Multiencoder VAE for cross-subject alignment of brain responses

Angeliki Papathanasiou (angeliki.papathanasiou@kellogg.ox.ac.uk)
Centre for Neural Circuits and Behaviour, University of Oxford, 13 Mansfield Road Oxford OX1 3SR, United Kingdom

Jascha Achterberg (jascha.achterberg@dpag.ox.ac.uk)
Centre for Neural Circuits and Behaviour, University of Oxford, 13 Mansfield Road Oxford OX1 3SR, United Kingdom

Ian Cone (ian.cone@dpag.ox.ac.uk)
Centre for Neural Circuits and Behaviour, University of Oxford, 13 Mansfield Road Oxford OX1 3SR, United Kingdom

Thomas E. Nichols (thomas.nichols@bdi.ox.ac.uk)
Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
Nuffield Department of Population Health, University of Oxford, Old Road Campus Oxford OX3 7LF, United Kingdom

Rui Ponte Costa (rui.costa@dpag.ox.ac.uk)
Centre for Neural Circuits and Behaviour, University of Oxford, 13 Mansfield Road Oxford OX1 3SR, United Kingdom

Abstract

Neural responses to identical stimuli vary considerably across individuals despite similar behavioral outcomes. Recent research demonstrates preserved latent neural dynamics in motor cortical populations across monkeys performing identical motor tasks. Inspired by these observations we introduce a multiencoder variational autoencoder (VAE) to model visual cortex responses. Our approach transforms subject-specific fMRI responses from natural scene viewing into a common latent space while predicting artificial neural network (ANN) activations elicited by identical stimuli. Using the Natural Scenes Dataset (NSD), our method outperforms traditional alignment techniques by capturing cross-subject representational similarities. The VAE architecture implements subject-specific encoders which project occipitotemporal cortex responses into a shared latent manifold that preserves semantic organization while accommodating neuroanatomical variability. Simultaneously, the decoder establishes a computational correspondence between this latent representation and ResNet-50 activations. This approach creates a framework for investigating shared neural representations across individuals while quantifying systematic relationships between biological and artificial NNs.

Keywords: representational alignment; cross-subject alignment; fMRI; variational autoencoders

Introduction

Neural circuits exhibit substantial individual variability, yet humans demonstrate remarkably similar perceptual capabilities. Safaie et al. (2023) demonstrated that low-dimensional neural population dynamics in motor cortex are preserved across animals performing identical tasks, suggesting a species-wide lower-dimensional "neural landscape" that constrains possible neural activity patterns. Extending this principle to human visual processing presents methodological challenges due to fMRI's temporal limitations and inter-subject variability in functional organization (Finn et al., 2020). While traditional alignment methods including Canonical Correlation Analysis (CCA) (Thompson, 2000), Partial Least Squares (PLS), and Procrustes analysis (Ross, 2004) offer linear transformations, they inadequately capture complex inter-subject neural relationships (Helmer et al., 2024).

To address these challenges we developed a multiencoder VAE framework to align fMRI responses across subjects viewing identical natural scenes. Our approach employs subject-specific encoders that map individual neural responses to a common latent space, establishing cross-subject alignment within this shared representation. Simultaneously, and to achieve this, we train the decoder to reconstruct activations from a pre-trained ResNet-50 convolutional neural network (pre-trained on ImageNet) presented with the same images, creating a computational bridge between human visual cortex activity and artificial neural representations. This architecture maintains distinct pathways: subject-specific encoders project fMRI data to a shared latent space for cross-subject alignment, while the decoder maps from this latent space to ResNet-50 activations, establishing correspondence between biological and artificial vision systems. The decoding to ResNet-50 activations acts as a unifying force in the latent space.

Methods

Dataset and Preprocessing

We utilized the Natural Scenes Dataset, comprising fMRI recordings from 8 participants each viewing 10,000 natural images across multiple sessions. Cross-subject alignment analyses focused on 872 images viewed by all participants, while VAE training leveraged the complete dataset ($\sim$ 70,500 distinct images). Neural responses were restricted to occipitotemporal cortex (OTC) voxels following established anatomical criteria in (Conwell, Prince, Kay, Alvarez, & Konkle, 2022), resulting in 20,732 voxels per subject.

Multiencoder VAE Architecture

Our framework employs separate encoder networks for each subject that map individual fMRI patterns to a shared 64-dimensional latent space. The decoder network reconstructs activations from a pre-trained ResNet-50 model (8,000 units) presented with the same images. This architecture explicitly models subject-specific transformations while enabling a common representational framework. The model was trained with a VAE loss function and optimized using Adam (learning rate=$1e-4$) with hidden layer dimensionality of 512.

Evaluation Metrics

We evaluated our approach using both reconstruction quality and latent space organization metrics on held-out test data (10% of the dataset):

1. **Reconstruction loss** to assess cross-domain mapping quality, i.e., correspondence between biological and artificial visual systems.

2. **Silhouette score**, quantifies semantic clustering integrity within the latent space by measuring how category-specific neural response patterns maintain distinctiveness during dimensionality transformation. Higher values indicate preserved stimulus-category boundaries (e.g., animals, vehicles) within the shared representational manifold, validating the retention of behaviorally relevant topographic organization during the encoding process.

3. **Cross-subject alignment metrics:**

   (a) *Component-wise correlation:* Correlation of individual latent dimensions across subjects.

   (b) *Inter-subject correlation (ISC):* Similarity of neural representations for identical stimuli across subjects.

   (c) *Representational similarity analysis (RSA):* Preservation of stimulus relationship patterns.

We implemented CCA, PLS and Procrustes alignment after dimensionality reduction of the raw fMRI data, namely, reduction using PCA to 200 dimensions followed by UMAP to 10 dimensions. For the VAE, the 64-dimensional latent space was reduced to 10 dimensions using UMAP for fair comparison.

Results

Training Dynamics and Bounds

The VAE framework demonstrated clear convergence properties, with reconstruction loss diminishing progressively during training (Figure 1B). We established theoretical bounds for performance based on the inherent properties of the cross-domain mapping problem. The NN→NN mapping established the theoretical performance ceiling by eliminating cross-domain translation requirements, while shuffled fMRI→NN mappings, i.e., deliberately destroying the correspondence between inputs and outputs, defined the upper error bound, as the random baseline. Our multiencoder model significantly outperformed the random baseline, confirming effective cross-domain representational alignment.

For latent space organization assessment (Figure 1C), the NN→NN configuration provided the optimal clustering benchmark due to the ANN's inherently higher signal fidelity and deterministic category-specific response properties. Conversely, the fMRI→fMRI configuration established the lower bound, as a result of the fMRI signal's characteristics, e.g., physiological artifacts, state-dependent variability. Our model's silhouette scores consistently positioned between these boundaries, validating the preservation of semantic clustering despite inherent biological signal variability.

Cross-Subject Alignment

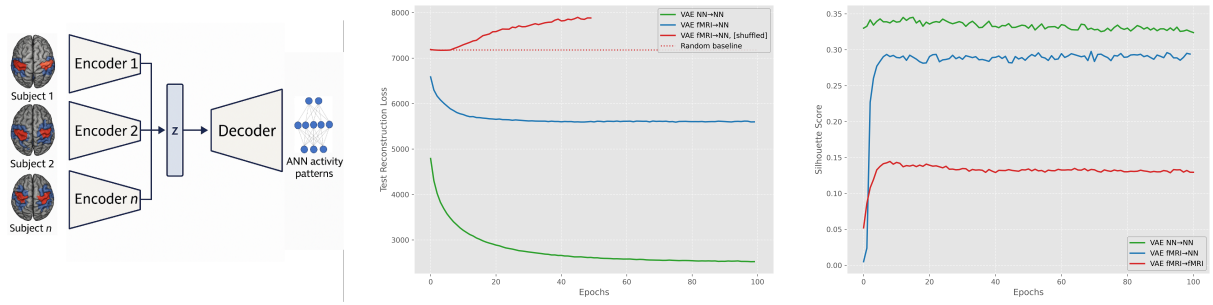
The multiencoder VAE outperformed alternative alignment methods across all metrics, as seen in Table 1. For reference, RSA on the raw data is 0.3866 and after dimensionality reduction, i.e., PCA followed by UMAP, RSA=0.4553. Notably, we attempted to enhance conventional methods by concurrently aligning all subject fMRI data with ANN activations to determine if this mechanism could drive superior performance as observed in our VAE framework; however, this modification actually diminished alignment metrics for CCA, PLS and Procrustes approaches.

Table 1: **Comparison of Cross-subject Alignment Metrics**

Method	Comp-wise Corr.	ISC	RSA
VAE	**0.8298**	**0.9558**	**0.7824**
Procrustes	0.5128	0.9314	0.4762
PLS	0.3780	0.4473	0.4250
CCA	0.2919	0.4520	0.4225

Discussion & Conclusion

Our multiencoder VAE framework outperforms conventional alignment methods in capturing shared neural representations across subjects. The performance advantages across all quantitative metrics demonstrate that complex transformations are essential for effective cross-subject fMRI alignment. High inter-subject correlation values reveal that despite idiosyncratic neural activity patterns, individuals share fundamental representational architectures for visual processing—extending Safaie et al.'s findings from motor control to visual perception. The decoding to ResNet-50 activations acts as a unifying force in the latent space, while establishing computational correspondence between fMRI responses and ResNet-50 activations and creating a principled bridge between biological and artificial visual systems. Future work should explore correlations between preserved neural dynamics and behavioral performance, while, also, investigate this framework as a methodology for mapping bidirectional relationships between biological and artificial NNs.



A B C

Figure 1: **Multiencoder VAE achieves cross-subject alignment.** (**A**) Schematic of multiencoder-VAE framework. (**B,C**) Model learning curves with loss (B) and silhouette score (C). Here we show results for the following models: (B-C) blue: VAE fMRI $\rightarrow$ NN, green: VAE NN $\rightarrow$ NN, (B) red: VAE fMRI $\rightarrow$ NN after shuffling the order, (C) red: VAE fMRI $\rightarrow$ fMRI

Acknowledgements

A.P. is supported by the EPSRC Centre for Doctoral Training in Health Data Science [EP/S02428X/1].

References

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*, 116828.

Helmer, M., Warrington, S., Mohammadi-Nejad, A.-R., Ji, J. L., Howell, A., Rosand, B., ... Murray, J. D. (2024). On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications biology*, *7*(1), 217.

Ross, A. (2004). Procrustes analysis. *Course report, Department of Computer Science and Engineering, University of South Carolina*, *26*, 1–8.

Safaie, M., Chang, J. C., Park, J., Miller, L. E., Dudman, J. T., Perich, M. G., & Gallego, J. A. (2023). Preserved neural dynamics across animals performing similar behaviour. *Nature*, *623*(7988), 765–771.

Thompson, B. (2000). Canonical correlation analysis.