Learning Latent Spaces for Individualized Functional Neuroimaging with Variational Autoencoders

Kajal Singla (singla@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig/Dresden, Germany

Dr. Pierre-Louis Bazin (piloubazin@fullbrainpicture.nl)
Full brain picture Analytics, Leiden, The Netherlands

Dr. Nico Scherf (nscherf@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig/Dresden, Germany

Abstract

Functional Magnetic Resonance Imaging (fMRI) studies often use dimensionality reduction methods like independent component analysis or diffusion map embedding to identify group-level brain networks and dynamics. These approaches struggle to capture individual-specific differences. To address this gap, we explore the use of variational autoencoders (VAEs) to model Blood Oxygen Level Dependent (BOLD) signals in a subject-specific latent space. Our approach effectively denoises fMRI data using a compressed, low-dimensional latent representation, enhancing the separation of signals from distinct functional networks without directly aligning them to specific latent axes. While direct alignment of latent dimensions across subjects is not straightforward, we observe shared geometric patterns across subjects' latent spaces, enabling meaningful cross-subject comparisons. Deep latent modeling offers a promising avenue for individualized fMRI analysis, providing new insights into the brain's complex functional architecture.

Introduction

Functional Magnetic Resonance Imaging (fMRI) measures brain activity over time by detecting changes in blood flow. Functional networks and brain states can be revealed using all three fMRI paradigms (rs: resting-state, tb: task-based, and naturalistic). Naturalistic and tb-fMRI paradigms offer the advantage of better experimental control and engagement compared to rs-fMRI, allowing for the identification of time-locked neural activity patterns that are comparable across subjects.

In this study, we leverage a naturalistic fMRI dataset (Finn et al., 2018), where participants listened to a narrative. Using minimally preprocessed fMRI data, we explore the capacity of deep latent models—specifically Convolutional Variational Autoencoders (CVAEs) (Wang et al., 2024) to reveal individual-specific patterns in the Blood Oxygen Level Dependent (BOLD) signals. CVAEs are effective for fMRI time series analysis because their 1D convolutional architecture captures temporal patterns and dependencies in BOLD signals while reducing noise through dimensionality reduction, whereas the variational component allows us to model the underlying probability distribution of BOLD signals, accounting for the inherent variability in neural responses across time and subjects. Additionally, by processing the time course for each voxel independently, the model can learn consistent temporal features across the brain while preserving spatial variations that reflect individual-specific functional organization.

Methods

Dataset We re-analyzed a naturalistic fMRI dataset from OpenNeuro (ds001338) featuring 22 healthy participants (Finn et al., 2018). During the experiment, participants listened to an original 22-minute audio narrative describing an ambiguous social scenario designed to elicit varying interpretations and emotional responses. We worked with the minimally preprocessed version of the data provided by the Naturalistic Data Analysis repository. Preprocessing involved `fmriprep` for standard corrections, followed by smoothing and General Linear Model (`GLM`) fitting.

The preprocessed fMRI data consists of 4D whole-brain volume activations over time. We reshaped the data into a 2D matrix with dimensions (voxels × time points) for each subject. To extract the time series and anatomical labels of relevant voxels, we used the Schaefer atlas (2018) with 200 parcellations mapped to the Yeo 17 networks (Schaefer et al., 2018) and applied the MNI152NLin2009cAsym brain mask. This resulted in a 139,501 (voxels) × 1,310 (time points) matrix for each participant. We normalized the BOLD time series to a range of [0,1] using the MinMaxScaler from `sklearn`.

Model Architecture and Training We used a 1D CVAE to process the temporal BOLD signals. Both the encoder and decoder networks consist of eight convolutional layers and two fully connected layers. The data for each subject was randomly split into training (70%), validation (10%), and test (20%) sets. We trained the models using mini-batches, where each batch comprised 280 voxels × 1,310 time points. The standard VAE loss function combines a Reconstruction Loss and a KL Divergence (Kingma et al., 2019). However, for this study, we set the KL divergence term to zero for two main reasons: (i) Training Stability: KL divergence often requires annealing strategies to ensure stable training (Joas et al., 2024).

(ii) Focus on Latent Structure: Our primary goal was not to train a generative model but to analyze the latent embeddings produced by the encoder. We trained a separate CVAE model for each of the 22 subjects to capture subject-specific latent representations. The models were implemented using PyTorch Lightning and trained on an A100 GPU. We used Adam with a learning rate of 0.001 and early stopping with a patience of 100 epochs.

To explore the impact of latent space dimensionality on reconstruction quality, we trained the CVAE for each subject using latent dimensions ranging from 2 to 16. We evaluated the reconstruction performance on the test set for each configuration, measuring the Mean Squared Error (MSE) between the input and reconstructed BOLD signals. Trading off reconstruction fidelity and model complexity, we empirically selected a 9D latent space for all analyses.

Results

Deep autoencoders for BOLD signal denoising To assess the denoising performance of the CVAE, we analyzed reconstructed BOLD signals from one subject (Subject 1). We compared the mean squared error (MSE) between input and reconstruction on a subset of 2,000 randomly selected voxels. The average MSE in the sample was 0.0011, indicating that the CVAE effectively reconstructs the temporal dynamics of the BOLD signals. We conducted a cross-subject analysis for validation, where the second subject's data was passed through the pre-trained model of the first subject. The MSE error between the input and the decoded data was 0.019, about an order of magnitude higher than the previous result, indicating that learned latent embeddings of BOLD signals are not trivially aligned.

To check for spatial inhomogeneities in reconstruction, we examined whole-brain reconstructions at an arbitrary time point ($t = 900$), which produced an empirical mean square error (EMSE) of 0.0015 between the input and reconstructed volumes. This suggests that the reconstruction quality is consistent across both voxels and time points. To quantitatively assess the denoising capability of the CVAE, we computed the temporal Signal-to-Noise Ratio (tSNR) for each voxel. In Subjects 1–5, the reconstructed BOLD signals demonstrated higher tSNR values, confirming that CVAE effectively reduces noise while preserving relevant neural signals. The average tSNR improvements were 4.98%, 4.83%, 6.46%, 4.61%, and 5.36%, respectively. These improvements in tSNR were statistically significant ($p < 1\mathrm{e}{-308}$, Wilcoxon signed rank test) in all cases. These results demonstrate that CVAE can successfully embed, reconstruct, and denoise BOLD time series at the voxel level. The learned latent representations not only capture essential temporal dynamics, but also enhance signal quality.

The structure of encoded latent representations We further assessed relationships between latent dimensions by calculating pairwise correlations. The correlations were relatively low (all below 0.3), suggesting that latent dimensions capture generally distinct features, despite the absence of orthogonality constraints. Compared to diffusion embedding, which typically emphasizes large-scale cortical gradients, our CVAE-based representations appear to reflect more individual-level variations and finer-scale structures. PCA and ICA visualizations of the latent space show a more uniform distribution of voxels with respect to the Yeo 17 networks, suggesting that CVAE reduces noise and captures generalized patterns across networks. However, we did not observe clear clustering based on network labels, indicating that latent space preserves functional structure without explicitly aligning to predefined network boundaries. In general, CVAE encodes BOLD signals in a low-dimensional space that captures smooth, anatomically coherent patterns while maintaining minimal correlations between dimensions. Although the latent dimensions do not directly map onto known functional networks, they appear to retain aspects of brain organization at the subject level.

Aligning latent representations across subjects Since each subject was trained independently with a CVAE, the resulting latent representations were not initially aligned across subjects, leading to subject-specific embeddings. To examine the alignment of latent spaces across subjects, we applied Orthogonal Procrustes (Sasse et al., 2024) analysis to evaluate the consistency of the learned embeddings. The latent spaces did not align well between the individuals, probably reflecting differences in brain function rather than random noise, suggesting that the embeddings capture meaningful, subject-specific information. This misalignment indicates that our method preserves individual differences that would be lost in models forcing cross-subject alignment during training.

Discussion

In this study, we demonstrated that convolutional variational autoencoders (CVAEs) can effectively denoise BOLD signals at the voxel level and generate subject-specific latent representations. These embeddings 1) accurately reconstruct fMRI time series, 2) capture brain organization patterns that improve the separability of Yeo17 networks without strictly aligning to known gradients, 3) provide a reference frame for comparing latent spaces across subjects, and 4) enable geometric methods to assess functional connectivity. Beyond denoising, CVAEs offer a flexible framework for embedding fMRI data into lower dimensional latent spaces useful for visualization and similarity analysis. They also hold potential for detecting outliers and generating new BOLD samples—areas that warrant future exploration.

In future work, we aim to examine the influence of the KL divergence term on reconstruction performance and to explore the extent to which it affects the separability of latent representations (Higgins et al., 2017). A key limitation of our approach is the need to train separate models for each subject, which limits scalability. A potential solution is to develop models that learn shared embedding spaces across subjects Huang et al. (2022), enabling generalization to new participants and making the method more practical for large-scale studies.

Acknowledgments

N.S. and K.S. are supported by BMBF (Federal Ministry of Education and Research) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI). N.S. is supported by BMBF through ACONITE (01IS22065) and by the European Union and the Free State of Saxony through BIOWIN. K.S. is supported by the Max Planck IMPRS CoNI Doctoral Program. We also acknowledge the MPCDF for providing access and support to computing infrastructure.

References

Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., & Constable, R. T. (2018). Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature communications*, *9*(1), 2043.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., . . . Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, *3*.

Huang, J., Busch, E., Wallenstein, T., Gerasimiuk, M., Benz, A., Lajoie, G., . . . Krishnaswamy, S. (2022). Learning shared neural manifolds from multi-subject fmri data. In *2022 ieee 32nd international workshop on machine learning for signal processing (mlsp)* (pp. 01–06).

Joas, M., Jurenaite, N., Praščević, D., Scherf, N., & Ewald, J. (2024). A generalized and versatile framework to train and evaluate autoencoders for biological representation learning and beyond: Autoencodix. *bioRxiv*, 2024–12.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392.

Sasse, L., Paquola, C., Dukart, J., Hoffstaedter, F., Eickhoff, S. B., & Patil, K. R. (2024). Procrustes alignment in individual-level analyses of functional gradients. *bioRxiv*, 2024–11.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., . . . Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, *28*(9), 3095–3114.

Wang, Y., Li, D., Li, L., Sun, R., & Wang, S. (2024). A novel deep learning framework for rolling bearing fault diagnosis enhancement using vae-augmented cnn model. *Heliyon*, *10*(15).