# Predicting EEG responses to visual stimuli using Deep Neural Networks

Satyam Bhardwaj

22210040, M.Tech CSE, IIT Gandhinagar
bhardwajsatyam@iitgn.ac.in

**Abstract.** The human brain is able to recognize and classify objects on a timescale of milliseconds. Deep neural networks have proven effective in modeling these rapid neural dynamics. However, they require massive amount of data to train and collecting EEG data for thousands of images in a traditional setting is infeasible. Here we use the THINGS EEG1 dataset [5] of EEG responses to visual stimuli in 50 participants collected using the RSVP paradigm. As a first step, we test a linearizing encoding model on this dataset, both within and between participants, and evaluate the results via a correlation analysis. We also test an end-to-end encoding model from a randomly initialized AlexNet and compare to the linearizing encoding model.

**Keywords:** EEG encoding · Visual Object Recognition · Deep Neural Networks

## 1 Introduction

### 1.1 Motivation

Humans are capable of visually recognizing and meaningfully interacting with a vast array of objects, despite significant variations in the angle of observation, lighting conditions, and the presence of visual clutter. This process of visual object recognition happens on the timescale of milliseconds, and is a complex multi-stage cognitive process involving many linear and non-linear transformations. To capture such fast neural dynamics, we require techniques with high temporal resolution such as EEG. In order to understand and explain these rapid transformations, we need predictive models that can synthesize EEG responses to arbitrary visual stimuli.

Recently, computational neuroscientists have begun to use Deep Learning to predict the brain responses to visual stimuli. Such techniques work best when trained with a large amount of data. For example, the ImageNet dataset [17] used to train the AlexNet CNN architecture [12] contains more than 1.4 million images. Collecting the EEG responses to each image in the dataset at a rate of one image per second will take weeks for a single participant, making this highly infeasible. However, Grootswagers et al. have shown that it is possible to extract meaningful information from the EEG signals about the images presented in Rapid Serial Visual Presentation (RSVP) paradigm [4,11]. We now look at two large EEG datasets using this paradigm.

### 1.2 Relevant literature review

There is rising interest in utilizing large-scale image databases for neuroimaging studies. The THINGS database [9] is a large database consisting of high-quality curated images of objects on a natural background with 1854 object concepts (e.g. hat, tiger, apple), each belonging to one of 27 higher-level categories (e.g. clothing, animal, food). Each object concept contains 12 or more images, for a total of 26,107 images.

Grootswagers et al. [5] have collected a large dataset of 64-channel EEG responses from 50 participants to 22,248 images from the THINGS database. The images were shown using the Rapid Serial Visual Presentation (RSVP) paradigm [4,11] to reduce the experiment time. Each image was shown only once. Another stage of the experiment included 200 validation images shown to each participant in 12 sequences in a random order. This constituted the validation dataset for performing the noise ceiling analysis.

To assess the quality of this dataset, the authors compute decoding accuracy using leave-one-out cross validation by training a regularized linear discriminant classifier on the EEG visual responses to distinguish between pairs of images. This resulted in a $1854 \times 1854$ Representational Dissimilarity Matrix [4]. This dataset has a huge potential for being the testing ground of various deep learning approaches for EEG encoding of visual stimuli. As of now, there are no published analysis that build encoding models using this dataset. Hence, I decided to work on this dataset in this paper. From now on, this dataset is referred to as **THINGS EEG1**.

In a similar vein, Gifford et al. [3] collected a dataset consisting of 64-channel EEG responses from 10 participants to 18540 images from the THINGS database, the main difference being a larger number of trial repetitions. The authors divided the images into 1654 training and 200 testing concepts and built linearizing encoding models that synthesize EEG visual responses using four different Deep Neural Networks (DNNs): AlexNet [12], ResNet-50 [7], CORnet-S [13] and MoCo [6], pre-trained for object classification on ILSVRC-2012 training set [17]. To test the accuracy of the linearizing encoding models, they used multiple strategies such as

- Correlation between the actual EEG response (BioTest) and synthesized EEG response (SynTest)
- Pairwise decoding, which involved training one-vs-one linear SVMs on BioTest and testing them on SynTest. The decoding accuracy is a measure of similarity between the two.
- Zero-shot Identification of BioTest image conditions among SynTest augmented with varying set sizes of SynImageNet, synthesized EEG responses using ILSVRC-2012 [17].

They also trained a randomly initialized AlexNet to learn the mapping between training set images and the corresponding EEG responses, thus building an end-to-end encoding model that predicts the EEG visual response per time point. From now on, this dataset is referred to as **THINGS EEG2**. This dataset

is also very recent and as of now, no other analysis has been published, besides the work already done by the original authors.

### 1.3 Contribution

To validate the approach of Gifford et al. [3] for predicting EEG visual responses, we train a linearizing encoding model using pre-trained AlexNet feature maps as predictors using the THINGS EEG1 dataset [5]. We also train a randomly initialized AlexNet to build end-to-end encoding models. As a measure of how well our models synthesize the EEG data, we perform a correlation analysis using Pearson's-r metric. We observe correlation peaks significantly above chance level for models trained both within and between participants.

We chose to work with the THINGS EEG1 dataset as it contains five times more participants and a larger number of image conditions (12 per concept) as compared to THINGS EEG2 dataset. In the discussion section of [3], Gifford et al. find that the prediction accuracy of encoding models is more dependent on the number of image conditions than the number of repetitions. We verify this by showing that even when trained with a single image condition repetition during the training, we obtain correlation significantly above chance level on the test set.

## 2 Methodology

### 2.1 Dataset description

The THINGS EEG1 dataset [5] consists of 64-channel EEG recording of 50 participants when presented with 22,248 images from the THINGS database, where the 64-electrodes were placed according to the international standard 10-10 system. The images were presented at 10 Hz with a 50% duty cycle. This means each image was shown for 50 ms, followed by a blank screen for 50 ms. The analog signal was digitized at 1000 Hz sample rate.

The first 12 images were selected from each 1854 image concepts to obtain 22,248 images, which were divided into 72 sequences of 309 stimuli belonging to different concepts. A set of 200 images were selected for validation (different from the previous images, but overlapping image concepts) and presented to the participants in 12 sequences, each time in random order.

### 2.2 Preprocessing

The authors of the dataset provide both raw and minimally preprocessed data, in the BrainVision and EEGLAB [2] formats respectively. We work with the preprocessed data in the interest of time. Their preprocessing steps included a Hamming windowed FIR filter with 0.1Hz high-pass and 100 Hz low-pass, re-referencing to the average reference, and downsampling to 250Hz. Out of the 50 participants, 6 were excluded either due to missing validation data, poor signal quality or equipment failure as suggested in the dataset paper [5].

We load the EEGLAB-format preprocessed data using the `mne` python library [1]. We restrict our analysis to the 17 electrodes above the Occipital and Parietal lobes of the brain, as we are primarily interested in the response of the visual cortex. The dataset was annotated with a boundary event signaling the start of the experiment, stimulus onset event (marked as E1), stimulus offset event (marked as E2) and start of each sequence (marked as E3).

We extract these events from annotations, then epoched the continuous data using the E1 events in the interval -100 ms to 400 ms, the primary stimulus onset event of interest being 0 ms. Baseline correction was applied using the 'mean' method with 100 ms prior to stimulus onset as the baseline. We resample the epochs down to 200 Hz and load the data, resulting in a `numpy` array of shape $24648 \times 17 \times 101$ (events $\times$ number of EEG electrodes retained $\times$ time points). Since each element of the matrix is a voltage in the units of $V$, we multiply by $10^6$ to store the voltage in units of $\mu V$.

The 24,648 events are further divided into 22,248 events from the main sequences and 2,400 events from the validation sequences. For our purposes, we perform a train-test split by excluding the 200 validation image concepts (the last 2400 events) from the main sequences, resulting in 19,848 image conditions in our training set, 12 images per 1654 image concept. We reshape our arrays to include the repetition dimension. See Table 1 for a summary of the array shapes.

| Partition | Images | Repetitions | Channels | Samples |
|---|---|---|---|---|
| Train EEG data | 19848 | 1 | 17 | 101 |
| Test EEG data | 200 | 12 | 17 | 101 |

**Table 1:** Summary of the data array shapes

### 2.3 Extracting feature maps using AlexNet

To predict the EEG responses given the visual stimuli, the first step is to apply a non-linear transformation to obtain the feature map of the input image. A feature map is essentially a vector in a high-dimensional space that is representative of the image for some particular task e.g. image classification. Each component of the vector reflects the degree to which a specific feature is present in the image.

We have used the AlexNet architecture, pretrained on the ILSVRC-2012 training image partition [17] for the image classification task provided by the PyTorch library [16]. The feature vector for an image consists of the activations of a select subset of layers in the AlexNet architecture, flattened and appended together, resulting in a 205672-dimensional vector. The selected layers include `maxpool1`, `maxpool2`, `ReLU3`, `ReLU4`, `maxpool5`, `ReLU6`, `ReLU7`, and `fc8`.

Since a single linear layer between the feature map and the 1717-dimensional EEG data (17 channels $\times$ 101 time-points) would consist of more than 353M

parameters, we perform non-linear dimensionality reduction using the Kernel PCA technique with a polynomial kernel of degree 4 and retain the first 1000 independent components. In the analysis by Gifford et al. [3], they have found that increasing the number of independent components beyond 1000 leads to diminishing returns. Thus, we obtain a 1000-dimensional feature vector for each image in the stimulus set.

### 2.4 Linearizing encoding models

A linearizing encoding model is a linear regression with the dimensionality-reduced feature maps as predictors and the recorded EEG data as the target. We compute the coefficients of the linear regression by solving the normal equation separately for each EEG channel and time-point using the training data.

$$W_{t,c} = \left(X^T X\right)^{-1} X^T y_{t,c}$$

Here, the matrix $X$ is of shape $19848 \times 1000$, each row being the kPCA components of the feature maps. $y_{t,c}$ is 19848-dimensional vector, each element being the EEG voltage at time $t$ and channel $c$. The learned weight matrix $\mathbf{W}$ of shape $17 \times 101 \times 1001$ is used to synthesize the EEG responses for the images in the test set.

Since we have the data of multiple participants, we can perform the analysis in two different ways. The default approach is to perform the training procedure separately for each participant in isolation. This is termed the *within* participants approach; the model is trained and tested on data of the same participant. However, to test how well the model generalizes to new participants having never seen their data during training, we also follow the *between* participants approach; the model is trained using the averaged EEG training data of $n - 1$ participants and tested against test EEG data of the left-out participant. We later compare the two approaches using a correlation analysis.

### 2.5 End-to-end encoding models

Owing to the vastness of the dataset, it is feasible to build an end-to-end encoding model. We train a randomly-initialized AlexNet using PyTorch [16] to predict the entire EEG response during the epoch interval when given the stimulus image as input. We replace the output linear layer of 1000 neurons with a linear layer of 1717 neurons, corresponding to the EEG voltage at each channel and time-point in the epoch interval. The training objective of the model is to minimize the Mean-Squared Error (MSE) loss between the predicted EEG voltages and the training EEG data. With a batch size of 64 images, we train the model using the Adam optimizer with `lr=1e-5` for 50 epochs and save the model with the lowest validation loss. We use this model to synthesize the EEG response given images from the stimulus test set.

### 2.6 Correlation analysis of synthesized EEG responses

As a measure of how well our models synthesize the EEG data, we perform a correlation analysis using Pearson's-$r$ metric. We first average the test EEG data across repetitions and for each EEG channel and time-point, we correlate the synthesized EEG data with the test EEG data across the dimension of image conditions. This gives us a $17 \times 101$ matrix with correlation values for each participant, which we show as a heat-map and the average over channels as a plot of correlation vs. time.

## 3 Results

### 3.1 Visualizing the Event-Related Potentials

We visualized our dataset by plotting the Event-Related Potentials (ERPs) by averaging the data arrays over the image conditions and repetitions (see Fig. 1). It was observed that the ERPs of many participants contained significant artifacts and large voltage spikes, and thus their data was rejected from the analysis. The ERP plots of all the participants are available in our repository.

The ERPs show peaks of activity every 100ms, consistent with the 100ms stimulus onset asynchrony (SOA) of the rapid serial visual presentation (RSVP) paradigm used to collect the EEG data.

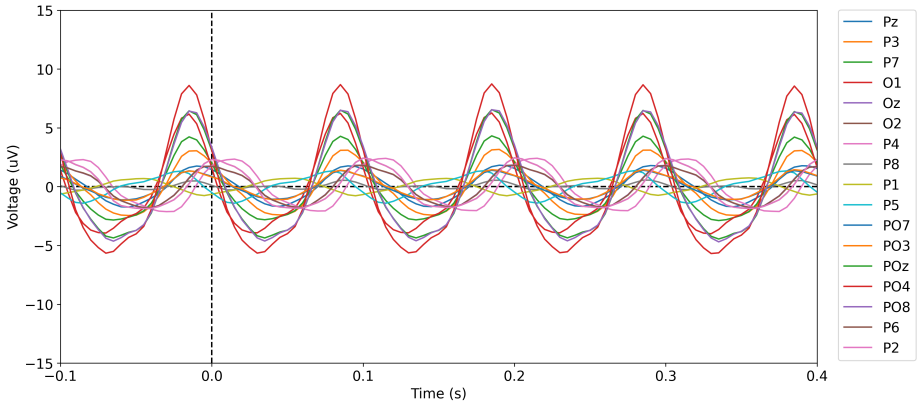


**Fig. 1:** ERP of participant 34, obtained by averaging over image conditions and repetitions.

### 3.2 Correlation plots of Linearizing encoding model

We plot the Pearson's-$r$ correlation vs. time for the EEG data synthesized using the linearizing encoding model. The dots below zero-line in figure 2 indicate

statistical significance computed with $P < 0.05$ using one-sample one-sided t-test after Fisher's z-transform, Bonferroni-corrected [3]. The error margins show the 95% confidence interval.

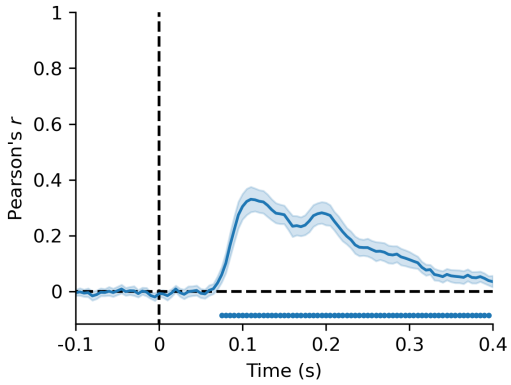


**Fig. 2:** Pearson's-$r$ correlation over time, averaged over participant for linearizing encoding models trained *within* participants. The error margins reflect 95% confidence interval. The dots show statistical significance ($P < 0.05$).

See figure 2 for the plot averaged over participants and 7 for plots of individual participants for the model trained *within* participants. In the averaged plot, we observe significant correlation from around 70 ms after the stimulus onset, and remain significant till the end of the epoch. The peak correlation of 0.33 occurs at 110 ms after the stimulus onset. Another peak occurs approximately 100ms after the primary peak. Participant 34 shows the highest peak correlation of 0.609 occurring 115 ms after the stimulus onset, while participant 41 shows the weakest correlation peak of 0.162 at 145 ms.

See figure 3 for the plot averaged over participants and 8 for plots of individual participants for the model trained *between* participants. In the averaged plot, we observe significant correlation from around 70 ms after the stimulus onset, and remain significant till the end of the epoch. The peak correlation of 0.274 occurs 195 ms after the stimulus onset. Participant 38 shows the highest peak correlation of 0.485 occurring 215 ms after the stimulus onset, while participant 27 shows the weakest correlation peak of 0.167 at 285 ms. It is observed that the primary peak around 100 ms is consistently weaker than the second peak around 200 ms, which reflects in figure 3.

### 3.3 Correlation plots of End-to-end encoding model

Similarly, we plot the Pearson's-$r$ correlation vs. time for the EEG data synthesized using the end-to-end encoding model trained *within* participants. The dots below zero-line in figure 2 indicate statistical significance computed with $P <$

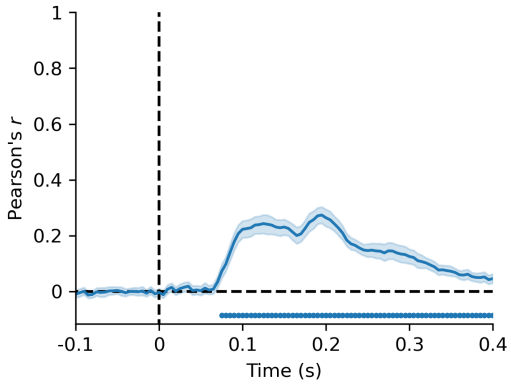


**Fig. 3:** Pearson's-$r$ correlation over time, averaged over participant for linearizing encoding models trained *between* participants. The error margins reflect 95% confidence interval. The dots show statistical significance ($P < 0.05$).

0.05 using one-sample one-sided t-test after Fisher's z-transform, Bonferroni-corrected [3]. The error margins show the 95% confidence interval.

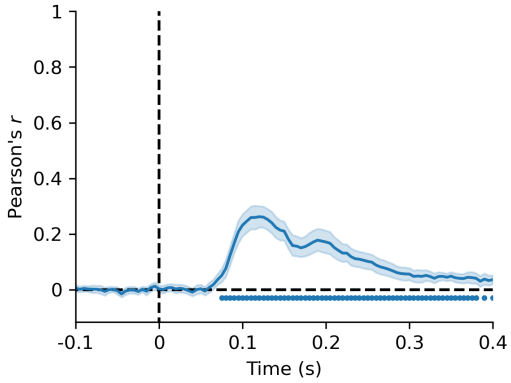


**Fig. 4:** Pearson's-$r$ correlation over time, averaged over participant for the end-to-end encoding models trained *within* participants. The error margins reflect 95% confidence interval. The dots show statistical significance ($P < 0.05$).

See figure 4 for the plot averaged over participants and 9 for plots of individual participants. In the averaged plot, we observe significant correlation from around 80 ms after the stimulus onset, and remain significant till the end of the epoch, with the exception of a few points. The peak correlation of 0.262 occurs 120 ms after the stimulus onset. Participant 34 shows the highest peak corre-

lation of 0.568 occurring 110 ms after the stimulus onset, while participant 41 shows the weakest correlation peak of 0.121 at 225 ms.

### 3.4 Comparison between the different approaches

The peak correlation and their occurrence times are compiled in table 2. See figure 5 and 6 for the correlations of different approaches in the same plot for comparison. We observe that the shape of the plot for linearizing encoding model (*within* participants) and the end-to-end encoding model (also trained *within* participants) are very similar. The correlation in the end-to-end encoding model is consistently weaker compared to the linearizing encoding model for both *within* and *between* subjects approach.

| Encoding type | Peak correlation | Time (ms) |
|---|---|---|
| Linearizing (within) | 0.33 | 110 |
| Linearizing (between) | 0.274 | 195 |
| End-to-end (within) | 0.262 | 120 |

**Table 2:** Peak correlation value and the its time of occurrence after the stimulus onset, averaged over participants.

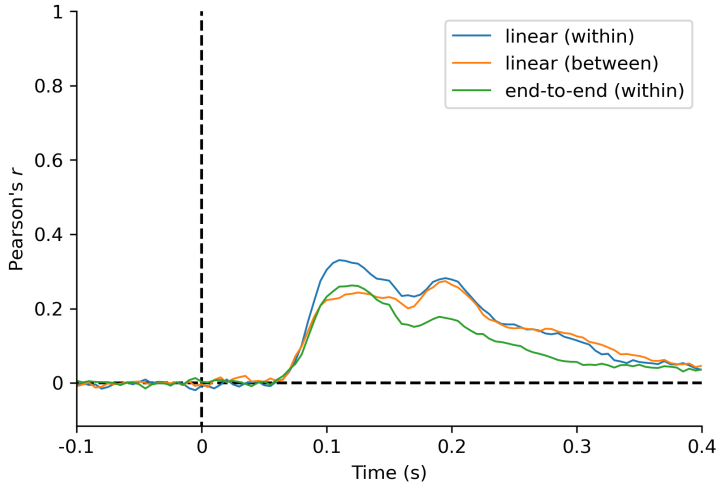


**Fig. 5:** Single participant correlation comparison between *within*/*between* approach for linearizing encoding models and *within* for end-to-end encoding models.

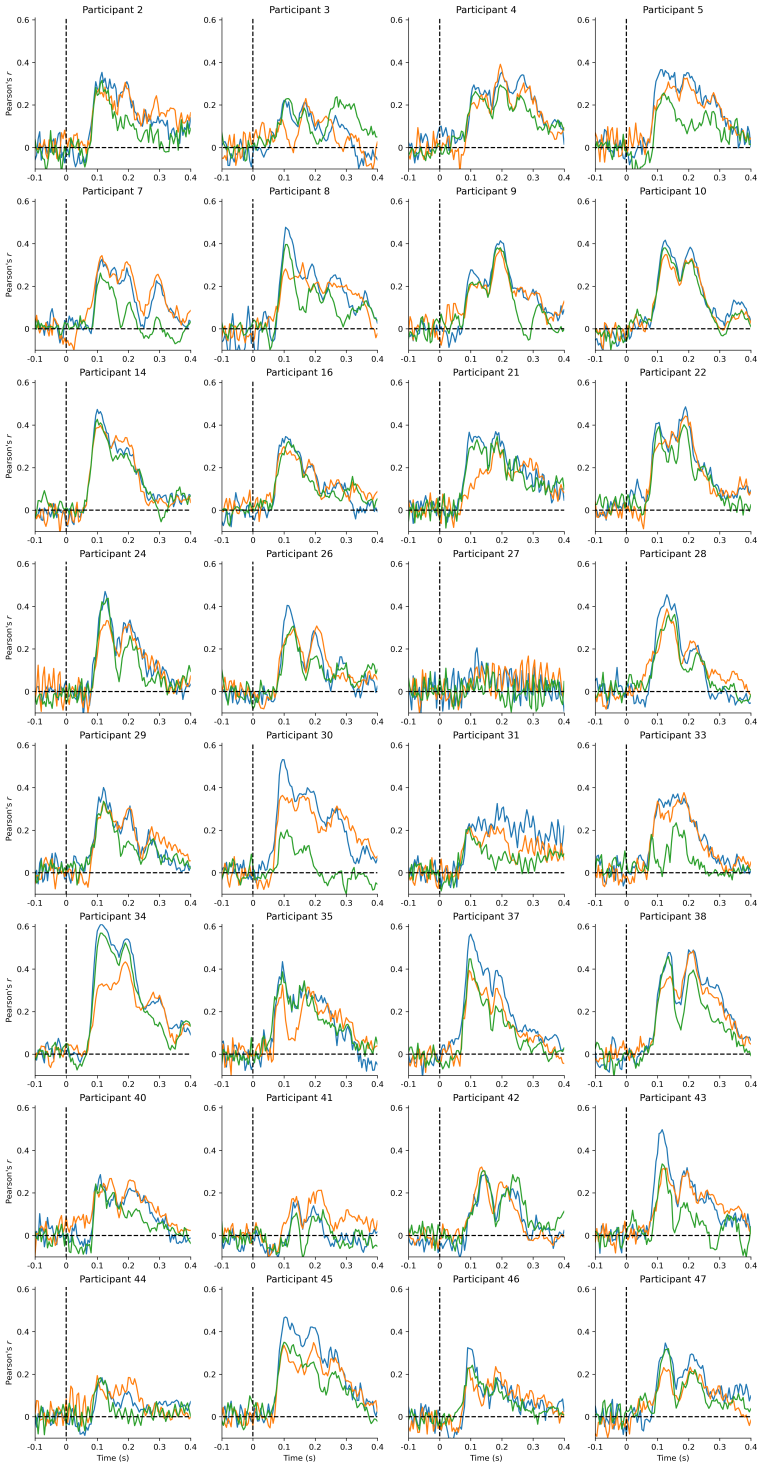


**Fig. 6:** Average correlation comparison between *within*/*between* approach for linearizing encoding models and *within* for end-to-end encoding models. Legend same as Figure 5

# 4 Discussion

## 4.1 Comparison to Gifford et al.

Comparing our results those obtained by Gifford et al. on the THINGS EEG2 dataset [3], we obtain a lower peak correlation averaged over subjects for both the linearizing encoding models and the end-to-end encoding models. The reason could be a single trial repetition per image in the training set. Due to the quick image presentation rate of 10 Hz, the RSVP paradigm introduces significant forward and backward masking in the signal. By averaging over the trial repetitions, Gifford et al. were able to mitigate the effect of this noise in the training set which is not possible for the THINGS EEG1 dataset.

## 4.2 Generalizing to novel participants

The *between* participants approach effectively asks the model to predict the EEG response of a participant it had never seen the EEG data of during training. Better encoding models that generalize well to new participants will enable us to study the core computational processes that happen in the brain during visual processing that are common to all humans. The peak correlation of 0.274 around 195 ms after the stimulus onset is significantly higher than chance level (which is 0), suggesting that the approach generalizes well to novel participants.

## 4.3 Linearizing encoding models vs end-to-end encoding models

We observe that the linearizing encoding models perform better than the end-to-end encoding models based on the correlation analysis. This could be due to the fact that the validation loss of the end-to-end encoding models did not significantly dip below their initial random value. Even then, this means that the end-to-end encoding model with the best validation loss is similar to a randomly initialized AlexNet, in which case the correlation is better than expected. This suggests that the inductive biases inherent in the architecture of the model are already sufficient to explain much of the variance in the EEG data.

## Acknowledgment

I thank Shivam Chaudhary and Manan Shah for their help in the initial stage of the project, and Shriraj for helping me find the annotation codes in the dataset. I thank the CVIG lab for providing me with the computational resources, without which this work was impossible.

### Code and data availability

We adapt the publicly available code ((link)) by Gifford et al. [3] for this dataset. The code to reproduce the results in this paper is available here along with all the figures. The THINGS EEG1 dataset [5] can be found here.

## References


1. Alexandre, Gramfort, M., Luessi, E., Larson, D., Engemann, D., Strohmeier, C., Brodbeck, R., Goj, M., Jas, T., Brooks, L., Parkkonen, M., Hämäläinen: Meg and eeg data analysis with mne-python. Frontiers in Neuroscience **7**(267), 1–13 (2013)
2. Delorme, A., Makeig, S.: Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. Journal of Neuroscience Methods **134**(1), 9–21 (3 2004). https://doi.org/10.1016/j.jneumeth.2003.10.009
3. Gifford, A.T., Dwivedi, K., Roig, G., Cichy, R.M.: A large and rich eeg dataset for modeling human visual object recognition. NeuroImage **264**, 119754 (2022). https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119754, bluehttps://www.sciencedirect.com/science/article/pii/S1053811922008758
4. Grootswagers, T., Robinson, A.K., Carlson, T.A.: The representational dynamics of visual objects in rapid serial visual processing streams. Neuroimage **188**, 668–679 (Dec 2018). https://doi.org/10.1016/j.neuroimage.2018.12.046
5. Grootswagers, T., Zhou, I., Robinson, A.K., Hebart, M.N., Carlson, T.A.: Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. Scientific Data **9**(1), 3 (Jan 2022). https://doi.org/10.1038/s41597-021-01102-7, bluehttps://doi.org/10.1038/s41597-021-01102-7
6. He, Kaiming: Momentum contrast for unsupervised visual representation learning. arXiv (2020). https://doi.org/10.48550/arXiv.1911.05722DOI
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
8. He, Y., Evans, A.: Graph theoretical modeling of brain connectivity. Curr Opin Neurol **23**(4), 341–350 (Aug 2010). https://doi.org/10.1097/WCO.0b013e32833aa567
9. Hebart, M.N., Dickter, A.H., Kidder, A., Kwok, W.Y., Corriveau, A., Van Wicklin, C., Baker, C.I.: Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. PLOS ONE **14**(10), 1–24 (10 2019). https://doi.org/10.1371/journal.pone.0223792, bluehttps://doi.org/10.1371/journal.pone.0223792
10. Imamoglu, F., Kahnt, T., Koch, C., Haynes, J.D.: Changes in functional connectivity support conscious object recognition. NeuroImage **63**(4), 1909–1917 (2012). https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.07.056, bluehttps://www.sciencedirect.com/science/article/pii/S1053811912007860
11. Intraub, H.: Rapid conceptual identification of sequentially presented pictures. Journal of Experimental Psychology: Human Perception and Performance **7**(3), 604–610 (6 1981). https://doi.org/10.1037/0096-1523.7.3.604DOI
12. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2012). https://doi.org/10.1145/3065386
13. Kubilius, Jonas: Brain-like object recognition with high-performing shallow recurrent anns. arxiv. arXiv.org (10 2019). https://doi.org/10.48550/arXiv.1909.06161DOI
14. Lang, E.W., Tomé, A.M., Keck, I.R., Górriz-Sáez, J.M., Puntonet, C.G.: Brain connectivity analysis: a short survey. Comput Intell Neurosci **2012**, 412512 (Oct 2012). https://doi.org/10.1155/2012/412512DOI
15. Pandey, P., Tripathi, R., Miyapuram, K.P.: Classifying oscillatory brain activity associated with indian rasas using network metrics. Brain Inform **9**(1), 15 (Jul 2022). https://doi.org/10.1186/s40708-022-00163-7

16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, vol. 32, pp. 8024–8035. Curran Associates, Inc (2019)
17. Russakovsky, Olga: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (12 2015). https://doi.org/10.1007/s11263-015-0816-yDOI
18. Xia, Mingrui: Brainnet viewer: A network visualization tool for human brain connectomics. PLoS ONE **8**(7), e68910 (7 2013). https://doi.org/10.1371/journal.pone.0068910DOI

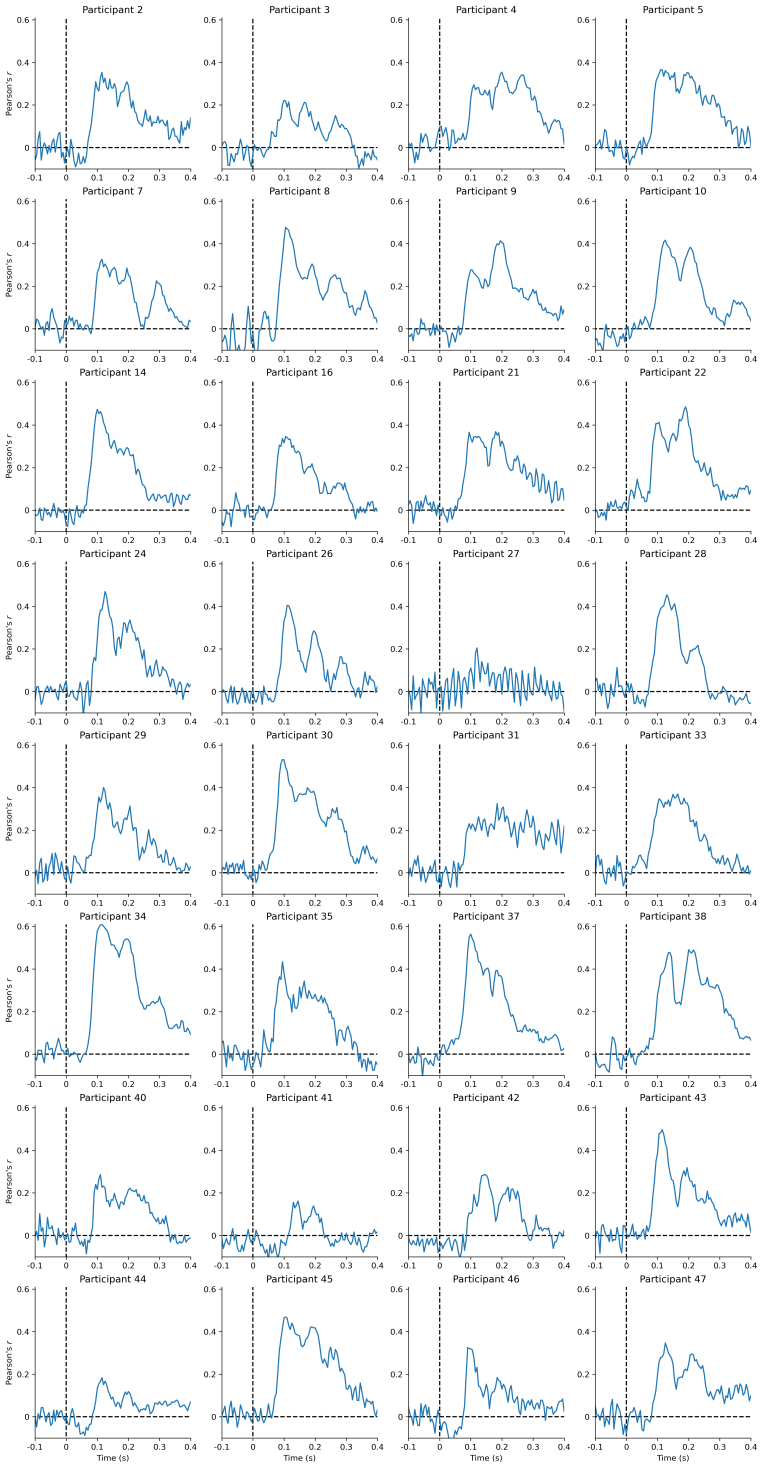


**Fig. 7:** Pearson's-$r$ correlation over time, for linearizing encoding models trained *within* participants.

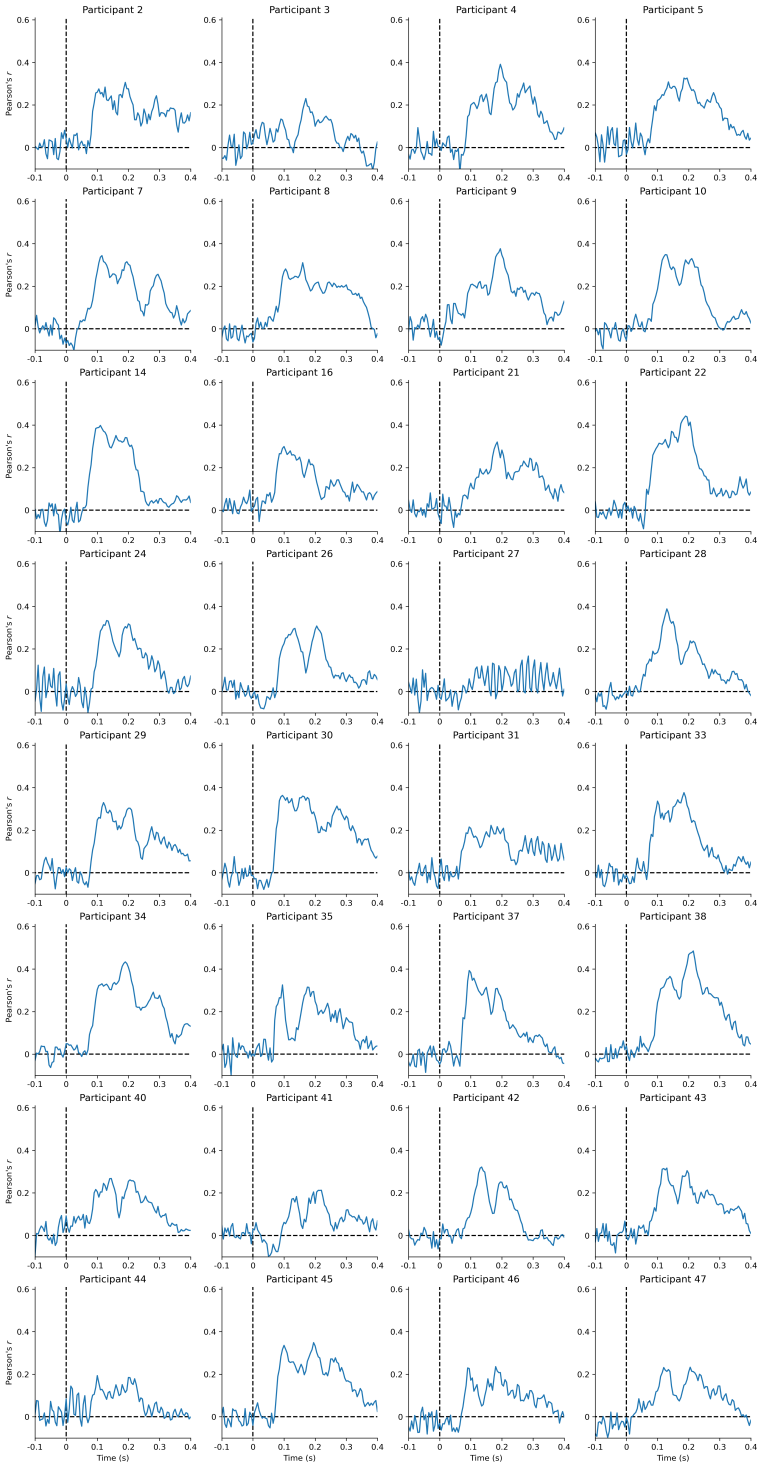


**Fig. 8:** Pearson's-$r$ correlation over time, for linearizing encoding models trained *between* participants.

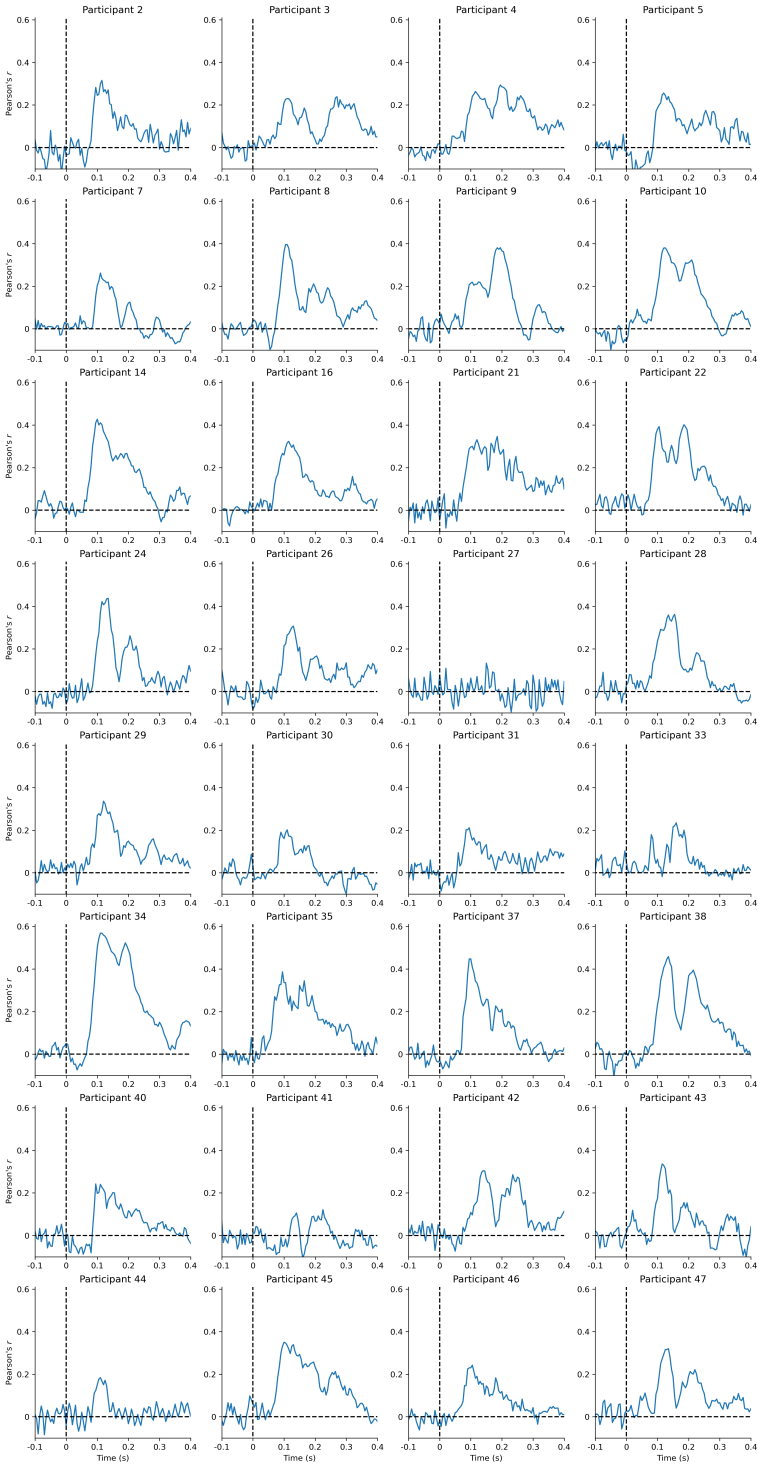


**Fig. 9:** Pearson's-$r$ correlation over time, for the end-to-end encoding models trained *within* participants.