FMRI Data Analysis Preserving Map Variability Via Unsupervised Object-Centric Learning

Rui Jin and Seung-Jun Kim*
Dept. of Computer Science & Electrical Engineering
University of Maryland, Baltimore County, Baltimore, MD 21250
E-mail: {rjin1,sjkim}@umbc.edu

***Abstract*—A novel data-driven functional magnetic resonance imaging (fMRI) data analysis method is proposed using a deep object-centric learning paradigm. The method can faithfully estimate the variabilities in the spatial neural activation maps, which capture functional interconnections in the brain, over fMRI volumes. The key idea is to treat the component maps composing individual fMRI volumes as "objects," whose latent representations are separately learned by a set of autoencoders. Numerical tests using synthetic and real data sets verify the advantages of the proposed method compared to existing matrix factorization-based approaches.**

***Index Terms*—Deep learning, dynamic functional connectivity, fMRI data analysis, object-centric learning, subject individuality.**

I. Introduction

Functional magnetic resonance imaging (FMRI) can reveal brain neural activities non-invasively. The fMRI data can be analyzed based on pre-defined regions-of-interest (ROIs) or reference signals in a hypothesis-driven manner. While readily interpretable, the results may be limited by the a priori modeling assumptions. Alternatively, data-driven approaches such as the independent component analysis (ICA) and dictionary learning techniques allow fully multivariate analysis in a matrix factorization framework [1].

A tacit assumption made in such algorithms is the existence of a common signal subspace over fMRI volumes. This, however, inevitably neglects possible variabilities of functional networks. In a similar framework, a group analysis of fMRI data can uncover common functional networks among multiple subjects [2], [3]. Notably, group/subject-specific spatial activation maps can also be obtained. Still, the maps are typically assumed to be fixed within the set of volumes corresponding to individual groups. Estimating fine-grained map variability is challenging but can reveal informative brain dynamics [4].

Recently, deep learning techniques have been actively employed for fMRI data analysis. A variational autoencoder (VAE) was used for resting-state fMRI data to analyze dynamic changes in neural activities [5]. A deep Markov factor analysis model was developed to explain spatial and temporal patterns, yet assuming common spatial activations across time [6]. A transformer framework was utilized to analyze fMRI scans, and shown to perform well for age and gender prediction tasks [7]. However, these works did not explicitly estimate spatial activation maps that preserve variabilities across volumes.

Object-centric learning aims at improving sample efficiency and generalization capability of machine learning models by decomposing complex structures in the input data into interactions of multiple objects [8]. In the compositional scene models for image and video data, for instance, individual objects constituting a scene can be segmented in an unsupervised fashion and their latent variables obtained by encoders with shared parameters across different objects [9]–[11]. The variability in the attributes of the objects such as the shape, color, and texture, can be readily preserved in the latent representations and reconstructed by decoders.

In this work, a fMRI data analysis method is developed based on the object-centric learning paradigm by treating the spatial activation maps that compose the individual fMRI volumes as objects. Inspired by [10], [11], an attention mechanism is adopted which generates spatial masks recursively for efficient downstream processing. Then, VAEs are adopted to extract individual component maps from a fMRI volume using the masks. This way, the model can learn a manifold structure, rather than a rigid common subspace, that can flexibly preserve map variabilities across volumes. The proposed method was tested with both synthetic and real data sets and compared with existing algorithms to verify the benefit of the approach.

The rest of the paper is organized as follows. The fMRI data analysis problem is stated in Sec. II. The proposed method is presented in Sec. III. The employed DNN architecture is described in Sec. IV. The evaluation results are presented in Sec. V. Conclusions are provided in Sec. VI.

II. FMRI Data Analysis With Map Variability

Let $\mathcal{X} := \{\mathbf{x}[n] \in \mathbb{R}^V : n = 1, \ldots, N\}$ be a set of N fMRI volumes with V voxels. Upon constructing a matrix $\mathbf{X} \in \mathbb{R}^{N\times V}$ whose n-th row is $\mathbf{x}[n]$, a matrix factorization approach for fMRI data analysis decomposes $\mathbf{X}$ into $\mathbf{AU}$, where $\mathbf{A} \in \mathbb{R}^{N\times K}$ is a matrix of activation coefficients $\{a_k[n]\}$ and $\mathbf{U} \in \mathbb{R}^{K\times V}$ contains the spatial activation maps $\mathcal{U} := \{\mathbf{u}_k \in \mathbb{R}^V\}$ as the rows. That is, a representation $\mathbf{x}[n] \approx \sum_{k=1}^K a_k[n]\mathbf{u}_k$ is obtained for each n. The identifiability of the factors often hinges on additional assumptions such as statistical independence or sparsity on $\mathbf{U}$ and orthogonality or norm constraints on $\mathbf{A}$ [12], [13]. Furthermore, since a single

*Corresponding author. This work was supported in part by US National Science Foundation grants 1631838 and 2242412.

set $\mathcal{U}$ is shared for all n, it can be noted that the variability of the component maps across the volumes is neglected.

To account for the map variability, in this work we aim at a representation given by $\mathbf{x}[n] \approx \sum_{k=1}^{K} a_k[n]\mathbf{u}_k[n]$, where a map $\mathbf{u}_k[n]$ may depend on n. In fact, one can simply estimate the product $\mathbf{s}_k[n] := a_k[n]\mathbf{u}_k[n]$, leading to the representation

$$\mathbf{x}[n] \approx \sum_{k=1}^{K} \mathbf{s}_k[n],\ n = 1, \ldots, N. \tag{1}$$

Obviously, the set of maps $\mathcal{S}[n] := \{\mathbf{s}_k[n]\}$ needs to be estimated in such a way that there is consistency of the maps across n. This prior is naturally encouraged in our object-centric learning model, where the individual maps are taken as "objects" and are encoded to latent vectors using K encoders. This representational bottleneck enforces learning of consistent yet variability-preserving maps while allowing efficient representation of the input.

III. Object-Centric Learning Formulation

We adopt the deep object-centric learning paradigm to estimate the maps $\mathcal{S}[n]$ for each fMRI volume $\mathbf{x}[n]$. Since the size V of the voxel space is quite large, an attention mechanism is first employed to figure out roughly the parts in the input related to the component activations. A recurrent architecture is designed to obtain the attention masks for the components sequentially. Then, the component maps are extracted from the input through a VAE architecture.

A. Attention Mechanism

The attention mechanism produces $K+1$ masks $\mathcal{M} := \{\mathbf{m}_k \in [0,1]^V\}_{k=1}^{K+1}$ from $\mathbf{x}$. (The volume index n is dropped for notational simplicity.) For $k = 1, \ldots, K$, the v-th element $m_{k,v}$ of $\mathbf{m}_k$ models the probability that the v-th voxel in $\mathbf{x}$ represents the neural activity corresponding to the k-th component map $\mathbf{s}_k$. Mask $\mathbf{m}_{K+1}$ corresponds to the background (non-gray matter) areas. Thus, upon denoting the component map/background index related to the v-th voxel as C_v, we have $m_{k,v} = P[C_v = k|\mathbf{x}]$ for $k = 1, \ldots, K+1$, and $\sum_{k=1}^{K+1} m_{k,v} = 1$ holds for all v. Define $\mathbf{c} \in \{1, \ldots, K+1\}^V$ to be a vector collecting $\{C_v\}$, and $q_{\boldsymbol{\psi}}(\mathbf{c}|\mathbf{x})$ the probability mass function (PMF) represented by $\mathcal{M}$.

To generate all masks efficiently, they are obtained sequentially using a recurrent architecture. Let $\mathbf{r}_{k-1} \in [0,1]^V$ represent the scope in which the activations not captured by the masks generated in the previous $k-1$ stages. Then, a deep neural network (DNN) $\boldsymbol{\alpha}_{\boldsymbol{\psi}}$ parameterized by $\boldsymbol{\psi}$ is used to generate the k-th mask as

$$\mathbf{m}_k = \mathbf{r}_{k-1} \odot \boldsymbol{\alpha}_{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{r}_{k-1}),\ k = 1, \ldots, K \tag{2}$$

where $\odot$ denotes element-wise multiplication. The background mask is simply given by $\mathbf{m}_{K+1} = \mathbf{r}_K$, which ensures that all voxels are accounted for, i.e., $\sum_{k=1}^{K+1} \mathbf{m}_k = \mathbf{1}$, where $\mathbf{1}$ is an all-one vector. With $\mathbf{r}_0 = \mathbf{1}$, the scopes are updated as

$$\mathbf{r}_k = \mathbf{r}_{k-1} \odot [\mathbf{1} - \boldsymbol{\alpha}_{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{r}_{k-1})],\ k = 1, \ldots, K. \tag{3}$$

B. VAE for Component Maps

A VAE is employed to extract component maps $\mathcal{S}$ from $\mathbf{x}$ with the help of the masks $\mathcal{M}$ [14]. First, an approximate posterior distribution of the latent vectors $\mathcal{Z} := \{\mathbf{z}_k \in \mathbb{R}^L\}$ for the maps is estimated through variational inference. Specifically, it is assumed that $\mathcal{Z}$ has a standard Gaussian prior $p(\mathcal{Z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{z}_k; \mathbf{0}, \mathbf{I})$, and the posterior for $\mathbf{z}_k$ can be approximated as Gaussian with mean and covariance obtained from $\mathbf{x}$ and $\mathbf{m}_k$ via a DNN parameterized by $\boldsymbol{\phi}$. That is,

$$q_{\boldsymbol{\phi}}(\mathbf{z}_k|\mathbf{x}, \mathbf{m}_k) = \mathcal{N}(\mathbf{z}_k; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{m}_k), \mathrm{diag}\{\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{m}_k)\}) \tag{4}$$

where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix with the diagonal entries listed in $\{\cdot\}$. Since $\mathbf{m}_k$ is itself estimated from $\mathbf{x}$ (cf. Sec. III-A), one can rewrite $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{m}_k) = \boldsymbol{\mu}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x})$ and $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{m}_k) = \boldsymbol{\sigma}^2_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x})$. Upon assuming conditional independence of the latents, the posterior of $\mathcal{Z}$ is expressed as

$$q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathcal{Z}|\mathbf{x}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{z}_k; \boldsymbol{\mu}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x}), \mathrm{diag}\{\boldsymbol{\sigma}^2_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x})\}). \tag{5}$$

The generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{Z})$ is parameterized by a decoder DNN $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$. In particular, the DNN generates the component maps as $f_{\boldsymbol{\theta}}(\mathbf{z}_k) = \mathbf{s}_k$ for $k = 1, \ldots, K$. Then, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{Z})$ is modeled as a voxel-wise independent Laplace distribution with location $\boldsymbol{\mu}_{\boldsymbol{\theta}} := \sum_{k=1}^{K} \mathbf{s}_k$ and scale b, which mitigates blurriness in the reconstructed maps [15]. That is,

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{Z}) = \prod_{v=1}^{V} \frac{1}{2b} \exp\left(-\frac{|x_v - \mu_{\boldsymbol{\theta},v}|}{b}\right) \tag{6}$$

where x_v and $\mu_{\boldsymbol{\theta},v}$ are the v-th entries of $\mathbf{x}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, respectively. For simplicity, b is pre-specified in this work.

C. Training Loss Function

The overall training is done using a loss function based on the β-VAE loss augmented by a term constraining the masks. First, the loss function per sample for β-VAE is given by [16]

$$\begin{aligned} l_{\beta\text{-VAE}}(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{x}) = &- \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathcal{Z}|\mathbf{x})} \{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{Z})\} \\ &+ \beta \mathcal{D}(q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathcal{Z}|\mathbf{x})||p(\mathcal{Z})) \end{aligned} \tag{7}$$

where $\mathcal{D}(\cdot||\cdot)$ denotes Kullback-Leibler (KL) divergence and $\beta > 0$ is a parameter encouraging disentanglement of latent variables. The first term in (7) corresponds to the reconstruction error of the VAE and the second term is a regularizer that ensures that the approximate posterior is close to the prior. The expectation can be evaluated using the reparameterization trick to facilitate the computation of gradients for training [14].

It is also necessary to encourage the attention masks $\mathcal{M}$ to match well with the estimated map areas as otherwise the masks are not spatially constrained. For this, a PMF $p_{\boldsymbol{\theta}}(\mathbf{c}|\mathcal{Z})$ represented by $\{\pi_{k,v} = P[C_v = k|\mathcal{Z}], k = 1, \ldots, K+1, v = 1, \ldots, V\}$ is constructed from $\mathcal{S}$ as

$$\pi_{k,v} = \frac{|s_{k,v}|}{\max_{v'} \sum_{k=1}^{K} |s_{k,v'}|},\ k = 1, \ldots, K \tag{8}$$

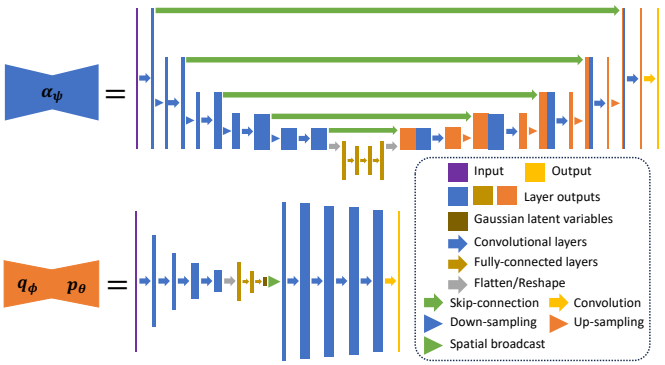

Fig. 1: Employed DNN architecture.

and $\pi_{K+1,v} = 1 - \sum_{k=1}^{K} \pi_{k,v}$ for all v. Then, the loss for the masks is defined as

$$l_{\text{mask}}(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathcal{Z}|\mathbf{x})} \left\{ \mathcal{D}(q_{\boldsymbol{\psi}}(\mathbf{c}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{c}|\mathcal{Z})) \right\}. \quad (9)$$

The overall loss function for sample $\mathbf{x}$ is then given by

$$l(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{x}) = l_{\beta\text{-VAE}}(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{x}) + \gamma l_{\text{mask}}(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{x}) \quad (10)$$

where γ is a positive weight parameter.

IV. DNN Architecture

The attention network $\boldsymbol{\alpha}_{\boldsymbol{\psi}}$ is based on the U-Net architecture as shown in Fig. 1 [17]. Each convolutional layer includes 3D bias-free convolutions with kernels of size $3 \times 3 \times 3$, stride 1, and one zero padding, followed by a group normalization over 8 groups and a ReLU activation. The spatial down- and up-sampling operations are done by a factor of 2. The output of the encoding path is passed through fully-connected (FC) layers with a hidden dimension of 128. The numbers of kernels are $[64, 128, 256, 512, 512]$ in the encoding path and $[512, 256, 128, 64, 64]$ in the decoding path. Then, a $1 \times 1 \times 1$ convolution and the sigmoid nonlinearity are applied.

The encoder $q_{\boldsymbol{\phi}}$ of the VAE gets the fMRI volume $\mathbf{x}$ and the log of the mask $\log \mathbf{m}_k$ as the input, which are processed by convolutional layers consisting of 3D convolutions with kernels of size $3 \times 3 \times 3$, stride 2, one zero padding, group normalization, and ReLU activation. The numbers of the kernels are $[32, 32, 64, 64]$. Then, one FC layer having an output dimension of 128 with ReLU and layer normalization, followed by a linear FC layer, is employed. The dimension L of the latent vector $\mathbf{z}_k$ was set to 16. The decoder $p_{\boldsymbol{\theta}}$ first employs a spatial broadcast on $\mathbf{z}_k$ [18]. Thus, spatial copies of $\mathbf{z}_k$ are made in 3D (with padding for subsequent convolutions), which are concatenated by the 3D coordinates normalized to the interval $[-1, 1]$. The resulting tensor is processed by convolutional layers with the same configuration as in the encoder except that the stride is 1 and there is no padding. In each layer, 64 kernels are employed. The output has the same size as the fMRI volume, which is processed by a $1 \times 1 \times 1$ convolution to obtain the desired $\mathbf{s}_k$.

The model was trained using the RMSProp optimizer with batch size 32. To avoid the KL divergences vanishing at the early training stage, β and γ were gradually increased from 0 to 0.5 over the first 20 epochs [19].

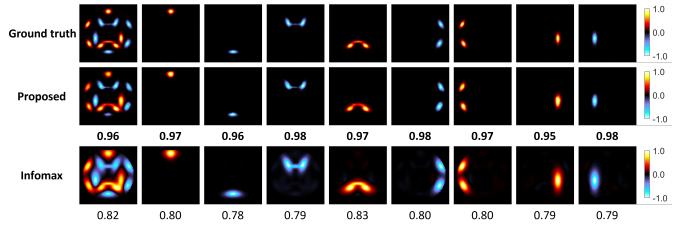

Fig. 2: Maps from the proposed and the Infomax algorithms.

V. Evaluation

A. Evaluation with Synthetic Data

The method was evaluated with both synthetic and real fMRI data sets. The simulated fMRI volumes were generated using $K = 8$ component maps $\{\mathbf{u}_k[n]\}$ in 2D of size $V = 64 \times 64$ in SimTB [20]. The spatial spread of the maps was varied over n by sampling the spatial spread parameter in SimTB from a uniform distribution over $[0.25, 1.75]$. Each volume was generated by $\mathbf{x}[n] = \sum_{k=1}^{K} a_k[n]\mathbf{u}_k[n]$, where $a_k[n]$ was sampled from a Bernoulli (with $p = 0.5$)-Uniform (over $[-1, 1]$) distribution. A total of $N = 71,000$ volumes ($70,000$ for training and $1,000$ for testing) were generated.

For comparison, the Infomax algorithm for ICA was also tested [21]. The mean-square error (MSE) between the ground truth maps and the estimated maps was calculated. The proposed algorithm and the Infomax algorithm yielded -38 dB and -35 dB, respectively, for the map estimation MSE. The MSE was also computed between the input volumes $\mathbf{x}[n]$ and the reconstructed volumes $\sum_{k=1}^{K} \mathbf{s}_k[n]$. Our method yielded -30 dB against -28 dB of Infomax for the volume reconstruction MSE. The performance advantage is due to the ability to better preserve map variability across volumes. Fig. 2 shows an example case with the spatial spread parameter equal to 0.25. The first column corresponds to the volume, and the rest the component maps. The number beneath each estimated map is the correlation coefficient against the ground truth. It can be seen that the proposed method obtains maps that are closer to the ground truth.

B. Evaluation with Real Data

1) Data Set and Data Augmentation: For real fMRI data, resting-state scans from the Center for Biomedical Research Excellence (COBRE) consisting of 74 healthy control (HC) subjects and 71 schizophrenic (SZ) subjects were used[1]. The raw scans were processed using the preprocessing pipeline in the Neuroimaging Analysis Kit (NIAK) [22], which included slice time correction, T1 normalization, BOLD-T1 co-registration, spatial resampling, artifact removal, and spatial smoothing, resulting in $V = 32 \times 32 \times 32$. The data set was split into 105 subjects for training, 10 validation, and 30 testing.

[1] https://fcon_1000.projects.nitrc.org/indi/retro/cobre.html

Since the real fMRI data did not have much temporal variability, one volume at around the middle time point was sampled from each subject's scan. The sampled volumes were collected in matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times V}$, to which an ICA algorithm was applied to extract spatial components $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times V}$ with $\tilde{\mathbf{X}} \approx \tilde{\mathbf{A}}\tilde{\mathbf{S}}$. Then, subject-specific residual spatial variation was computed as $\mathbf{X}_r = \tilde{\mathbf{X}} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$. Finally, the augmented volumes were generated as $\mathbf{X}_a = \mathbf{A}'\tilde{\mathbf{S}} + \mathbf{X}_r$, where the elements of $\mathbf{A}'$ were sampled from a uniform distribution with the mean and variance matched to those of $\tilde{\mathbf{A}}$. In this way, 700 additional volumes were generated for each subject in the training set.

2) Results: Since the ground truth component maps are not available for the real data set, the method is first assessed based on the similarity between the input and the reconstructed volumes. Fig. 3 shows a box plot of the correlation coefficients due to the proposed method. Also shown for comparison are the results from i) the ICA-EBM algorithm [23] based on the same data set without data augmentation; ii) the group ICA method without back-reconstruction; and iii) group ICA with back-reconstruction. For ii) and iii), the volumes over the entire time points were used as the input. The back-reconstruction procedure in iii) is a post-processing step added to ii), where the input data is regressed again on the estimated subject-specific time courses to obtain subject-specific spatial maps [2]. It can be seen that our method achieves higher correlation values than the methods that do not use back-reconstruction. Group ICA with back-reconstruction achieves very high correlations, but this is simply because back-reconstruction amounts to least-squares regression to match the input volumes. As we will see in Fig. 5, the step may not lead to very interpretable component maps. Fig. 4 depicts some example reconstructions together with the inputs sampled from the HC and the SZ groups. The numbers below the maps indicate the correlations toward the inputs. The proposed method is seen to faithfully capture the differentiating features between the groups, better than other ICA-based methods without back-reconstruction. Note that the group differences in the reconstructions for methods i) and ii) are due to the group differences in the coefficients $a_k[n]$, not the maps.

Fig. 5 shows $K = 8$ spatial activation maps from the proposed method for the same pair of subjects chosen for Fig. 4. The maps have been Z-scored and thus are independent of global scaling. Thus, only the maps from the proposed method and the group ICA with back-reconstruction can have group differences. However, the maps from group ICA with back-reconstruction look quite noisy and less interpretable than maps from other methods. It can also be observed that the maps from our method can capture component-wise group differences that are reflected in the input volume. For instance, The lower left slices in map $k = 7$ from our algorithm detect higher activation in the frontal area in the HC subject compared to the SZ, which is also visible in the input volumes in Fig. 4.

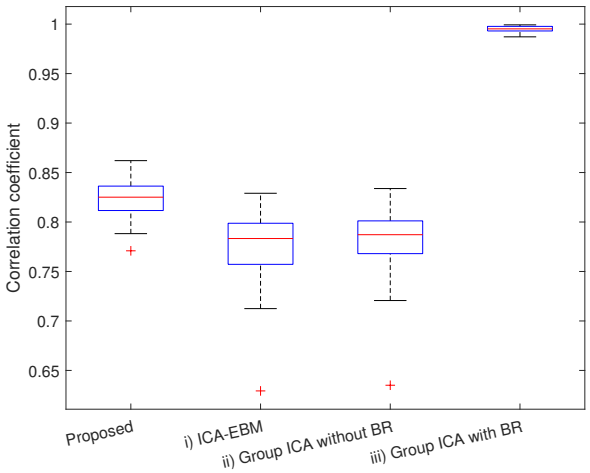

Fig. 3: Correlation coefficients between the input and the reconstructed volumes.

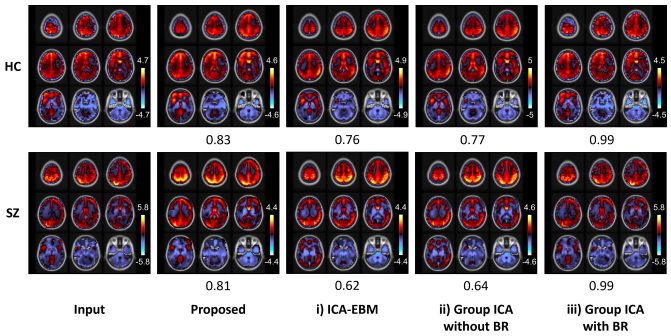

Fig. 4: Input and reconstructed volumes of HC/SZ subjects.

VI. Conclusion

A deep learning-based fMRI data analysis method has been proposed in an object-centric learning paradigm. By treating the spatial activation maps composing the fMRI volumes as "objects," the method can effectively capture the variabilities in the component maps across fMRI volumes. The proposed architecture includes an attention network that generates masks for the desired components, based on which VAEs can efficiently learn the latent representations of the maps. The numerical tests with synthetic data showed that the proposed method can faithfully estimate the components, and the real data results indicated that meaningful component maps that preserve volume-wise variabilities can be produced. Our future work will be on fully exploiting temporal dynamics in the input and applying the method for analysis of dynamic functional connectivity.

References

[1] S.-J. Kim, V. D. Calhoun, and T. Adalı, "Flexible large-scale fMRI analysis: A survey," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, Mar. 2017.

[2] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1, pp. S163–S172, Mar. 2009.

Fig. 5: Component maps obtained for the HC and SZ subjects.

[3] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, Pacific Grove, CA, Jun.–Jul. 2011, pp. 562–573.

[4] A. Iraji, R. Miller, T. Adali, and V. D. Calhoun, "Space: A missing piece of the dynamic puzzle," *Trends Cogn. Sci.*, vol. 24, no. 2, pp. 135–149, 2020.

[5] J.-H. Kim, Y. Zhang, K. Han, Z. Wen, M. Choi, and Z. Liu, "Representation learning of resting state fMRI with variational autoencoder," *NeuroImage*, vol. 241, Nov. 2021, Art. no. 118423.

[6] A. Farnoosh and S. Ostadabbas, "Deep markov factor analysis: Towards concurrent temporal and spatial analysis of fMRI data," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 17876–17888.

[7] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler, "Self-supervised transformers for fMRI representation," in *Proc. Int. Conf. Med. Imag. with Deep Learning*, Jul. 2022, vol. 172, pp. 895–913.

[8] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 11525–11538.

[9] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, "Multi-object representation learning with iterative variational inference," in *Proc. Int. Conf. Machine Learning.*, Long Beach, CA, Jun. 2019, pp. 2424–2433.

[10] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "MONet: Unsupervised scene decomposition and representation," *arXiv preprint arXiv:1901.11390*, 2019.

[11] M. Engelcke, A. R. Kosiorek, O. P. Jones, and I. Posner, "GENESIS: Generative scene inference and sampling with object-centric latent representations," in *Proc. of the International Conf. on Learning Representations*, Apr. 2020.

[12] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, Aug. 2012.

[13] K. Lee, S. Tak, and J. C. Ye, "A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *IEEE Trans. Med. Imaging*, vol. 30, no. 5, pp. 1076–1089, May 2011.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[15] J. Neri, R. Badeau, and P. Depalle, "Unsupervised blind source separation with variational auto-encoders," in *Proc. IEEE Eur. Signal Process. Conf.*, Aug. 2021, pp. 311–315.

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. of the International Conf. on Learning Representations*, Apr. 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Imag. Comput. Computer-Assisted Intervention*, Munich, Germany, Oct. 2015, Springer, pp. 234–241.

[18] N. Watters, L. Matthey, C. P. Burgess, and A. Lerchner, "Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs," in *ICLR Workshop on Learning from Limited Labeled Data*, New Orleans, LA, May 2019.

[19] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "How to train deep variational autoencoders and probabilistic ladder networks," in *Proc. Int. Conf. Machine Learning.*, New York City, NY, Jun. 2016.

[20] E. B. Erhardt, E. A. Allen, Y. Wei, T. Eichele, and V. D. Calhoun, "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *NeuroImage*, vol. 59, no. 4, pp. 4160–4167, Feb. 2012.

[21] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[22] P. Bellec, F. M. Carbonell, V. Perlbarg, C. Lepage, O. Lyttelton, V. Fonov, A. Janke, J. Tohka, and A. Evans, "A neuroimaging analysis kit for matlab and octave," in *Proc. Int. Conf. Functional Mapp. Hum. Brain*, Quebec City, Canada, Jun. 2011.

[23] X.-L. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5151–5164, 2010.