



Discovering Functional Brain Networks with 3D Residual Autoencoder (ResAE)

Qinglin Dong¹, Ning Qiang², Jinglei Lv³, Xiang Li^{1,5}, Tianming Liu⁴,
and Quanzheng Li^{1,5}(✉)

¹ Center for Advanced Medical Computing and Analysis, Department of Radiology,
Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
li.quanzheng@mgh.harvard.edu

² School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China

³ School of Biomedical Engineering and Sydney Imaging, Brain and Mind Centre,
The University of Sydney, Camperdown, Australia

⁴ Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and
Bioimaging Research Center, The University of Georgia, Athens, GA, USA

⁵ MGH & BWH Center for Clinical Data Science, Boston, MA, USA

Abstract. Functional MRI has attracted increasing attention in cognitive neuroscience and clinical mental health research. Towards understanding how brain give rises to mental phenomena, deep learning has been applied to functional MRI (fMRI) dataset to discover the physiological basis of cognitive process. Considering the unsupervised nature of fMRI due to the complex intrinsic brain activities, an encoder-decoder structure is promising to model hidden structure of latent signal sources. Inspired by the success of deep residual learning, we propose a 68-layer 3D residual autoencoder (3D ResAE) to model deep representations of fMRI in this paper. The proposed model is evaluated on the fMRI data under 3 cognitive tasks in Human Connectome Project (HCP). The experimental results have shown that the temporal representations learned by the encoder matches the task design and the spatial representations can be interpreted to be meaningful functional brain networks (FBNs), which not only include tasks based FBNs, but also intrinsic FBNs. The proposed model also outperforms a 3-layer autoencoder, showing the key factor for the performance improvement is depth. Our work demonstrates the feasibility and success of adopting 2D advanced deep residual networks in computer vision into 3D fMRI volume modeling.

Keywords: Deep learning · fMRI · Brain networks · 3D spatiotemporal model

1 Introduction

It has been decades since the neuroscience community started to research the neural connections that are involved in cognition process, aiming for a comprehensive understanding of brain functions. Functional MRI (fMRI) provides a powerful non-invasive

Q. Dong and N. Qiang—Equally contribution to this work.

© Springer Nature Switzerland AG 2020

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12267, pp. 498–507, 2020.

https://doi.org/10.1007/978-3-030-59728-3_49

tool to model cognitive behaviors of the whole brain and offers a useful information source to understand the intrinsic functional networks and the architecture of the human brain function [1–4]. In respect of modeling task-related brain function, growing evidence from fMRI (tfMRI) data [5–8] has revealed that these cognitive functions can be represented as a set of functional brain networks (FBNs), which are a collection of regions showing functional connectivity committed to different tasks [3, 4, 9]. Various computational algorithms/methods have been successfully exploited for tfMRI, such as independent component analysis (ICA) [10–13], general linear model (GLM) [9, 14] and sparse dictionary learning (SDL) [15–17]. Yet limited by their shallow nature, these existing machine learning models cannot extract fMRI intrinsic features in a hierarchical way. What’s more, GLM relies on prior knowledge of task designs, ICA has independence assumption and SDL has sparsity assumption.

Deep learning has attracted much attention in the fields of machine learning and data mining. It has been proven with great performance in multiple tasks that deep learning approach is superb at learning high-level and mid-level features from low-level raw data. [18–21] Considering the complexity of fMRI data and its intrinsic weak supervised nature, unsupervised deep models have gained great popularity in fMRI data modeling due to its superior representation power in learning latent features and association representations in a hierarchical way. Among those unsupervised deep models that have been applied to fMRI data analysis, the Deep Belief Nets (DBN) [22–24], Convolutional Autoencoder (CAE) [25–27] and Recurrent Autoencoder (RAE) [28, 29] have shown great promise in yielding a compact representation of brain activity. Recently, deep residual network has achieved significant performance improvement on natural image classification datasets with a substantially deeper structure than previous deep models [30, 31]. Despite that deeper neural networks are more difficult to train the gradient vanishing issue, they have greater representation powers, and the deep residual networks solve the issue using shortcuts between layers. Inspired by the success of deep residual networks, this paper exploited the possibility of learning representations of fMRI data with a very deep model. More specifically, a 68-layer residual autoencoder (ResAE) was designed to model the task-based fMRI in an unsupervised way.

In this paper, a group-wise scheme that aggregated multiple subjects’ fMRI data was designed for the effective training of ResAE models. The contribution of this work is three-fold. First, it presented an new approach to utilizing very deep models for learning meaningful FBNs from fMRI volumes. In addition, a comparison study with GLM showed that out proposed ResAE generates meaningful functional networks. Second, the enormous feature dimension challenge is tackled with convolution and pooling filters in the proposed method. Despite these recent investigations of the feature extraction and classification of MRI/fMRI data using deep networks, no study has explicitly employed whole-brain fMRI volume as an input and blindly extracted hidden features from the fMRI data. The curse of dimensionality is evident when the deep neural networks with tens of thousands of input nodes. Third, to address the inherent unsupervised nature of fMRI data, which comes with only coarse-grained labels or no labels at all, an autoencoder scheme is designed for fMRI analysis. Due to the unsupervised framework, many intrinsic FBNs were also found besides the task related FBNs, which implies the complexity of human brain activity.

2 Methods

The proposed computational framework is summarized in Fig. 1. In Sect. 2.1, fMRI data of all subjects are registered to a standard space and concatenated after preprocessing. In Sect. 2.2, the ResAE model consists of a pair of encoder and decoder which takes 3D fMRI volumes as input. The model is trained on a large-scale task fMRI dataset by reconstructing the input volumes. In Sect. 2.3, the feature representations of fMRI data were generated from the trained encoder with the latent nodes and were further visualized into interpretable FBNs.

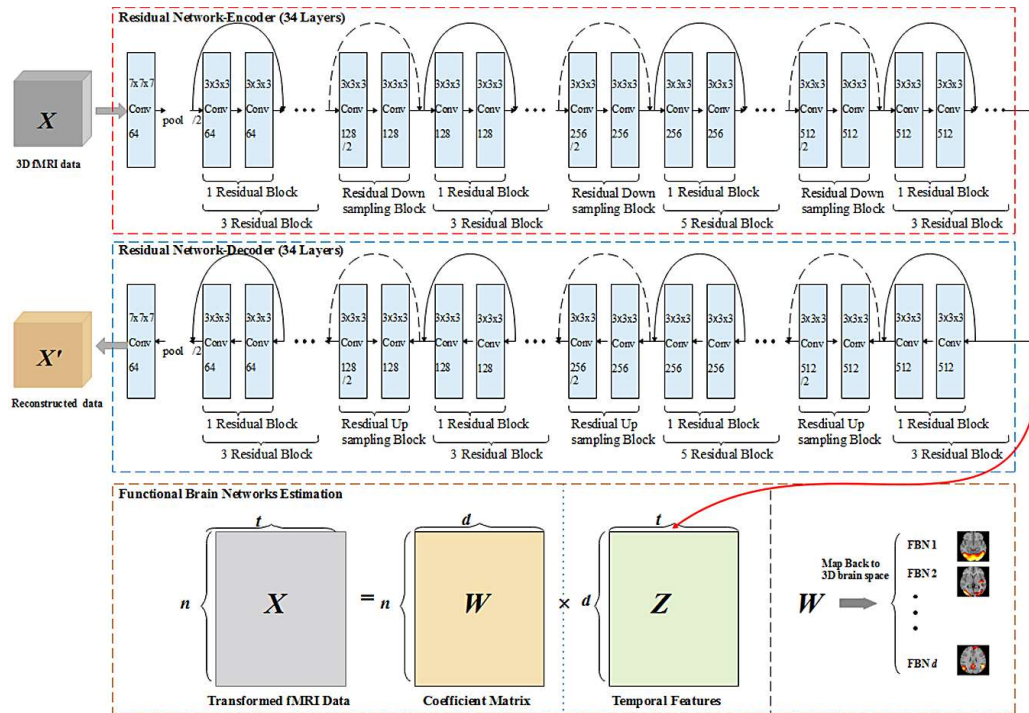


Fig. 1. Illustration of ResAE learning representations of fMRI data. The preprocessed 3D fMRI volumes are temporally concatenated as input. The proposed ResAE consists of a 34-layer encoder and a 34-layer decoder. Each residual block consists of two 3D convolution layers and an up/down pooling layer. With linear regression, the learned temporal features are used to build FBNs, which are further illustrated in Sect. 3.

2.1 Dataset and Preprocessing

The HCP task fMRI dataset is a systematic and comprehensive brain mapping collection of connectome-scale over a large population [6]. The primary goals of the HCP datasets were to identify as many core functional nodes in the brain as possible that can be correlated to structural and functional connectomes and behavior measurements. In the HCP Q3 public release, 900 subjects' fMRI datasets are available. In this paper, our experiments are based on three tasks: Emotion, Gambling and Social. Among these 900 subjects, 35 are excluded from our experiment for consistency of all tasks. The

acquisition parameters of tfMRI data are as follows: 90×104 matrix, 220 mm FOV, 72 slices, $TR = 0.72$ s, $TE = 33.1$ ms, flip angle = 52° , $BW = 2290$ Hz/Px, in-plane $FOV = 208 \times 180$ mm, 2.0 mm isotropic voxels. For tfMRI images, the preprocessing pipelines are implemented by FSL FEAT including skull removal, motion correction, slice time correction, spatial smoothing, global drift removal (high-pass filtering). All of these steps are implemented by FSL FEAT. [32, 33].

To perform group-wise ResAE training, all subjects' data were registered to the MNI152 $4 \times 4 \times 4$ mm³ standard template space, making sure that data from all subjects are in one same template space. The MNI152 template image is with $2 \times 2 \times 2$ mm³ spatial resolution originally and was down-sampled to $4 \times 4 \times 4$ mm³ before the registration. All volumes were variance normalized, concatenated along time dimension and shuffled. The size of the dataset is shown in Table 1. The dimension of the volumes is $49 \times 58 \times 47$ and was padded to $64 \times 64 \times 64$ for the convenience of the multiple down-sampling and up-sampling operations.

Table 1. Size of tfMRI data in HCP Q3

Task	Volumes	Duration (min)	Subjects	Samples
Emotion	176	2:16	865	152,240
Gambling	253	3:12	865	218,845
Social	274	3:27	865	237,010

2.2 Residual Module and ResAE

In this work, the deep residual module is adopted to address the notorious vanishing gradients problem in the training of deep neural networks. As shown in Fig. 2, given the input x of a layer, instead of fitting the regular mapping $H(x)$ of a layer, the residual module fits a residual mapping of $F(x) = H(x) - x$. Thus, the original mapping is transformed into an identity mapping $F(x) + x$ and is realized by shortcut connections of feedforward neural networks.

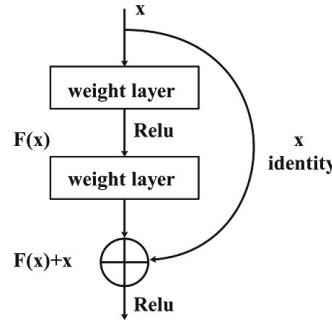


Fig. 2. Illustration of a residual block consisting of two weight layers.

Inspired by the success of deep residual nets on natural images, we propose an extension of deep residual networks for 3D neuroimage reconstruction. As illustrated in Fig. 1, the residual autoencoder network architecture creates a feature representation with its encoder. The encoder consists of 17 down-residual blocks and the decoder consists of 17 up-residual blocks. The down-residual block is composed of residual down sampling block following two 3D residual blocks, whereas the down-residual block is composed of residual down sampling block followed two 3D residual blocks. The solid side arrow in the standard residual block is a shortcut connection performing identity mapping. The dotted arrows in the up and down residual blocks are projection connections done using upsampling and max pooling, respectively. The up and the down residual blocks increase and decrease the output size, respectively.

To improve the gradient flow, the batch normalization (BN) [34] was applied to convolutional layer output before activation, which explicitly forces the activations to be unit gaussian distributed. All convolution filters are of size $3 \times 3 \times 3$. All rectified linear units (ReLU) in up-residual blocks had leak 0.3. We use Adam optimizer with a mini-batch size of 20 [35]. Mini batches take the advantage of GPU boards better and accelerate training with a proper size. However, if the batch size is too large, it may end up with less efficiency or even not converging, unless learning rate is decreased even larger. With a learning rate of 0.001, full cohort of data is trained with 100 epochs from scratch for full convergence in 20 h with a NVIDIA GTX 1080 GPU. The implementation can be found at <https://github.com/QinglinDong/ResAE>.

2.3 FBN Estimation

To explore the representation on the task fMRI data, we apply Lasso regression to estimate the coefficient matrix which is further used to build spatial maps. As shown in Fig. 1, the group-wise fMRI data \mathbf{X} is fed into the trained encoder, yielding the latent variables \mathbf{Z} from the output of encoder. Next, the FBNs \mathbf{W} are derived from latent variables and group-wise input via Lasso regression as follow:

$$\mathbf{W} = \min \|\mathbf{Z} - \mathbf{X}\mathbf{W}\|_2^2 + \lambda \|\mathbf{W}\|_1 \quad (1)$$

After Lasso regression, \mathbf{W} is regularized and transposed to a coefficient matrix, then each row of coefficient matrix is mapped back to the original 3D brain image space, which is the inverse operation of masking in data preprocessing. [36] Thus, the FBNs are generated and interpreted in a neuroanatomically meaningful context.

For comparison study, the GLM-based activation result was performed individually using FSL FEAT and group-wise averaged. Task designs were convoluted with the double gamma hemodynamic response function and set as the repressors of GLM. The contrast-based statistical parametric mapping was carried out with T-test and $p < 0.05$ (with cluster correction) is used to reject false positives. All the FBNs were thresholded at $Z > 2.3$ after transformation into “Z-scores” across spatial volumes.

3 Results

3.1 Temporal Features from ResAE

To further investigate the temporal feature corresponds to the FBNs, the ResAE-derived temporal features of a random subject are illustrated. As shown in Fig. 3, the ground truth is the Hemodynamic Response Function (HRF) Response to the task stimulus, indicating the brain activity corresponds the task design. The temporal features resulted from ResAE are compared with the HRF Response using Pearson correlation and it is shown that ResAE matches the task design well. To illustrate the effect of depth and residual model, a fully connected, 3-layer Autoencoder is adopted as comparison. The temporal features resulted from the 3-layer AE are also compared with the HRF Response and it shows an inferior match with task design compared to ResAE. For temporal features, the correlation between ResAE and HRF Response (average of 0.758) is greater than the correlation between AE and HRF Response (average of 0.404) at a significance level of 0.006%. It can be implied that ResAE has better capability to model the temporal information than the 3-layer AE.

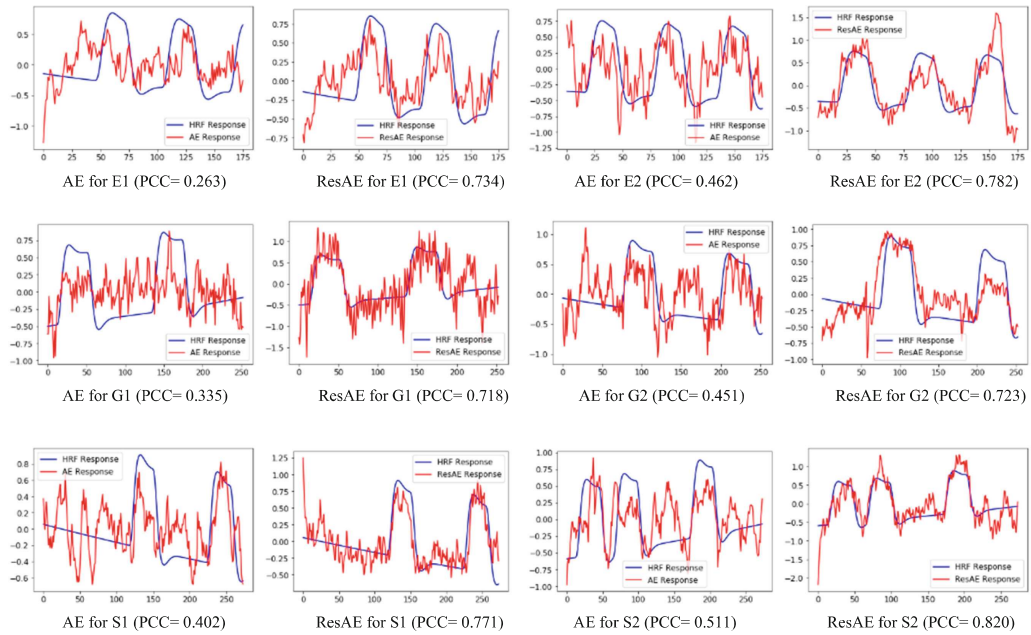


Fig. 3. Pearson correlation of temporal features based on AE and ResAE. The blue lines are the HRF responses, which are ground truth. (Color figure online)

3.2 Task-Related FBNs from ResAE

After the ResAE training, the temporal features are regressed and mapped back to the MNI152 space and superimposed onto the T1-weighted MRI image, so that the functional spatial maps are visualized and interpreted. For each node in the hidden layer, there is one functional network learned by ResAE. Due to space limit, some representative networks

that are task related are selected and visualized in Fig. 4. By visual inspection, these FBNs can be well interpreted, and they agree with domain knowledge of functional network atlases in the literature. To quantitatively evaluate the performance of ResAE in modeling tfMRI data, a comparison study between the proposed ResAE, a 3-layer AE and GLM (considered as benchmark) is also provided.

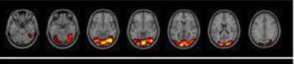
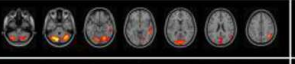
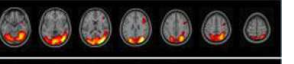
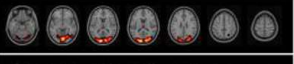
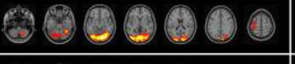
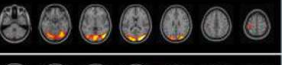
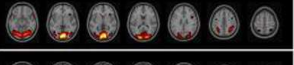
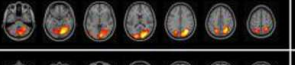
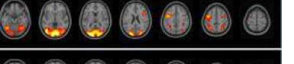
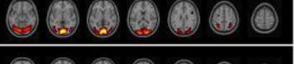
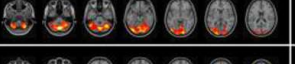
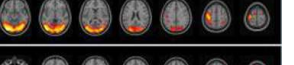
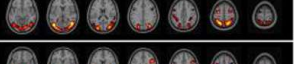
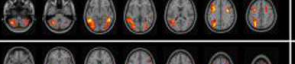
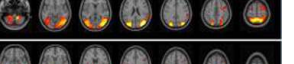

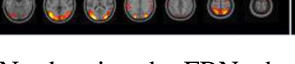

	GLM	AE	Res-AE
E1		IoU=0.334 	IoU=0.587 
E2		IoU=0.326 	IoU=0.635 
G1		IoU=0.274 	IoU=0.604 
G2		IoU=0.285 	IoU=0.579 
S1		IoU=0.403 	IoU=0.687 
S2		IoU=0.337 	IoU=0.712 

Fig. 4. Illustration of task-related FBNs showing the FBNs derived from GLM (benchmark), 3-layer AE, ResAE. For each of the tasks involved in this paper (Emotion, Gambling, Social), the original two explanatory variables (EVs) and corresponding results are shown.

To compare the FBNs derived by these three methods, the spatial overlap rate is defined to measure the similarity of two spatial maps. The spatial similarity is defined by the intersection over union rate (IoU) between two FBNs $N^{(1)}$ and $N^{(2)}$ as follows, where n is the volume size:

$$IoU(N^{(1)}, N^{(2)}) = \frac{\sum_{i=1}^n |N_i^{(1)} \cap N_i^{(2)}|}{\sum_{i=1}^n |N_i^{(1)} \cup N_i^{(2)}|} \quad (2)$$

With the similarity measure defined above, the similarities $IoU(N_{ResAE}, N_{GLM})$ and $IoU(N_{AE}, N_{GLM})$ are quantitatively measured. All the networks by ResAE have similar spatial distributions as the GLM derived networks, as shown by the quantitative similarities by the side of FBNs in Fig. 4. It is evident that the ResAE-derived network maps are very similar to the GLM derived network maps. This result demonstrated that ResAE can identify all GLM-derived networks, partly suggesting the effectiveness of the proposed model. Comparing this ResAE with the 3-layer AE, it is shown that ResAE has a better match with GLM than the 3-layer AE. For task related FBNs, the IoU between ResAE and GLM (average of 0.634) is greater than the IoU between AE and GLM (average of 0.326) at a significance level of 0.0001%. It is shown that with a deeper network, the proposed ResAE can model FBNs better than the 3-layer AE, suggesting the importance of the depth effect.

3.3 Intrinsic FBNs from ResAE

In our experiment results, it was also observed that the intrinsic FBNs, or resting state networks (RSNs), were continuously dynamically active even when subjects are doing task, which provides evidence supporting the conclusion in [2].

With the similarity measure defined above, the similarities $IoU(N_{ResAE}, N_{ICA})$ and $IoU(N_{AE}, N_{ICA})$ are quantitatively measured, where ICA is considered as benchmark for the intrinsic FBNs. Comparisons of pairs by these two methods are shown in Fig. 5, and the quantitative comparison are shown with the FBNs. RSN1, RSN2 and RSN3 correspond to visual network, RSN4 correspond to default mode network (DMN), RSN5 correspond to cerebellum, RSN6 correspond to sensorimotor network, RSN7 correspond to auditory network, RSN8 correspond to executive control network, RSN9 and RSN10 correspond to frontoparietal network. For intrinsic FBNs, the IoU between ResAE and ICA (average of 0.607) is greater than the IoU between AE and ICA (average of 0.365) at a significance level of 0.002%. It is shown that with a deeper network, the proposed ResAE can model FBNs better than the 3-layer AE, suggesting the importance of the depth effect, again. Major RSNs are all covered in the ResAE derived FBNs, which shows ResAE can discover intrinsic RSNs besides task related FBNs.

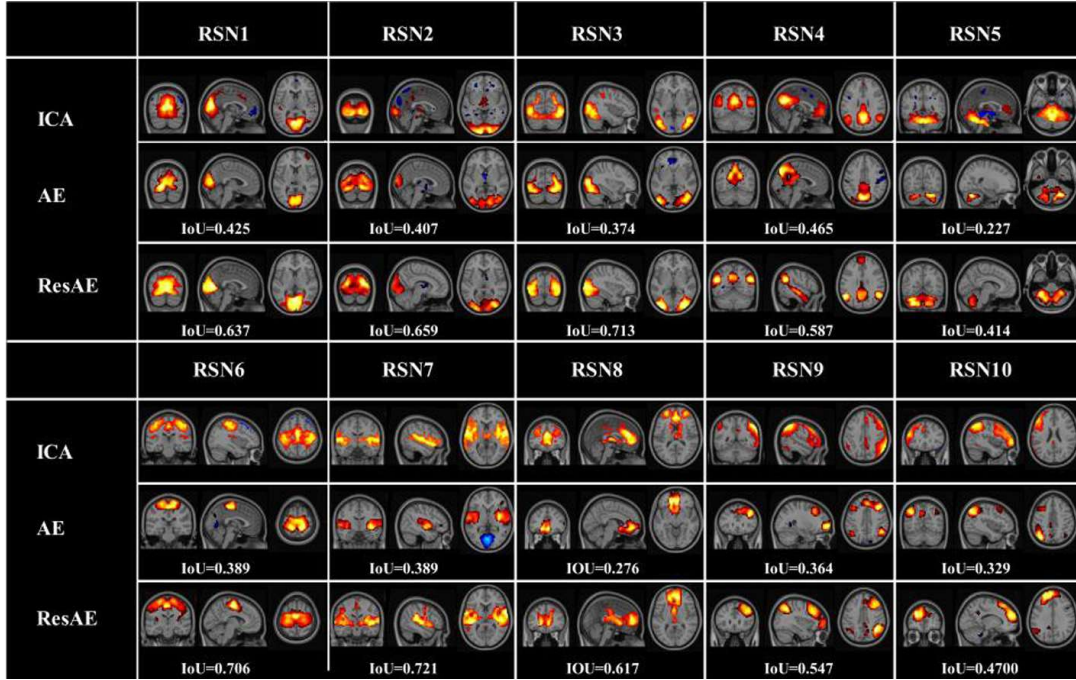


Fig. 5. Illustration of intrinsic FBNs showing the RSNs derived from ICA (benchmark), FBNs from 3-layer AE and ResAE.

4 Discussions

This paper is the first study that model fMRI networks with deep residual network to our best knowledge. In this paper, we proposed to adopt the encoder-decoder structure to exploit the deep residual network for this unsupervised task. With a group-wise experiment on massive tfMRI data, the ResAE model quantitatively and qualitatively showed its capability to learn FBNs. A comparison study with GLM, AE and ResAE shows that the FBNs learned by ResAE are meaningful and can be well interpreted.

One limitation of our current approach is that the effects of hyperparameters is not fully shown, including the model depth, which we plan to illustrate the model depth effects in further study. One promising future study is to apply the encoder representation and corresponding functional connectivity as biomarkers to brain disorder characterization such as Alzheimer's disease, ADHD, Autism, etc.

References

1. Huettel, S.A., et al.: Functional Magnetic Resonance Imaging, vol. 1. Sinauer Associates, Sunderland (2004)
2. Smith, S.M., et al.: Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci.* **106**(31), 13040–13045 (2009)
3. Pessoa, L.: Understanding brain networks and brain organization. *Phys. Life Rev.* **11**(3), 400–435 (2014)
4. Lv, J., et al.: Task fMRI data analysis based on supervised stochastic coordinate coding. *Med. Image Anal.* **38**, 1–16 (2017)
5. Archbold, K.H., et al.: Neural activation patterns during working memory tasks and OSA disease severity: preliminary findings. *J. Clin. Sleep Med.* **5**(01), 21–27 (2009)
6. Barch, D.M., et al.: Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169–189 (2013)
7. Binder, J.R., et al.: Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *Neuroimage* **54**(2), 1465–1475 (2011)
8. Dosenbach, N.U., et al.: A core system for the implementation of task sets. *Neuron* **50**(5), 799–812 (2006)
9. Kanwisher, N.: Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl. Acad. Sci.* **107**(25), 11163–11170 (2010)
10. McKeown, M.J.: Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage* **11**(1), 24–35 (2000)
11. Calhoun, V.D., et al.: A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* **14**(3), 140–151 (2001)
12. Beckmann, C.F., et al.: Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**(1457), 1001–1013 (2005)
13. Calhoun, V.D., et al.: Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* **5**, 60–73 (2012)
14. Beckmann, C.F., et al.: General multilevel linear modeling for group analysis in FMRI. *Neuroimage* **20**(2), 1052–1063 (2003)
15. Jiang, X., et al.: Sparse representation of HCP grayordinate data reveals novel functional architecture of cerebral cortex. *Hum. Brain Mapp.* **36**(12), 5301–5319 (2015)
16. Lv, J., et al.: Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Trans. Biomed. Eng.* **62**(4), 1120–1131 (2015)
17. Li, X., et al.: Multiple-demand system identification and characterization via sparse representations of fMRI data. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE (2016)
18. Bengio, Y.: Learning deep architectures for AI. *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009)
19. Bengio, Y., et al.: Deep learning. *Nature* **521**(7553), 436–444 (2015)

20. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
21. Yamins, D.L., et al.: Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**(3), 356 (2016)
22. Hjelm, R.D., et al.: Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. *NeuroImage* **96**, 245–260 (2014)
23. Jang, H., et al.: Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks. *NeuroImage* **145**, 314–328 (2017)
24. Dong, Q., et al.: Modeling hierarchical brain networks via volumetric sparse deep belief network (VS-DBN). *IEEE Trans. Biomed. Eng.* (2019)
25. Huang, H., et al.: Modeling task fMRI data via mixture of deep expert networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
26. Huang, H., et al.: Modeling task fMRI data via deep convolutional autoencoder. *IEEE Trans. Med. Imaging* **37**(7), 1551–1561 (2018)
27. Zhao, Y., et al.: 4D modeling of fMRI data via spatio-temporal convolutional neural networks (ST-CNN). *IEEE Trans. Cogn. Dev. Syst.* (2019)
28. Wang, H., et al.: Recognizing brain states using deep sparse recurrent neural network. *IEEE Trans. Med. Imaging* **38**, 1058–1068 (2018)
29. Li, Q., et al.: Simultaneous spatial-temporal decomposition of connectome-scale brain networks by deep sparse recurrent auto-encoders. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 579–591. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_45
30. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
31. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
32. Glasser, M.F., et al.: The minimal preprocessing pipelines for the human Connectome project. *Neuroimage* **80**, 105–124 (2013)
33. Jenkinson, M., et al.: Fsl. *Neuroimage* **62**(2), 782–790 (2012)
34. Ioffe, S., et al.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015)
35. Kingma, D.P., et al.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
36. Abraham, A., et al.: Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8**, 14 (2014)