Predicting fMRI Response Dissimilarity to Natural Scenes

Lindsay Hexter and Mariona Puente Quera

30-01-2025,

Bar-Ilan University

Introduction

This project did not aim to replicate or directly compare results from existing papers. Instead, we adopted an exploratory approach to a novel task in an existing problem space.

Much of the existing literature focuses on predicting a stimulus given fMRI (decoding), or vice versa, predicting the fMRI response given a stimulus (encoding). For instance, the Algonauts 2023 Challenge invited participants to develop models that predict fMRI neural response patterns based on image features. We instead propose a novel task: directly predicting the dissimilarity between fMRI responses to image pairs. This approach offers new insights into neural representations and explores the patterns of dissimilarity within high-resolution fMRI data.

In this study, we utilized the Natural Scenes Dataset (NSD), which contains high-resolution fMRI data collected from eight participants as they viewed 10,000 natural scene images. Each image was presented approximately three times, and the fMRI responses were averaged across repetitions to enhance reliability.

The resulting model receives two images as input and outputs the dissimilarity value between their fMRI responses: the greater the value, the more dissimilar the responses.

As motivation for our problem, the paper *THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks* explores the correlation between various pretrained model embeddings and fMRI responses. The authors noted high correlation between a few models, namely CLIP-ViT, and actual fMRI responses, giving us the basis to begin our prediction work. For some of our analyses, we also employed the THINGSvision Python toolbox to extract features from deep neural networks.

Methodology

To simplify our preprocessing steps, we used the preallocated NSD data from the Algonauts 2023 Challenge (dataset access here). The given fMRI data is initially divided into left and right hemispheres. To preprocess the data, we first selected specific regions of interest (ROIs), such as V1, V2, or all available visual cortex regions. Our approach involved experimenting with multiple ROI selections and evaluating model performance. The selected ROIs from both hemispheres were then concatenated horizontally and standardized (using StandardScaler from sklearn).

Since our goal is to predict the dissimilarity between fMRI responses to pairs of images, we first needed to calculate the required labels to train the model, i.e. to calculate the degree of dissimilarity between fMRI responses to all pairs of images. To achieve this, we employed two different metrics to compare their effectiveness in this task:

- **1 - Pearson Correlation:** Produces values ranging from 0 to 2, with 0 indicating the highest possible similarity, and 2 indicating the highest dissimilarity.
- **Euclidean Distance:** Produces positive values without a specific range, which can extend to very high magnitudes– the greater the value the more dissimilar.

A Representational Dissimilarity Matrix (RDM) is a tool commonly used in neuroscience to visualize the dissimilarity between neural representations of different stimuli. It is a square matrix where each row and column corresponds to a stimulus—in our case, an image—and the values in the matrix represent the dissimilarity between the fMRI responses to each pair of stimuli.

For our project, we used RDMs to express the results of our analysis. The x-axis and y-axis of the RDM correspond to the images, and each cell in the matrix, represented as a heatmap, displays the dissimilarity metric (either 1 - Pearson correlation or Euclidean distance).

The heatmap visualization of the RDM provides an intuitive way to interpret and compare dissimilarity patterns between the fMRI responses to the image pairs. In Figure 1 we can see our RDMs and in Figure 2 their corresponding distributions of the values:

Figure 1
RDMs using two different metrics: 1 - Pearson correlation and Euclidean distance

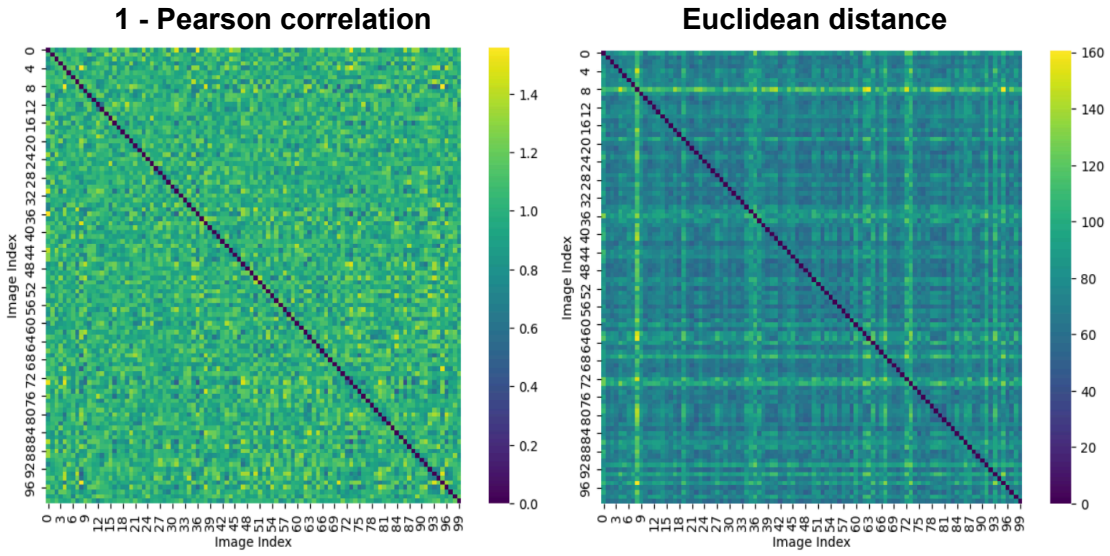
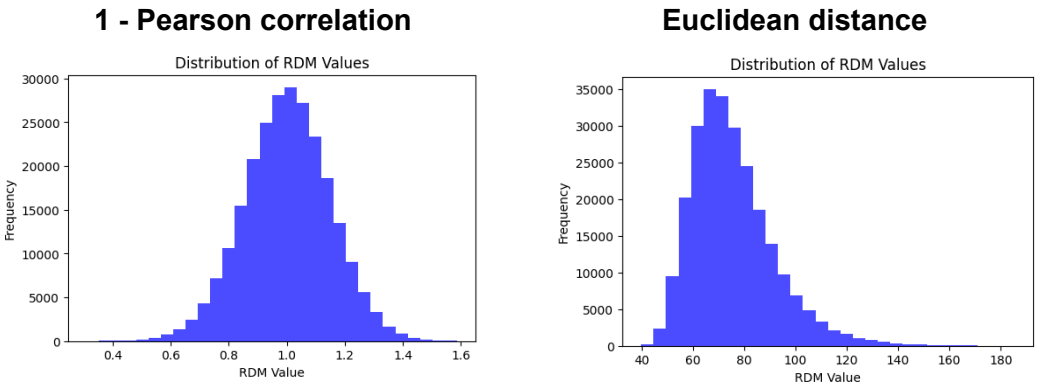


Figure 2

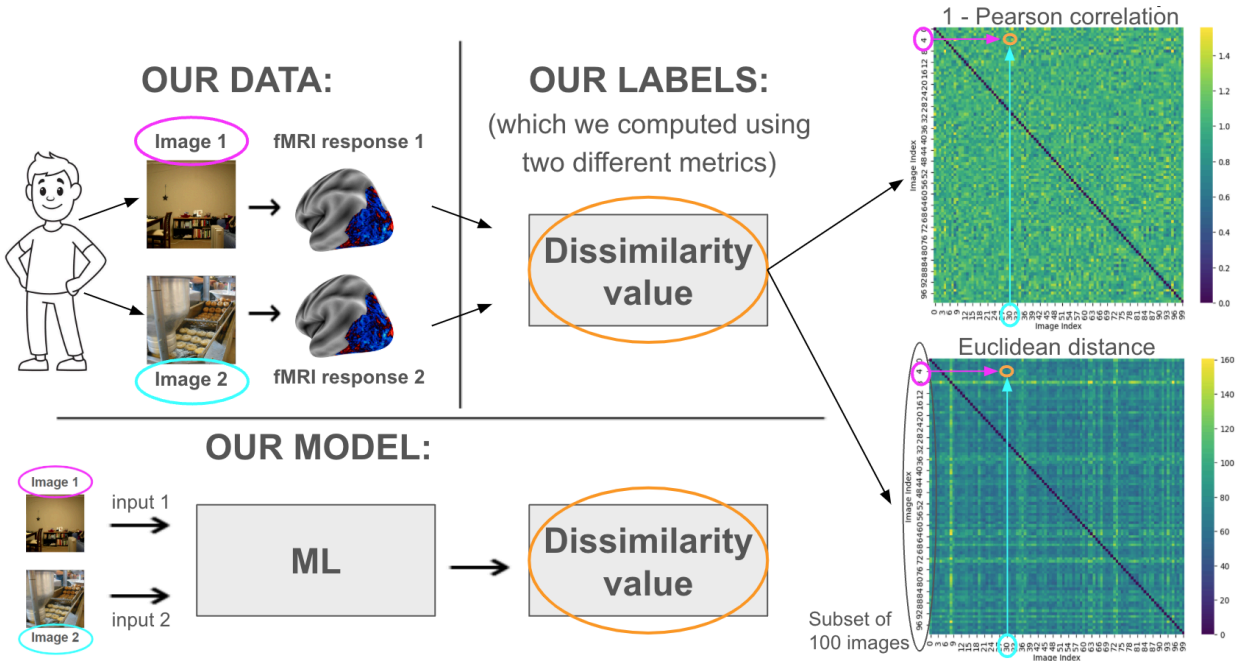
RDM distribution using two different metrics: 1 - Pearson correlation and Euclidean distance



To summarize our methodological process, Figure 3 shows a flow chart to better understand our data, labels, and model.

Figure 3

Flow chart summarizing our data, labels and model



Furthermore, as an initial step in our analysis, we aimed to explore the intuition of our approach – i.e., what does a highly dissimilar or minimally dissimilar pair of images look like in practice? To achieve this, we identified the smallest and largest values in both the RDM calculated from 1 - Pearson correlation and the RDM constructed with Euclidean distance. We then retrieved the corresponding image pairs associated with those selected RDM values.

Additionally, for the 1 - Pearson correlation metric, we also selected the image pair corresponding to the RDM value closest to 1. This allowed us to better understand what kind of image relationships correspond to intermediate dissimilarity values and evaluate whether these results align with our expectations. We completed this analysis using only V1 as well as separately with the entire available visual cortex (see here for more information on available ROIs).

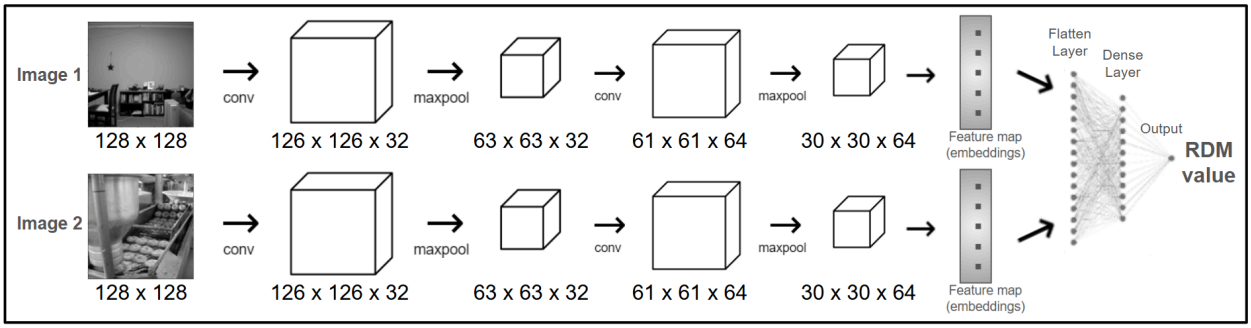
Moving on to the machine learning component, we implemented two models to tackle this task. We developed the second model after becoming more familiar with the problem space and seeing poor results in the first model.

For our first attempt, we developed a simple two-layer Siamese Convolutional Neural Network (CNN), per Figure 4. However, as this model yielded poor results, we realized that we had grossly underestimated the limitations of training a model from scratch; therefore we proceeded to architect a second model, which combined a pretrained model (to process images and extract their features) with a Multilayer Perceptron (MLP). The MLP component consisted of several hidden layers with varying configurations to optimize performance, per Figure 5.

First model: Simple Siamese 2-layer CNN with two images as inputs

- Focused only on the V1 region of the brain.
- Input: Grayscale images resized to 128x128 and normalized using min-max normalization.
- Distance between fMRI responses was only calculated using the metric 1 - Pearson correlation.
- Loss function: Combined Pearson correlation and mean squared error (MSE), weighted with a parameter alpha set to 0.5 (alpha x corr. + (1-alpha) x MSE).
- Activation functions: Linear and Sigmoid.
- Evaluation: R^2 and Spearman correlation.

Figure 4
Simple Siamese 2-layer CNN architecture

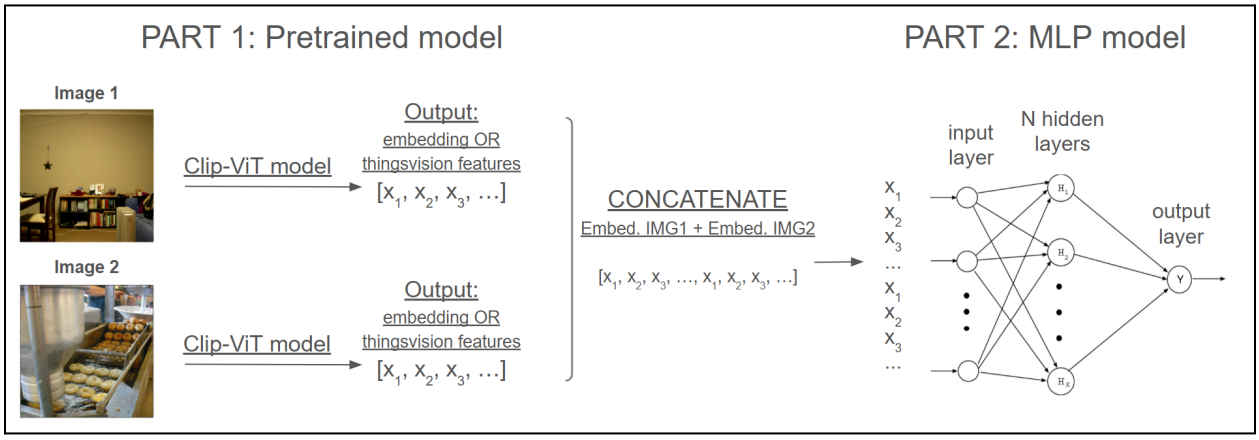


Second Model: MLP model with different number of hidden layers using a pretrained model to get the image embeddings

- Used all visual cortex ROIs (per THINGSvision paper correlation results).
- Embeddings for each image generated from pretrained Clip-ViT model either from vanilla CLIP python packages or THINGSvision package (resulting embedding is a row vector of length 512).
- Concatenated embeddings for each image pair.
- Processed the concatenated embeddings using an MLP model with a different number of hidden layers (0, 1, 2, 3).
- Distance between fMRI responses was calculated using the metric 1 - Pearson correlation and Euclidean distance.
- Loss function: MSE.
- Activation function: Linear.
- Evaluation: R^2.

Figure 5

MLP model using different numbers of hidden layers (with image embeddings from pretrained model as input)



In terms of training the models, we initially computed all pairs of images and then split the data into training and testing sets. However, after seeing extremely good model performance, we realized that computing pairs before splitting the sets creates data leakage: the model is trained on an image in one pair and tested on the same image in another pair. However, as we spent most of our time tuning model hyperparameters (e.g. layer number, learning rate, epochs) to account for results with data leakage, we present results both with and without exposed pairs in this report.

Results

Understanding our Data and Task

1 - Pearson Correlation

For the 1 - Pearson correlation metric, we can interpret the results in Table 1 as follows:

- Lowest RDM value: This corresponds to the highest positive correlation between two fMRI responses. It means that the relationship between these responses is directly proportional—these image pairs are viewed similarly by the brain, as the same areas are activated simultaneously when viewing them.
- Largest RDM value: This represents the lowest negative correlation value, indicating that these images are viewed most dissimilarly by the brain. The relationship between their fMRI responses is inversely proportional—areas activated when viewing one image are deactivated when viewing the other, and vice versa.
- Middle values: Image pairs with a correlation closer to 0 (corresponding to a dissimilarity value of 1) represent responses that are uncorrelated. These images are viewed by the brain in a way that is neither similar nor dissimilar—activation patterns for one image provide no predictive information about the other. Activation and deactivation occur independently.

Euclidean Distance

For the Euclidean distance metric, the interpretation is slightly different:

- Lowest RDM value: This represents the smallest difference between two fMRI response vectors. It indicates that the brain processes these image pairs in a very similar way, as their activation patterns are nearly identical.
- Largest RDM value: This corresponds to the greatest difference between two fMRI response vectors, meaning that these image pairs are processed very differently by the brain.
- Intermediate values: These reflect a moderate degree of difference in fMRI responses, suggesting that the images are neither highly similar nor highly dissimilar in terms of how they are represented in the brain.

Per only using V1 in Table 1, we can interpret the results via understanding which features are significant in the V1 response, namely orientation and frequency– which we can understand as edges in the images. When squinting at images corresponding to the lowest RDM values (mimicking edge detection in V1), we can see that negative and positive shapes are pronounced similarly in the pairs. Conversely, we do not see edge similarities in the images corresponding to high RDM values.

Table 1

Smallest and largest RDM values and their corresponding images using both metrics, only focusing on V1

ROI: V1			
1 - Pearson correlation of fMRI responses		**Euclidean distance of fMRI responses**	
	*Lowest RDM value* 0.288		*Lowest RDM value* 38.154
	*Closest to 1* 1.000		
	*Largest RDM value* 1.630		*Largest RDM value* 185.448

When analyzing our data using the whole visual cortex as shown in Table 2, we can see that the most similar images are indeed similar in a more general sense (i.e. more easily understood with the naked eye, not considering specific selectivity of an ROI). We see that animal images of a certain layout are most similar in the Euclidean matrix, and for Pearson correlation, we see that the images with the lowest RDM value are two very similar images of pizza.

Table 2

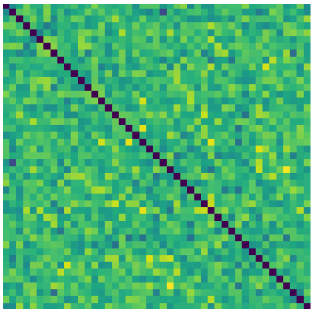
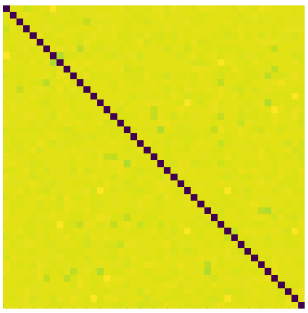
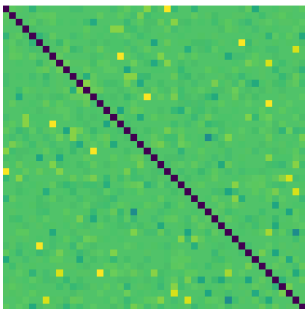
Smallest and largest RDM values and their corresponding images using both metrics, using all the visual cortex

<table>
<tr><td colspan="4">ROI: all visual cortex available</td></tr>
<tr><td colspan="2">1 - Pearson correlation of fMRI responses</td><td colspan="2">Euclidean distance of fMRI responses</td></tr>
<tr><td> 
</td><td>Lowest RDM value
0.356</td><td> 
</td><td>Lowest RDM value
181.478</td></tr>
<tr><td> 
</td><td>Closest to 1
1.000</td><td colspan="2"></td></tr>
<tr><td> 
</td><td>Largest RDM value
1.524</td><td> 
</td><td>Largest RDM value
573.075</td></tr>
</table>

First Model Results (only V1)

The results from our first model as shown in Table 3 clearly indicate a lack of predictive power for this approach. The Spearman correlation in the resulting predicted matrices is extremely low, even as the computation required for this task was expensive. Therefore, after reviewing these poor outcomes and understanding the computational and model-specific limitations, we proceeded with the second approach.

Table 3
Summary of first model results

GROUND TRUTH	Activation function: LINEAR	Activation function: Sigmoid * 2
		
	$R^2 = -0.0039$	$R^2 = -0.1116$
	Spearman = 0.0065	Spearman = 0.0053

Second Model (all ROIs available)

Results of the second model using our leaked data are plotted in Figure 6, 7, 8 and 9. We hypothesized that the Euclidean distance metric might have certain advantages over Pearson correlation, so we included both metrics in this study. Additionally, we obtained results using embeddings directly computed from the CLIP package as well as those extracted with the THINGSvision python package (in both cases we used the ViT-B/32 model). For THINGSvision, we simply utilized the provided tutorial Google Colab linked in the GitHub repo here. Essentially with THINGSvision, we are solely extracting the raw visual layer of the model, versus when using the CLIP transformers package, we are extracting the features of the final layer (which also includes information about semantic feature space).

Results for leaked data

Figure 6

Results using leaked data, 1 - Pearson correlation and directly computed embeddings from CLIP-ViT-B/32 model

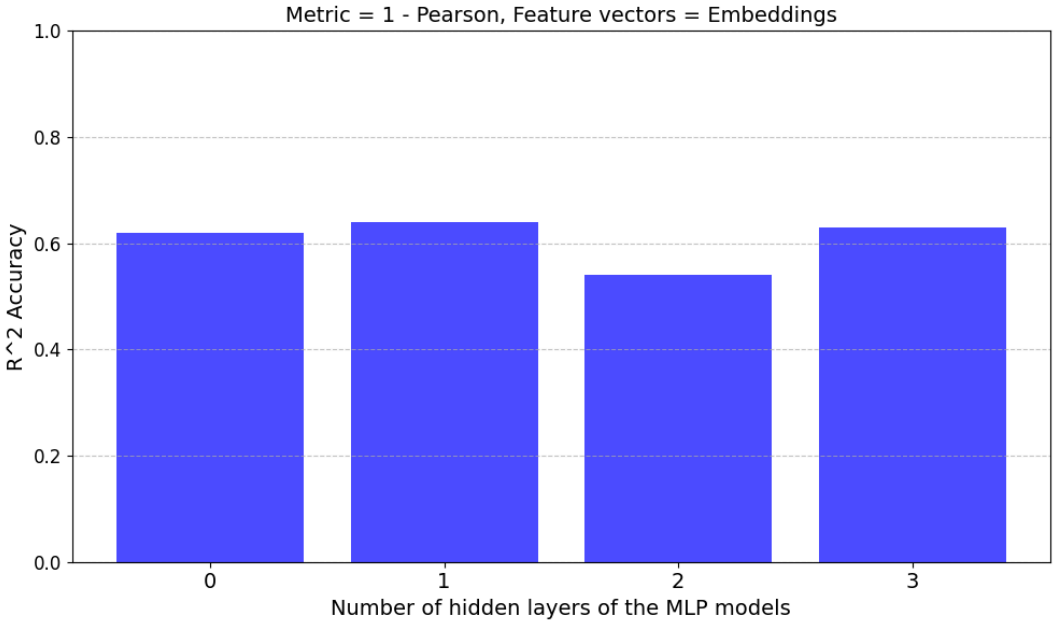


Figure 7

Results using leaked data, 1 - Pearson correlation and feature vectors from CLIP-ViT-B/32 model computed with THINGSvision toolbox

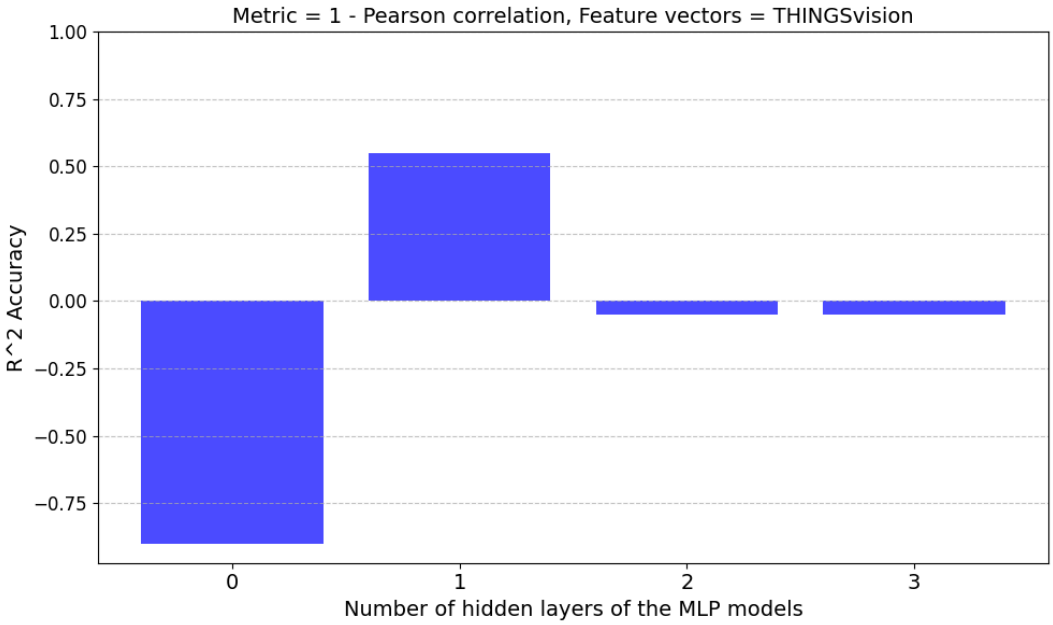


Figure 8

Results using leaked data, Euclidean distance and directly computed embeddings from CLIP-ViT-B/32 model

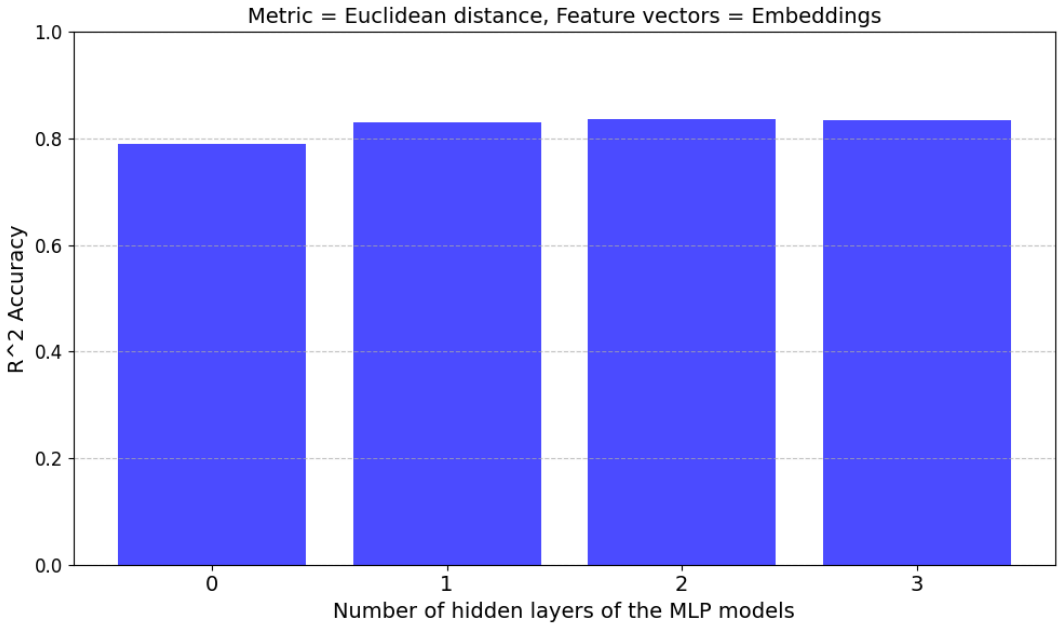
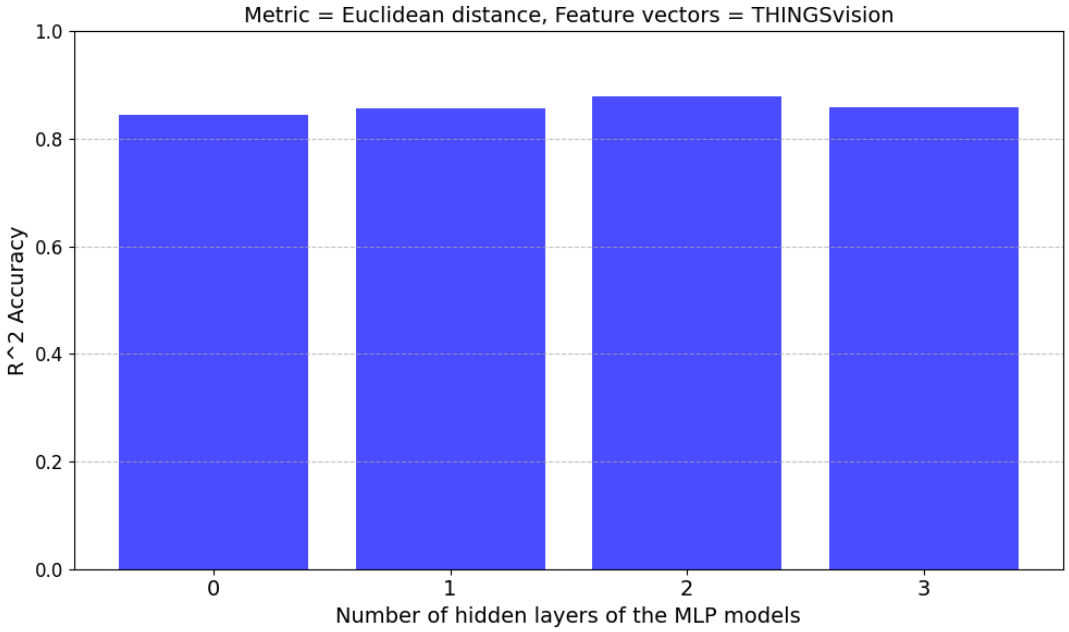


Figure 9

Results using leaked data, Euclidean distance and directly computed embeddings from CLIP-ViT-B/32 model



As shown in these 4 figures (using leaked image pairs):

- Euclidean distance performs significantly better than Pearson correlation for this task.
- For Pearson correlation, the direct CLIP-ViT embeddings outperform the visual features extracted using THINGSvision.
- Overall Euclidean distance seems to be a more suitable metric than 1 - Pearson correlation.

In comparison, per Figure 10, 11, 12 and 13, non-leaked results showed poor performance. We consistently saw negative R^2 values for accuracy, in both vanilla CLIP extraction and THINGSvision features.

Results for non-leaked data

Figure 10
Results using leaked data, 1 - Pearson correlation and directly computed embeddings from CLIP-ViT-B/32 model

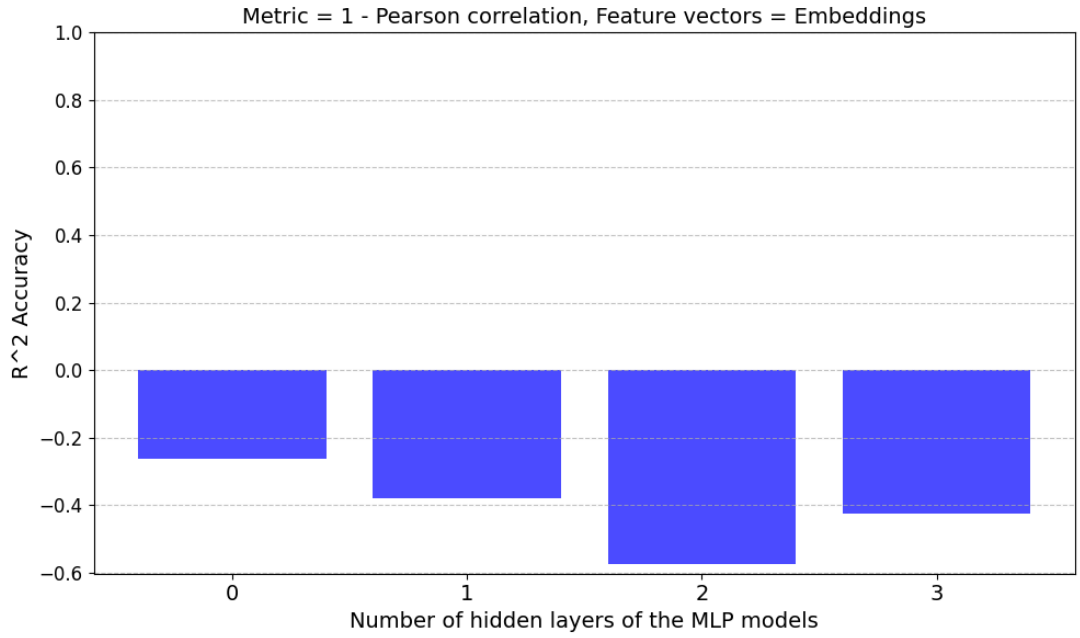


Figure 11

Results using leaked data, 1 - Pearson correlation and feature vectors from CLIP-ViT-B/32 model computed with THINGSvision toolbox

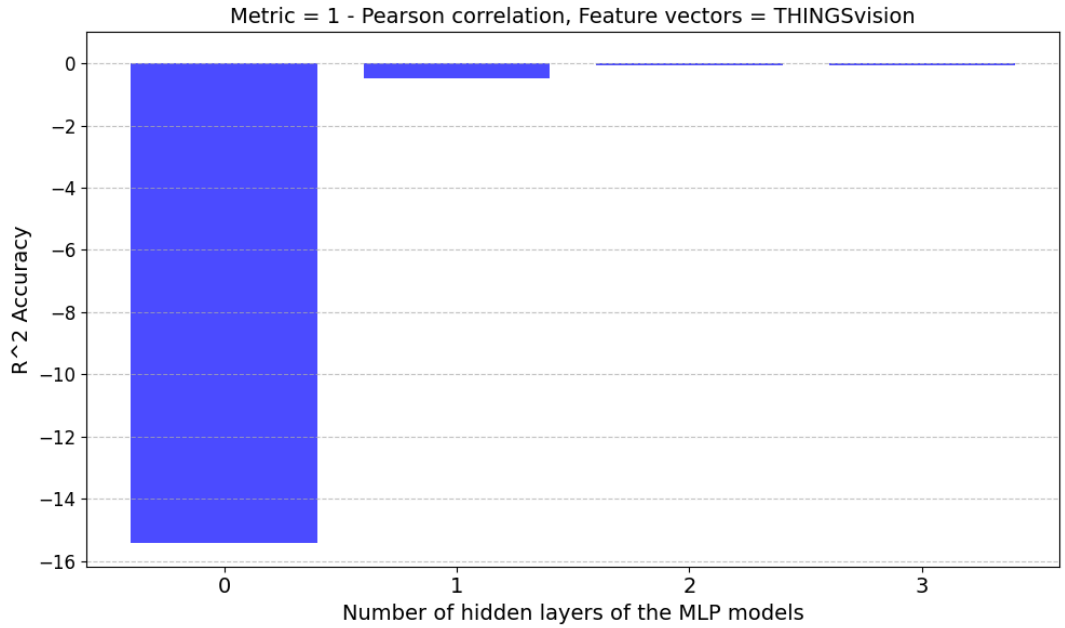


Figure 12

Results using leaked data, Euclidean distance and directly computed embeddings from CLIP-ViT-B/32 model

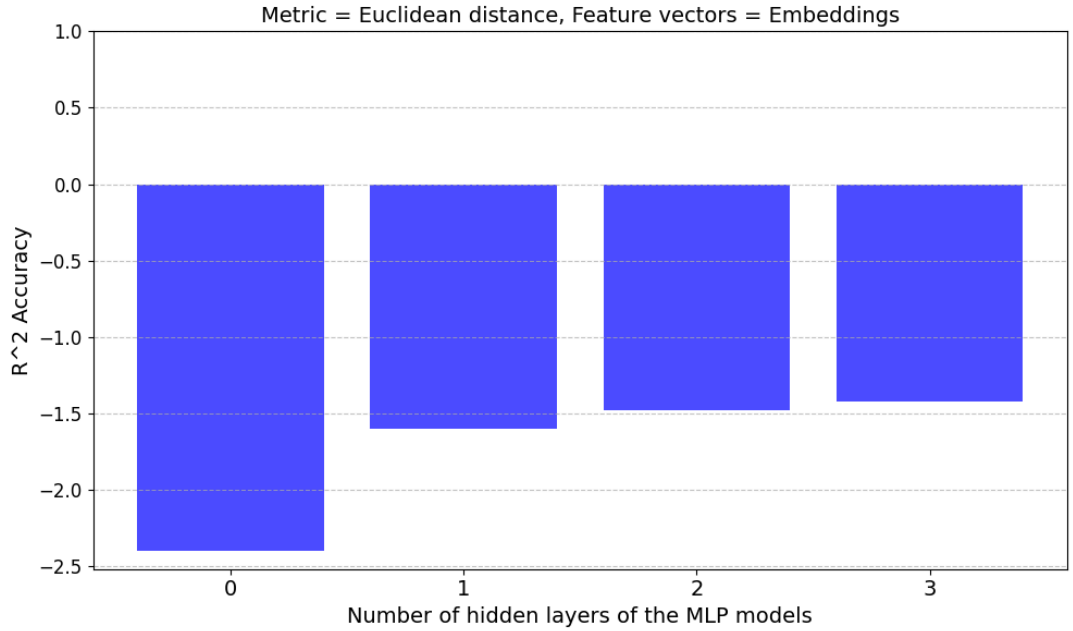
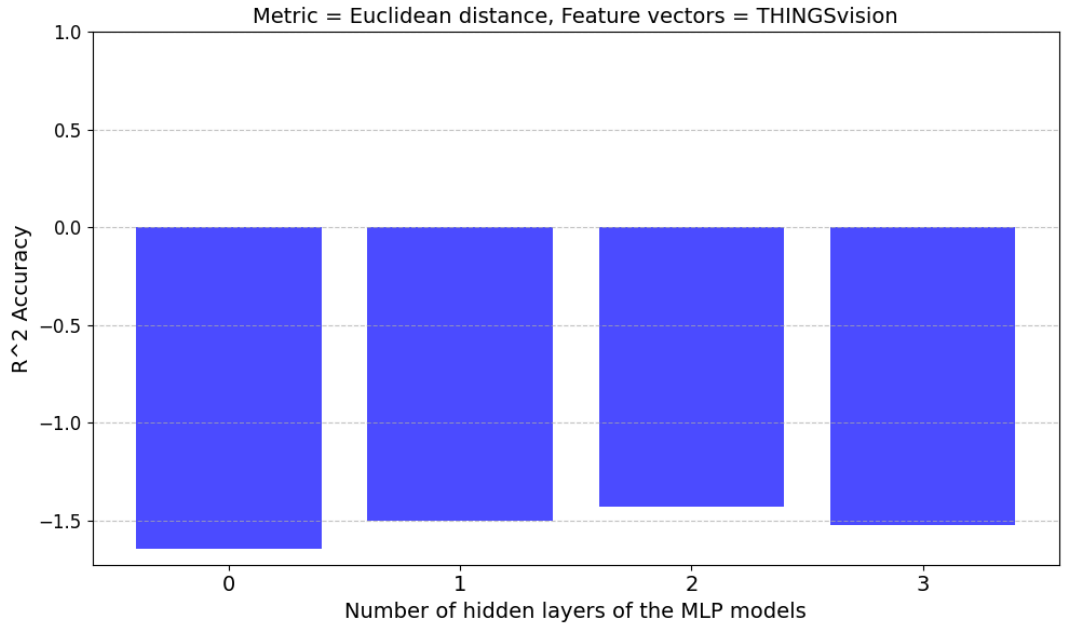


Figure 9

Results using leaked data, Euclidean distance and directly computed embeddings from CLIP-ViT-B/32 model



Conclusion and discussion

Unfortunately much of our time was spent tuning hyperparameters for a model trained on unintentionally leaked data, as described in methodology. Therefore there is much to be improved upon now that we have a better understanding of our model baseline performance, even with using a pretrained model to start.

Moreover, while the THINGSvision paper showed high correlation between RDMs created from model embeddings as compared to fMRI responses to pairs of images, this may not indicate that training a model further and predicting the actual, continuous dissimilarity metrics will be successful. We saw that for RDMs generated using all visual cortex information, the RDM Spearman correlations were higher at the onset using 1 - Pearson correlation as the dissimilarity metric (e.g. 0.288 for THINGSvision features) rather than Euclidean distance (e.g. 0.162 for THINGSvision features). As Euclidean outperformed Pearson in the leaked data results, we thus may not be able to base our hypotheses on initial RDM correlations. However, considering poor results obtained using either metric with non-leaked data, we need to reevaluate our explanation upon reaching better performance.

Additionally, as it does seem intuitive to compare images from just V1 fMRI data per Table 1, we could explore more thoroughly both V1 and other ROIs. However, using all ROIs available to obtain as much visual information as possible in predicting the outputs could improve performance as we could intuit from Table 2.

Challenges and Limitations

- We initially wanted to check if our task could be resolved using a simpler linear model instead of an MLP; however, computational resource constraints prevented the implementation of Support Vector Regression (SVR), one such example model we tried.
- The embeddings generated by the pretrained Clip-ViT model did not correlate strongly with the fMRI data, suggesting that the results from the THINGSvision paper may not apply to the NSD. We attempted to mitigate this by extracting only visual features using the THINGSvision package (Spearman correlation was still low, 0.288, but a bit improved).
- We wasted cycles tuning a model which was based on leaked data.

Future Directions

Future work could include:

- Input symmetry: the way in which our embeddings/ extracted features are concatenated may encourage the model to learn that input order is important. Instead, the model should be able to compute the dissimilarity value regardless of the order of the embeddings (e.g. training on both combinations of pair inputs).
- Explore contrastive learning approaches to train our model, which may be more successful than predicting a continuous value.
- Experiment with alternative pretrained models. We also tried VGG16, because it is solely a visual model; however, the resulting vectors were of size 200,504, compared to 512 from Clip-ViT. This was a blocker from a computational perspective.
- Enhance computational resources to implement more computationally demanding models like VGG16 and to handle more training data (complexity of pairs is always N choose 2, and we already have nearly 80,000 training data pairs when selecting 500 images with a 400 training/100 testing split).
- Further investigate the use of specific ROIs and compare.
- Once a trustworthy analysis is performed for one subject, repeat over all eight subjects.

This project highlights the complexity of predicting fMRI response dissimilarity and provides valuable insights for future work in this domain.