	S02	S03	S01	S02	S03	S01	S02	S03
Pt.	Pt.	96	75.6	94.2	54.6	36.15	67.13	71.7	58.1	85.6
Im.	Im.	0.0	1.2	3.6	0.0	1.2	1.2	1.2	1.2	4.2
Pt.	Im.	2.4	1.8	6.0	1.2	4.84	6.0	1.8	4.2	4.8
Pt. + Im.	Pt.	91.0	64.1	90.3	55.1	32.6	60.2	63.8	48.7	79.7
Pt. + Im.	Im.	0.0	1.2	4.2	2.4	3.6	4.2	1.2	3.0	4.2

Table 3: Results on pre-processed fMRI data for 3 subjects. The first model is baseline, Logistic Regression, the second model is 8-layer FC Net (written as 8-layer for space restrictions) and the last model is 2-layer MLP (written as MLP for space reasons). The S01, S02, S03 represents Subject 01, Subject 02 and Subject 03 respectively. Perception is represented by Pt. and Imagery is represented by Im., this is done due to space constraints. Best results are colored blue row-wise, this is because we focus on best results in each train-test setting as described.

Table 3 quantifies the result of experiments on pre-processed fMRI data type. We performed the experiments for 3 subjects using each data type as input to all 3 of the models. We can see that all the results are consistent across all subjects, with little to no variation in the trend. The highest accuracy score is achieved by Logistic Regression model trained and tested on perception data i.e. 96% accuracy on held-out test set. In the case of trained and tested on imagery, the highest result is achieved by 2-layer MLP model using subject 03's data i.e. 4.2%, this score is not significant. In case of trained on perception and test on imagery, Logistic Regression and 8-layer FC Net achieves 6% test accuracy on subject 01 and 03's data. In case of training on combination of perception and imagery and testing on perception, the highest score is achieved by Logistic Regression on subject 01's data i.e. 91% whereas when tested on imagery data the highest score 4.2% is achieved on subject 03's data by all three models.

From these results, we can conclude that the model is able to learn classification task on perception data and there is little to no learning on the imagery data. Note that this data is pre-processed and we hypothesize that pre-processing might not be in the favor of imagery data. We discuss more on this in the next sub-section.

5.2 Raw fMRI Data

Train ↓	***Test*** ↓	**LogReg**			**8-layer**			**MLP**		
		Test Acc (%)			Test Acc (%)			Test Acc (%)		
		S01	S02	S03	S01	S02	S03	S01	S02	S03
Pt.	Pt.	14.7	12.3	10.5	4.4	3.9	4.0	10.8	9.6	8.1
Im.	Im.	51.6	31.2	31.8	11.5	6.0	6.3	34.9	23.3	25.4
Pt.	Im.	2.1	2.9	1.4	1.8	1.7	2.6	2.9	2.6	2.4
Pt. + Im.	Pt.	14.4	11.8	10.7	4.5	3.9	4.0	8.3	8.2	7.5
Pt. + Im.	Im.	32.9	21.6	25.6	11.3	5.9	6.9	26.6	17.5	21.7

Table 4: Results on raw fMRI data for 3 subjects. The first model is baseline, Logistic Regression, the second model is 8-layer FC Net (written as 8-layer for space restrictions) and the last model is 2-layer MLP (written as MLP for space reasons). The S01, S02, S03 represents Subject 01, Subject 02 and Subject 03 respectively. Perception is represented by Pt. and Imagery is represented by Im., this is done due to space constraints. Best results are colored blue row-wise, this is because we focus on best results in each train-test setting as described.

Table 4 quantifies the result of experiments on raw fMRI data type. We performed the experiments for 3 subjects using each data type as input to all 3 of the models, just like in previous case where the data type was pre-processed fMRI. We can see that all the results are consistent across all subjects, with little to no variation in the trend. The highest accuracy score is achieved by Logistic Regression model when trained and tested on perception data i.e. 14.7% accuracy on held-out test set. In the case of trained and tested on imagery, the highest result is achieved again by Logistic Regression model i.e. 51.6%. In case of trained on perception and test on imagery, Logistic Regression and 2-layer MLP net achieves 2.9% test accuracy on subject 02 and 01's data, respectively. In case of training on combination of perception and imagery and testing on perception, the highest score is achieved by Logistic Regression on subject 01's data i.e. 14.4% whereas when tested on imagery data the highest score is achieved again by Logistic Regression i.e. 32.9%.

From these results, we can conclude that the model is now able to learn from both perception and imagery. The experiment where we train on a combination of perception and imagery and individually test on each explains that the models are able to build a shared representation of both perception and imagery. The results are encouraging, although there could be improvements made. We assume that this is because this data is raw fMRI and there is no pre-processing involved. As it is observed in Table 3, when pre-processed perception fMRI data is fed to the model, it achieves accuracy of over 90%. Therefore we hypothesize that pre-processing for perception data could be essential. However, it is interesting to note that when same pre-processing is applied on imagery data, the model is unable to learn anything, same Logistic Regression returns 0.0% accuracy when trained and tested on pre-processed imagery. In future, we would like to investigate this further and work out the exact reasons of this behavior.

5.3 Decoded Visual Features

Train ↓	***Test*** ↓	**LogReg**	**8-layer**	**MLP**
		Test Acc (%)	Test Acc (%)	Test Acc (%)
		S01	S01	S01
Dec. Im.	Dec. Im.	0.0	1.2	1.2

Table 5: Results on decoded imagery to visual features for only 1 subject. The first model is baseline, Logistic Regression, the second model is 8-layer FC Net (written as 8-layer for space restrictions) and the last model is 2-layer MLP (written as MLP for space reasons). The S01 represents Subject 01 and decoded Imagery is represented by Dec. Im., this is done due to space constraints. Best results are colored blue row-wise, this is because we focus on best results in each train-test setting as described.

We performed another experiment as suggested by [6] where the imagery fMRI representations are fed to Sparse Linear Regression model and are decoded to visual representations from convolutional neural network (CNN) layers and other classical feature extractors, i.e., SIFT, HMAX, GIST. We decode the imagery fMRI data voxel by voxel, following the design set-out by the work done in [6]. Given these decoded features, we then learn a 50-way classifier to see if decoded imagery representations are indicative enough. The results are illustrated in Table 5. We only trained for one subject since the results were not significant or satisfactory. In this case, Logistic Regression is not able to learn and returns 0.0% accuracy on the held-out test set (which is of those decoded imagery fMRI). However, 8-layer FC Net and 2-layer MLP only return 1.2% accuracy scores on the held-out test set. Therefore, we conclude that this setting is not preferable in learning a multi-class object classifier.

5.4 Visual Illustration of Results - Raw & Pre-processed fMRI Data

In Figure 4, the visual results of training 3 models on pre-processed fMRI data are shown. There are total 5 experiments performed, see Table 3 namely: trained on perception (pt.), tested on perception (pt.), trained on imagery (im.), tested on imagery (im.), trained on perception (pt.), tested on imagery (im.), trained on perception and imagery (pt. + im.) and tested on perception (pt.) and imagery (im.) respectively. The bar plots qualitatively illustrate these results. These results indicate that all the models trained on imagery, in any combination, do not give any results whereas there is a striking difference between the results of these models when trained on perception. We assume that

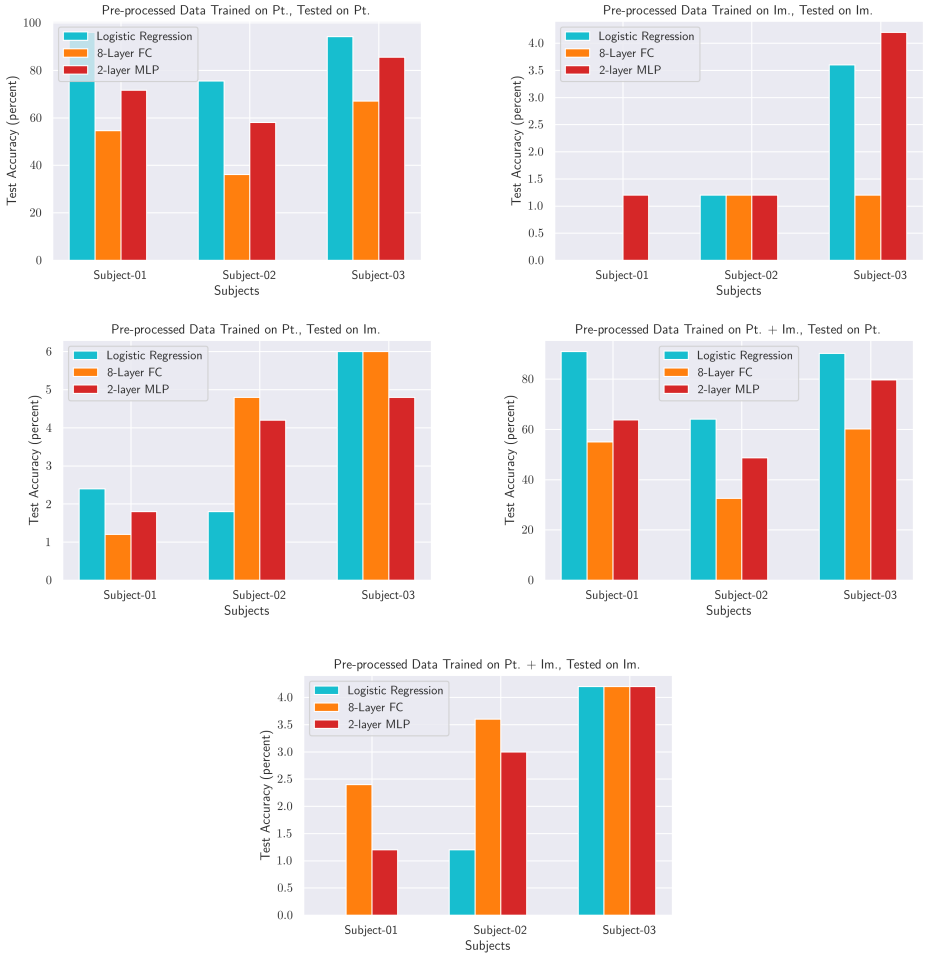


Figure 4: The qualitative analysis of all the models trained on pre-processed fMRI data type for each subject. Starting from top-left, the first plot is trained on perception (pt.), tested on perception (pt.), the next plot is trained on imagery (im.), tested on imagery (im.), while the next plot is trained on perception (pt.), tested on imagery (im.). The last two plots are trained on perception and imagery (pt. + im.) and tested on perception (pt.) and imagery (im.) respectively. Note that these bar plots are visual illustration of Table 3. Logistic Regression is represented with cyan color, whereas 8-Layer FC Net is orange and 2-Layer MLP is red. Best viewed in color.

this is because in perception test fMRI data [6] (the one we use as our training set. This is done to maintain similar classes in case of both imagery and perception, see Section 3), there is only 1 image per class (out of 50 classes) and each of the image is shown to each participant around 35 times. This consequently helps in getting indicative fMRI representations in case of perception, whereas in case of imagery the subject could have imagined a different shape/form/color of the same object in repeated experimentation rending less indicative imagery representations.

In Figure 5, the visual results of training 3 models on raw fMRI data are shown. There are total 5 experiments performed, see Table 4 namely: trained on perception (pt.), tested on perception (pt.), trained on imagery (im.), tested on imagery (im.), trained on perception (pt.), tested on imagery (im.), trained on perception and imagery (pt. + im.) and tested on perception (pt.) and imagery (im.) respectively. The bar plots qualitatively illustrate these results. These results on raw fMRI data are most promising since models trained on combination of perception and imagery are able to learn a shared representation which allows multi-class classification when tested on held-out sets of both imagery and perception.

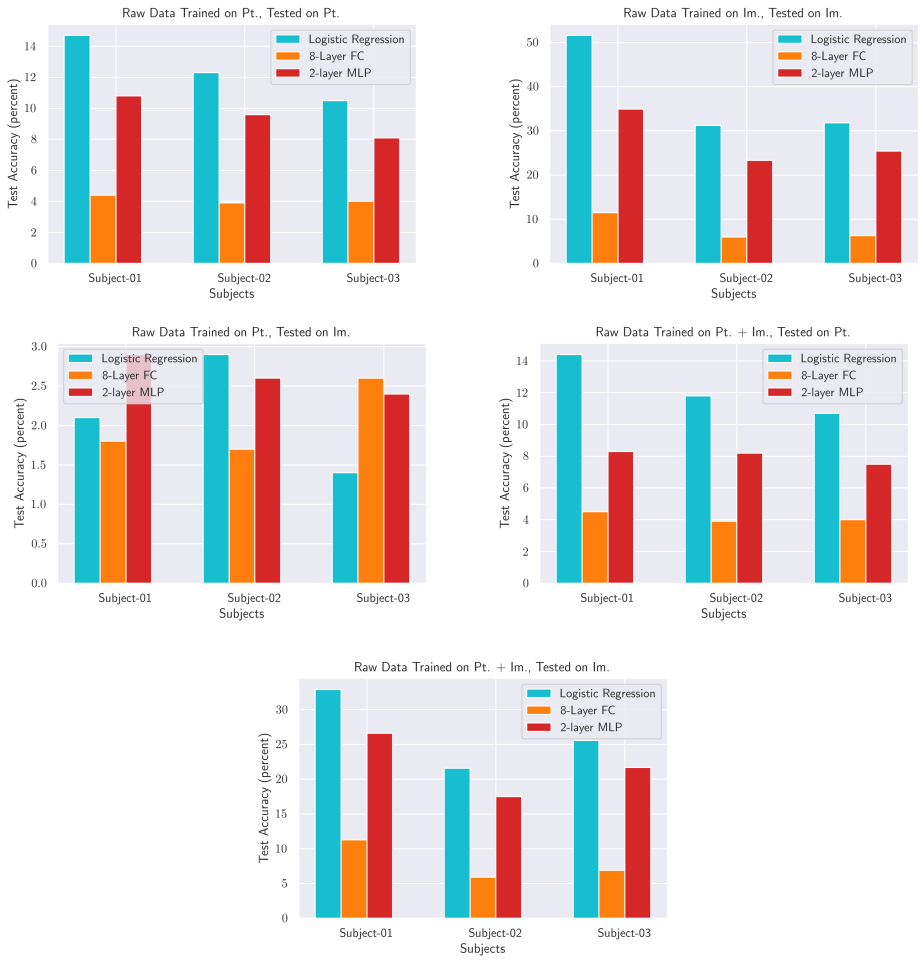
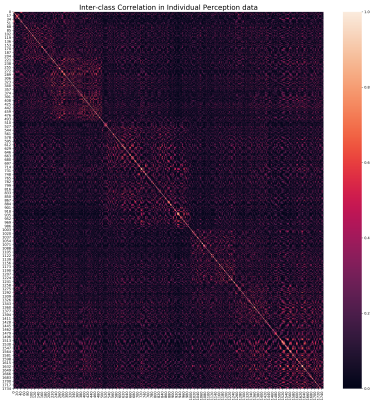


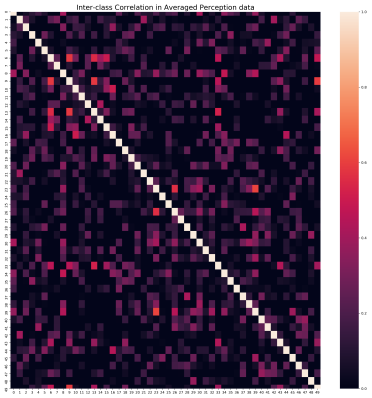
Figure 5: The qualitative analysis of all the models trained on raw fMRI data type for each subject. Starting from top-left, the first plot is trained on perception (pt.), tested on perception (pt.), the next plot is trained on imagery (im.), tested on imagery (im.), while the next plot is trained on perception (pt.), tested on imagery (im.). The last two plots are trained on perception and imagery (pt. + im.) and tested on perception (pt.) and imagery (im.) respectively. Note that these bar plots are visual illustration of Table 4. Logistic Regression is represented with cyan color, whereas 8-Layer FC Net is orange and 2-Layer MLP is red. Best viewed in color.

6 Discussion

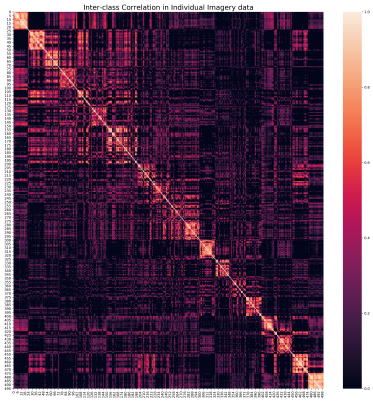
From the Results Section 5 on pre-processed fMRI data (Table 3), we observe that it is hard to learn a prediction model on imagery data. In contrast, the models trained on perception data gives more than 90% accuracy with base logistic regression model. To understand this behaviour we performed experiments to understand the data, i.e., by plotting correlation matrix to look at the inter class similarity in perception and imagery fMRI data. Figures 6 (a), (b) are the correlation plots for individual data points and the averaged (across class) data respectively for pre-processed perception data. It can be observed that the inter-class correlation is not that high for perception data. In contrast, Figures 6 (c), (d) are the correlation plots for individual and averaged (across class) data for pre-processed imagery data, and here it can be observed that inter-class correlation is much higher compared to perception correlation plots. These correlation plots can explain why models trained and tested on pre-processed perception data perform better than imagery, however, they still do not explain why the models were not able to learn anything on imagery data. Is high inter-class correlation the only reason why models are not able to learn anything on pre-processed imagery data? or is there some other factor at play? We further investigated by looking at the different



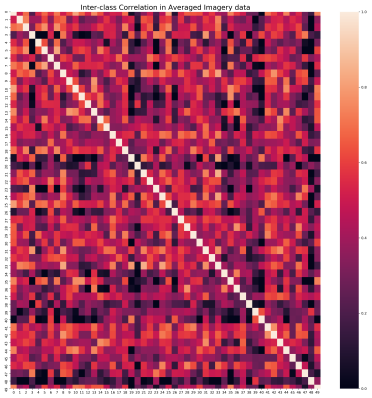
(a) Perception all samples



(b) Perception Averaged



(c) Imagery all samples



(d) Imagery Averaged

Figure 6: Correlation plots of perception and imagery data on all samples (a) and (c) respectively and also on averaged data (b) and (d) respectively.

pre-processing steps performed on the data by authors of [6]. We learned that the two most important things are (a) the four pre-processing steps mentioned in Section 3 and (b) averaging of volumes to get one scan corresponding to the stimuli. Our hypothesis is that averaging of volumes might be the reason why the prediction models are not able to learn anything on the imagery data. To test our hypothesis we performed experiments on raw fMRI data, without any pre-processing and without averaging of volumes. This means that we treat the multiple scans, 3 for perception, and 5 for imagery as individual data points corresponding to the object category label. This increased the amount of data, and also allowed us to learn on both perception and imagery data. These results (Table 4) are of most significance to us since they show that our methodology to use individual fMRI scans for prediction of object category using ANNs is possible and can give above chance accuracy. The results with combined (perception + imagery) training also encourages our hypothesis that a joint representation can be learned for both imagery and perception data. Although, we see a slight decrease in performance compared to models trained and tested on just perception and imagery data, the combined models still performs reasonably well for both perception and imagery data.

One interesting thing to note is that although using the raw fMRI data allows us to learn on both perception and imagery data, we see a drop in accuracy of perception models compared to the perception models trained on pre-processed fMRI data. This drop in accuracy can be attributed to

either the four missing pre-processing steps or not averaging the volumes. Another thing to keep in mind is that either the four pre-processing steps or the averaging of volumes maybe the reason of poor performance of models on pre-processed imagery data. Our hypothesis is that since a subjects look at the same static image of an object for 9 seconds in perception experiment, it makes more sense to average the 3 volumes, since the stimuli remains the same. In case of imagery, the subjects imagine the object for a total of 15 seconds, and it feels like a hard job imagining one specific representation of that object. Therefor averaging of volumes might not be very helpful in case of imagery. This of course is just a hypothesis and would require further investigation to have a concrete answer. One way to test this hypothesis would be to perform just the four pre-processing steps on the raw fMRI data and perform the same experiments as proposed in Section 4. One important thing to note is that in perception data collection the subjects are shown only 1 image per class for a total of 35 times. This means that the perception fMRI data corresponds to a single image of that object, and intuitively this should make the classification task relatively easy. Maybe this is one reason why we see above 90% accuracy in pre-processed fMRI experiments. However, if this is true, then either the averaging of volumes or the four pre-processing steps are very essential for perception data, and at the same time one of the two or both are not suitable for imagery data.

In Figures 4, 5 we observe that the overall trend across the three subjects is similar for all the experiments. The base logistic regression model performs the best in most cases, followed by the 2-layer MLP model, and then comes the 8-layer FC net. This trend can be attributed to the size of training data, training time, and the hyperparameter selection. We observed overfitting in all the models. To mitigate the effects of overfitting we added dropout layers and designed the 2-layer MLP model with just one hidden layer. This design choice is a step in the right direction since it mitigates overfitting and performs better than the 8-layer FC net.

6.1 Ethical Implications

Addressing the bias in design choices and/or data also necessitates the explanation(s) of any assumptions made. For this work, we build on top of the data provided by [6], as discussed at length before. One of the most fundamental assumption made is the representational assumption that states that the data sample is representative of the larger population of individuals i.e. the perception and imagery fMRI representations are consistent across a larger pool of individual. The dataset contains samples of only five individuals, out of which four are male and one is female. Since the dataset is limited, it is hard to conclude that if a joint representation can be consistently learned given fMRI data of various individuals. This in turn introduces a representational bias. Furthermore, the individuals who volunteered for the study (in [6]) are likely to be from where the authors are and therefore generalizing results across individuals of various cultures and societies is also not possible. Another important thing to consider is that in the paper [6], the individuals who volunteered for the data collection were very familiar with data collection procedures in such experiments and the authors call them experts in such data collection procedures. This also adds another dimension of bias and it is worth taking into account if having abundant familiarity with the data collection procedure effects the fMRI data that is being collected. Given all these considerations, the purpose of our study is tangential to such ethical implications in a sense that it only focuses on the possibility of learning a joint representation of perception and imagery in ANNs.

7 Conclusion & Future Work

Studies have shown that humans relate perception and imagination all too frequently and this link between the two is an integral component of the human brain. Subsequently, the ability of the human brain to imagine visual scenes, and their close similarity to perceptual visual scenes suggests a correlation in the neural mechanism of the two. In this study, we want to understand if such a correlation exists in ANNs, and if they can learn a joint representation for both imagery and perception. We perform exhaustive experiments to test this hypothesis on pre-processed, and raw fMRI data and imagery decoded to visual features. In case of pre-processed data, even baseline model such as logistic regression performs well on perception data, but fails to learn representations of imagery fMRI. However, with raw data ANNs are able to learn joint representation of both perception and imagery and results indicate above chance level performance when tested on imagery and perception. This is promising result since it quantitatively re-iterates our hypothesis.

We have worked out a case where averaging of samples (such as in pre-processed fMRI data) does not work very well on imagery and therefore we would like to conduct experiments with pre-processing steps but without averaging across available samples. Another possible prospective direction is to perform ablative studies on the topic with data collected from varied groups of individuals which also allows learning from increased set of available data. Another future direction would be to perform grid search over the hyperparameters with kFold cross validation. Another interesting future direction would be to compare our proposed methodology with existing decoding approaches as proposed in [6]. This would require updating the methodology of [6] to work for individual data points instead of averaged (across class) visual features.

References

[1] Radoslaw M Cichy, Jakob Heinzle, and John-Dylan Haynes. Imagery and perception share cortical representations of content and location. *Cerebral cortex*, 22(2):372–380, 2012.

[2] Federico De Martino, Giancarlo Valente, Noël Staeren, John Ashburner, Rainer Goebel, and Elia Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *Neuroimage*, 43(1):44–58, 2008.

[3] 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. Ieee, 2009.

[4] Giorgio Ganis, William L Thompson, and Stephen M Kosslyn. Brain areas underlying visual mental imagery and visual perception: an fmri study. *Cognitive Brain Research*, 20(2):226–241, 2004.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[6] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[7] Tomoyasu Horikawa and Yukiyasu Kamitani. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in computational neuroscience*, 11:4, 2017.

[8] Scott A Huettel, Allen W Song, Gregory McCarthy, et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[9] Matthew Robert Johnson and Marcia K Johnson. Decoding individual natural scene representations during perception and imagery. *Frontiers in Human Neuroscience*, 8:59, 2014.

[10] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[11] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

[12] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[13] Cheves West Perky. An experimental study of imagination. *The American Journal of Psychology*, 21(3):422–452, 1910.

[14] Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: decoding category information during mental imagery. *Neuroimage*, 50(2):818–825, 2010.

[15] Nigel J.T. Thomas. Mental Imagery. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.