FUNCTIONAL BRAIN NETWORK ANALYSIS

BASED ON UNSUPERVISED DEEP LEARNING

by

QINGLIN DONG

(Under the Direction of Tianming Liu)

ABSTRACT

In the neuroimaging and brain mapping communities, researchers have proposed a variety of computational methods and tools to learn functional brain network (FBN), such as general linear models (GLM), independent component analysis (ICA) and sparse dictionary learning (SDL). Recently, deep learning has attracted much attention in the fields of machine learning and data mining, and it has been proven that deep learning approach has superb representation power over traditional shallow models. In this research, three deep models, which are volumetric sparse deep belief networks (VS-DBN), neural architecture search based DBN (NAS-DBN) and recurrent autoencoder (RAE), were designed to explore representations of fMRI volumes. The quantitative analysis showed that these deep models have promising capability in learning meaningful FBNs and revealed novel insights into the organizational architecture of human brain.

INDEX WORDS: fMRI, functional brain networks, deep learning, deep belief networks, neural architecture search, recurrent autoencoder

FUNCTIONAL BRAIN NETWORK ANALYSIS

BASED ON UNSUPERVISED DEEP LEARNING

by

QINGLIN DONG

BS, Nankai University, China, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

FUNCTIONAL BRAIN NETWORK ANALYSIS

BASED ON UNSUPERVISED DEEP LEARNING

by

QINGLIN DONG

Major Professor:	Tianming Liu
Committee:	Hamid Arabnia
	Khaled Rasheed

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

DEDICATION

This dissertation is dedicated to my advisor Dr. Tianming Liu, for his tremendous help in my doctoral work. Without his kind guidance and support throughout my PhD, I cannot finish this dissertation.

ACKNOWLEDGEMENTS

T Liu was partially supported by National Institutes of Health (DA033393, AG042599) and National Science Foundation (IIS-1149260, CBET-1302089, BCS-1439051 and DBI-1564736).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Functional Brain Networks

Among the many imaging techniques available, non-invasive brain imaging techniques are showing great promises to reveal the intrinsic functional architecture of the brain. Functional magnetic resonance imaging (fMRI) records spontaneous fluctuations in the brain, yielding blood oxygenation level dependent (BOLD) based data. By revealing the synchronization of distant neural systems via correlations in neurophysiological measures of brain activity, functional brain networks (FBNs) have emerged as fundamental, organizational elements of human brain architecture.

Reconstruction and interpretation of FBNs from fMRI data, either resting state fMRI or task-based fMRI, has been under extensive active research (Bullmore & Sporns, 2009; Dosenbach et al., 2006; Duncan, 2010; Fox et al., 2005; Huettel, Song, & McCarthy, 2004; Kanwisher, 2010; Pessoa, 2012) in the past decade. In the neuroimaging and brain mapping communities, researchers have proposed a variety of computational methods and tools for brain network mapping, which can be generally categorized into model-driven or data-driven groups (Beckmann, DeLuca, Devlin, & Smith, 2005; Calhoun & Adali, 2012; Calhoun, Adali, Pearlson, & Pekar, 2001; Calhoun, Liu, & Adalı, 2009; B. Ge et al., 2016; Hu et al., 2015; Lv, Jiang, Li, Zhu, Chen, et al., 2015; Lv, Jiang, Li, Zhu, Zhang, et al., 2015; Lv, Jiang, Li, Zhu, Zhao, et al., 2015; Lv et al., 2017; McKeown, 2000; S. Zhang et al., 2016; W. Zhang et al., 2018; S. Zhao et al., 2015; Y.

Zhao et al., 2016). Among the data-driven methods, machine learning methodologies have played a central role in advancing both brain network reconstruction methods and their neuroscientific interpretations.

Independent component analysis (ICA) offers a methodology for exploring FBNs in human neuroimaging data (Beckmann et al., 2005; Calhoun & Adali, 2012; Calhoun et al., 2001; Calhoun et al., 2009; McKeown, 2000) . This exploratory approach identifies maximally statistically independent distributed spatial patterns depicting source processes. Evidence has shown that ICA can extract the intrinsic FBNs in both task fMRI and resting fMRI. Sparse dictionary learning (SDL) (B. Ge et al., 2016; Hu et al., 2015; Lv, Jiang, Li, Zhu, Chen, et al., 2015; Lv, Jiang, Li, Zhu, Zhang, et al., 2015; Lv, Jiang, Li, Zhu, Zhao, et al., 2015; Lv et al., 2017; S. Zhang et al., 2016; W. Zhang et al., 2018; S. Zhao et al., 2015; Y. Zhao et al., 2016) has also been widely used in many fMRI studies.

Recent studies (W. Zhang et al., 2018) comprehensively compared four variants of ICA methods and three variants of SDL methods using synthesized fMRI data with ground-truth. It was shown that ICA methods perform very well and slightly better than SDL methods when FBNs' spatial overlaps are minor, but ICA methods have difficulty in differentiating FBNs with moderate or significant spatial overlaps. In contrast, the SDL methods perform consistently well no matter how FBNs spatially overlap. SDL methods are significantly better than ICA methods when spatial overlaps between networks are moderate or severe. Despite the advantages of SDL methods over ICA methods, however, it is worth noting that SDL methods are still fundamentally limited as they are "shallow" models, meaning that they are not able to represent hierarchical organization of

FBNs, which is an intrinsic nature of the brain (Ferrarini et al., 2009; L. Lin, Osan, & Tsien, 2006; Meunier, Lambiotte, & Bullmore, 2010; Meunier, Lambiotte, Fornito, Ersche, & Bullmore, 2009; Zhou, Zemanová, Zamora, Hilgetag, & Kurths, 2006).

1.2 Review of Deep Learning

Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning. Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at orientations and locations in the image. The second layer typically detects motifs by spotting arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure. Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government.

Convolutional neural networks (CNN) are designed to process data that come in the form of multiple arrays, for example a colour image composed of three 2D arrays containing pixel intensities in the three colour channels.(Karpathy et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012; Lawrence, Giles, Tsoi, & Back, 1997; Simonyan & Zisserman, 2014) Many data modalities are in the form of multiple arrays: 1D for signals and sequences, including language; 2D for images or audio spectrograms; and 3D for video or volumetric images. There are four key ideas behind ConvNets that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers. The first few stages are composed of two types of layers: convolutional layers and pooling layers. Units in a convolutional layer are organized in feature maps, within which each unit is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. The result of this local weighted sum is then passed through a non-linearity such as a ReLU. All units in a feature map share the same filter bank. Different feature maps in a layer use different filter banks. The reason for this architecture is twofold. First, in array data such as images, local groups of values are often highly correlated, forming distinctive local motifs that are easily detected. Second, the local statistics of images and other signals are invariant to location. In other words, if a motif can appear in one part of the image, it could appear anywhere, hence the idea of units at different locations sharing the same weights and detecting the same pattern in different parts of the array. Mathematically, the filtering operation performed by a feature map is a discrete convolution, hence the name. Although the role of the convolutional layer is to detect local conjunctions of features from the previous layer, the role of the pooling layer is to merge semantically similar features into one. Because the relative positions of the features forming a motif can vary somewhat, reliably detecting the motif can be done by coarse-graining the position of each feature. A typical pooling unit computes the maximum of a local patch of units in one feature map (or in a few feature maps). Neighbouring pooling units take input from patches that are shifted by more than one row or column, thereby reducing the dimension of the representation and creating an invariance to small shifts and distortions. Two or three stages of convolution, non-linearity and pooling are stacked, followed by more convolutional and fully-connected layers. Backpropagating gradients through a ConvNet is as simple as through a regular deep network, allowing all the weights in all the filter banks to be trained. Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences.

Recurrent neural networks (RNN) process an input sequence one element at a time, maintaining in their hidden units a ‘state vector’ that implicitly contains information about the history of all the past elements of the sequence. (Han Wang; Sak, Senior, & Beaufays, 2014; Schuster & Paliwal, 1997)When we consider the outputs of the hidden units at different discrete time steps as if they were the outputs of different neurons in a deep multilayer network, it becomes clear how we can apply backpropagation to train RNNs. RNNs are very powerful dynamic systems but training them has proved to be problematic because the backpropagated gradients either grow or shrink at each time step, so over many time steps they typically explode or vanish. RNNs have been found to be very good at predicting the next character in the text or the next word in a sequence, but they can also be used for more complex tasks. For example, after reading an English sentence one word at a time, an English ‘encoder’ network can be trained so that the final state vector of its hidden units is a good representation of the thought expressed by the sentence. This thought vector can then be used as the initial hidden state of (or as extra input to) a jointly trained French ‘decoder’ network, which outputs a probability distribution for the first word of the French translation. If a first word is chosen from this distribution and provided as input to the decoder networks, it will then output a probability distribution for the second word of the translation and so on until a full stop is chosen. Overall, this process generates sequences of French words according to a probability distribution that depends on the English sentence. This rather naive way of performing machine translation has quickly become competitive with the state-of-the-art, and this raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules. It is more compatible with the view that everyday reasoning involves many simultaneous analogies that each contribute plausibility to a conclusion. Instead of translating the meaning of a French sentence into an English sentence, one can learn to ‘translate’ the meaning of an image into an English sentence. The encoder here is a deep CNN that converts the pixels into an activity vector in its last hidden layer. The decoder is an RNN like the ones used for machine translation and neural language modelling. There has been a surge of interest in such systems. RNNs, once unfolded in time, can be seen as very deep feedforward networks in which all the layers share the same weights. Although their main purpose is to learn long-term dependencies, theoretical and empirical evidence shows that it is difficult to learn to store information for very long.

1.3 Unsupervised Deep Learning on FMRI

Due to the weak supervision, in medical image analysis, the supervised deep models of CNN or RNN may not be directly applied to fMRI data. The lack of data is two-fold and more acute: there is general lack of publicly available data, and high-quality labelled data is even more scarce. The fMRI data are given with only coarse-grained labels or even no labels at all. What’s more, due to the complexity of human brain activity, many intrinsic FBNs could be activated at the same time and they cannot be labeled.

In the past several years, there have been growing bodies of literature (Cui et al., 2018; Han Wang; Hjelm et al., 2014; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018; Li, Huang, Chen, & Liu, 2018; Plis et al., 2014; Suk, Wee, Lee, & Shen, 2016; Y. Zhao et al., 2017; Y. Zhao, Ge, & Liu, 2018; Y. Zhao, Ge, Zhang, & Liu, 2018; Y. Zhao, Li, et al., 2018), including our own recent studies (Cui et al., 2018;

Han Wang; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018; Li et al., 2018; Y. Zhao et al., 2017; Y. Zhao, Ge, & Liu, 2018; Y. Zhao, Ge, Zhang, et al., 2018; Y. Zhao, Li, et al., 2018), that adopted deep learning models into fMRI data modeling and associated applications.

For instance, in (Hjelm et al., 2014), Hjelm et al. applied Restricted Boltzmann Machine (RBM) models (Fischer & Igel, 2012; Hinton, 2002; Yamashita, Tanaka, Yoshida, Yamauchi, & Fujiyoshii, 2014) to reconstruct FBNs from fMRI data and compared its performance to that of ICA methods. Based on synthetic and real task fMRI data, Hjelm et al. demonstrated that RBMs can be used to identify brain networks and their temporal activations with accuracy that is equal to or greater than that of ICA methods (Hjelm et al., 2014). Later, Hu et al. proposed to apply RBM models to fMRI time courses (Hu et al., 2018), instead of fMRI volume images. The proposed RBM method in (Hu et al., 2018) not only interprets fMRI time courses explicitly to take advantages of RBM in latent feature learning, but also substantially reduces model complexity and increases the scale of training set to improve model training. Their results based on Human Connectome Project (HCP) dataset demonstrated the superiority of the RBM method over ICA (Hu et al., 2018) in representing fMRI time series data. Moreover, the RBM method in (Hu et al., 2018) separated out components representing intermixed effects between task events, which could reflect inherent interactions among functionally connected brain regions. However, the RBM models in (Hu et al., 2018) and (Hjelm et al., 2014) for fMRI data were still shallow, that is, there were no deep structures of multiple layers of RBM layers used.

With deep learning method, hierarchical models for fMRI data can be constructed and high-level feature can be extracted which may contain more information than traditional methods. These computational deep methods can be generally classified into two categories: spatial approaches and temporal approaches.

Modeling fMRI with temporal features has already been explored in the literature (F. Ge et al., 2015; F. Ge et al., 2018; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018). For examples, Hu *et al. (Hu et al., 2018)* used restricted Boltzmann machine (RBM) to interpret fMRI temporal courses; Huang *et al. (Huang, Hu, Zhao, et al., 2018)* used deep convolutional auto-encoder (DCAE) to derive the features from task-based fMRI time series. Studies based on temporal approaches mostly focused on temporal features modeling while spatial information is overlooked.

Studies based on spatial approaches usually focused on the spatially decomposed components derived from fMRI data and typically ignored temporal dynamics information(Jiang et al., 2015; Lv, Jiang, Li, Zhu, Zhang, et al., 2015; Plis et al., 2014). For examples, Lv *et al.* (Lv, Jiang, Li, Zhu, Zhang, et al., 2015)used SDL to investigate the brain's spatial functional networks from fMRI data; Jiang *et al. (Jiang et al., 2015)* used the sparse representation to characterize the spatial functional regions with task-based fMRI data.

Due to the inter-subject variability is relatively more associated with the volatile time courses than with the spatial volumes in different imaging sessions, it appears that taking volumes as input possibly works better than time series in terms of modeling the FBNs for fMRI data in this case (Schmithorst & Holland, 2004). However, the spatial approach comes with two challenges despite the fact that it handle inter-subject variability better (Schmithorst & Holland, 2004). The first challenge is overfitting caused by data paucity. Considering the tremendous dimension of fMRI volumes, which can be more than 200K voxel per frame (based on MNI152 template (Fonov, Evans, McKinstry, Almli, & Collins, 2009)) and much more than a typical neuroimage dataset size, the overfitting can be serious. The second challenge is the lack of high-quality label. The rfMRI data is weak-supervised since the psychological label is coarse-grained and no frame-wise label is given, plus the complex co-activities of multiple ICNs.

In general, these previous studies focused on either spatial or temporal perspective of fMRI data and rarely modeled both domains simultaneously, thus, few of them has the ability to model the spatial-temporal variation patterns of FBNs. Therefore, a comprehensive and systematic framework is still in great need to recognize dynamic, temporal brain states at connectome-scale and model the brain's spatial-temporal dynamic activities simultaneously. However, development of such a comprehensive framework faces major challenges including the lack of ground truth of underlying neural activities and the inherent complexity associated with those spatial-temporal patterns of connectome-scale functional networks (Huang, Hu, Zhao, et al., 2018; Wang et al., 2018).

In a more recent study, Zhao et al. proposed a spatiotemporal convolutional neural network (ST-CNN) (Y. Zhao, Li, et al., 2018) to jointly learn the spatial and temporal patterns of targeted networks from training data and to perform automatic identification of functional networks in test data yet the temporal features were derived from the spatial features inherently.. The proposed ST-CNN is evaluated by the task of identifying the default mode network (DMN) from HCP fMRI data. Experimental results show that while the ST-CNN framework can capture the intrinsic relationship between the spatial and temporal characteristics of DMN and thus it ensures the accurate identification of DMN from independent datasets.

In another recent study (Cui et al., 2018), Cui et al. proposed a novel framework of Deep Recurrent Neural Network (DRNN) to model the FBNs from task fMRI data, and it was shown that the proposed DRNN can not only faithfully reconstruct FBNs, but also identify more meaningful brain networks with multiple time scales which are overlooked by traditional shallow models.

1.4 Dissertation Outline

Chapter 1 starts with an introduction of the concept of FBNs and the traditional approaches to learn the FBNs from fMRI data.

In Chapter 2, a novel volumetric sparse deep belief network (VS-DBN) model was designed and implemented through the popular TensorFlow open source platform to reconstruct hierarchical brain networks from volumetric fMRI. The experimental results showed that many interpretable and meaningful brain networks can be robustly reconstructed in a hierarchical fashion, and importantly, these brain networks exhibit reasonably good consistency and correspondence across multiple HCP task-based fMRI datasets.

In Chapter 3, It has been shown that deep neural networks are powerful and flexible models that can be applied on fMRI data with superb representation ability over traditional methods. However, a new challenge of neural network architecture design has also attracted attention: due to the high dimension of fMRI volume images, the manual process of network model design is very time-consuming and error prone. To tackle this problem, we proposed a Particle Swarm Optimization (PSO) based neural architecture search (NAS) framework for a deep belief network (DBN) that models volumetric fMRI data, named NAS-DBN. The core idea is that the particle swarm in our NAS framework can temporally evolve and finally converge to a feasible optimal solution. Experimental results showed that the proposed NAS-DBN framework can find robust architecture with minimal testing loss.

In Chapter 4, a novel deep sparse recurrent auto-encoder (DSRAE) was proposed in an unsupervised way to learn spatial patterns and temporal fluctuations of brain networks jointly. The proposed DSRAE were evaluated and validated on three tasks of the publicly available human connectome project (HCP) fMRI dataset with promising results. The proposed DSRAE is among the early efforts in developing unified models that can extract connectome-scale spatial-temporal networks from 4D fMRI data simultaneously.

In Chapter 5, the three models were summarized, and future work was discussed.

CHAPTER 2

MODELING HIERARCHICAL BRAIN NETWORKS VIA VOLUMETRIC SPARSE DEEP BLEIEF NETWORK

2.1 Overview

It has been recently shown that deep learning models such as convolutional neural networks (CNN), deep belief networks (DBN) and recurrent neural networks (RNN), exhibited remarkable ability in modeling and representing fMRI data for the understanding of functional activities and networks because of their superior data representation capability and wide availability of effective deep learning tools. For example, spatial and/or temporal patterns of functional brain activities embedded in fMRI data can be effectively characterized and modeled by a variety of CNN/DBN/RNN deep learning models as shown in recent studies. However, it has been rarely investigated whether it is possible to directly infer hierarchical brain networks from volumetric fMRI data using deep learning models such as DBN. The perceived difficulties of such studies include very large number of input variables, very large number of training parameters, the lack of effective software tools, the challenge of results interpretation, etc. To bridge these technical gaps, we designed a novel volumetric sparse deep belief network (VS-DBN) model and implemented it through the popular TensorFlow open source platform to reconstruct hierarchical brain networks from volumetric fMRI data based on the Human Connectome Project (HCP) 900 subjects release. Our experimental results showed that a large number of interpretable and meaningful brain networks can be robustly reconstructed from HCP 900 subjects in a hierarchical fashion, and importantly, these brain networks exhibit reasonably good consistency and correspondence across multiple HCP task-based fMRI datasets. Our work contributed a new general deep learning framework for inferring multiscale volumetric brain networks and offered novel insights into the hierarchical organization of functional brain architecture.

In this paper, a volumetric sparse Deep Belief Net (VS-DBN) was designed to explore the hierarchical organization of FBNs. As shown in Figure. 2.1, from tfMRI data of different tasks, the VS-DBN model learned features that can be interpreted as hierarchical task-specific FBNs.

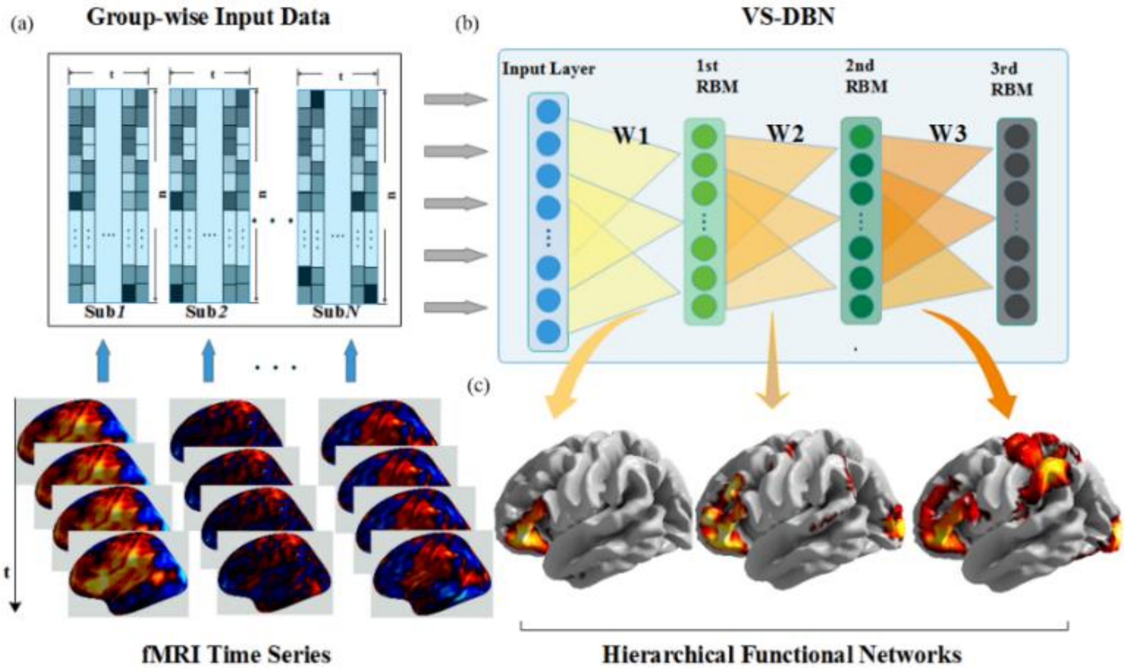


Figure. 2.1. Illustration of representing hierarchical structures of brain networks in tfMRI data by VS-DBN. (a) Preprocessed fMRI data was temporally concatenated in spatial fashion for input. Each fMRI volume data was treated as a single training sample for VS-DBN. (b) A VS-DBN with 3 layers was trained with volumes, and each layer has 100

hidden nodes. (c) The weights of each layer in the trained VS-DBN were considered as brain networks and visualized in the standard brain space.

2.2 Background

Recently, deep learning has attracted much attention in the fields of machine learning and data mining (Bengio, Courville, & Vincent, 2013; Karpathy et al., 2014; Krizhevsky et al., 2012; Lawrence et al., 1997; LeCun, Bottou, Bengio, & Haffner, 1998; Liu, Han, Liu, & Li, 2018; Liu, Han, Zhang, Wen, & Liu, 2015; Schmidhuber, 2015; Simonyan & Zisserman, 2014), and it has been proven that deep learning approach is superb at learning high-level and mid-level features from low-level raw data. A deep learning architecture usually consists of deep network layers by stacking multiple similar building blocks. The bottom layer receives input and then passes the transformed versions of the input to the next layer, all the way up to the top layer. As a result, the architecture of a deep learning model acts as a hierarchical feature extractor as a whole (Bengio, Goodfellow, & Courville, 2015; Li et al., 2018). In the past several years, there have been growing bodies of literature (Cui et al., 2018; Han Wang; Hjelm et al., 2014; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018; Li et al., 2018; Plis et al., 2014; Suk et al., 2016; Y. Zhao et al., 2017; Y. Zhao, Ge, & Liu, 2018; Y. Zhao, Ge, Zhang, et al., 2018; Y. Zhao, Li, et al., 2018), including our own recent studies (Cui et al., 2018; Han Wang; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018; Li et al., 2018; Y. Zhao et al., 2017; Y. Zhao, Ge, & Liu, 2018; Y. Zhao, Ge, Zhang, et al., 2018; Y. Zhao, Li, et al., 2018), that adopted deep learning models into fMRI data modeling and associated applications. For instance, in (Hjelm et al., 2014), Hjelm et al. applied Restricted Boltzmann Machine (RBM) models (Fischer & Igel, 2012; Hinton, 2002; Yamashita et al., 2014) to reconstruct FBNs from fMRI data and compared its performance to that of ICA methods. Based on synthetic and real task fMRI data, Hjelm et al. demonstrated that RBMs can be used to identify brain networks and their temporal activations with accuracy that is equal to or greater than that of ICA methods (Hjelm et al., 2014). Later, Hu et al. proposed to apply RBM models to fMRI time courses (Hu et al., 2018), instead of fMRI volume images. The proposed RBM method in (Hu et al., 2018) not only interprets fMRI time courses explicitly to take advantages of RBM in latent feature learning, but also substantially reduces model complexity and increases the scale of training set to improve model training. Their results based on Human Connectome Project (HCP) dataset demonstrated the superiority of the RBM method over ICA (Hu et al., 2018) in representing fMRI time series data. Moreover, the RBM method in (Hu et al., 2018) separated out components representing intermixed effects between task events, which could reflect inherent interactions among functionally connected brain regions. However, the RBM models in (Hu et al., 2018) and (Hjelm et al., 2014) for fMRI data were still shallow, that is, there were no deep structures of multiple layers of RBM layers used. Therefore, the hierarchical organization of functional brain activities and networks cannot be explored yet. In a more recent study, Zhao et al. proposed a spatiotemporal convolutional neural network (ST-CNN) (Y. Zhao, Li, et al., 2018) to jointly learn the spatial and temporal patterns of targeted networks from training data and to perform automatic identification of functional networks in test data. The proposed ST-CNN is evaluated by the task of identifying the default mode network (DMN) from HCP fMRI data. Experimental results show that while the ST-CNN framework can capture the intrinsic relationship between the spatial and temporal characteristics of DMN and thus it ensures the accurate identification of DMN from independent datasets. In another recent study (Cui et al., 2018), Cui et al. proposed a novel framework of Deep Recurrent Neural Network (DRNN) to model the FBNs from task fMRI data, and it was shown that the proposed DRNN can not only faithfully reconstruct FBNs, but also identify more meaningful brain networks with multiple time scales which are overlooked by traditional shallow models.

Prior studies of using deep learning models for fMRI data analysis, such as CNN/RBM/RNN (Cui et al., 2018; Han Wang; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018; Li et al., 2018; Y. Zhao et al., 2017; Y. Zhao, Ge, & Liu, 2018; Y. Zhao, Ge, Zhang, et al., 2018; Y. Zhao, Li, et al., 2018), have exhibited great promises, however, it has been rarely examined whether/how to infer and reconstruct hierarchical brain networks from volumetric fMRI data directly using deep learning models such as deep belief networks (DBN). A major advantage of using DBN for fMRI data modeling is that DBN can naturally represent the hierarchical patterns of FBNs in an unsupervised manner (Bengio, 2009; Hinton, Osindero, & Teh, 2006). Theoretically, the unsupervised learning via DBN has the solid interpretability based on maximizing likelihood estimation rather than minimizing the reconstruction error, which makes DBN attractive (Bengio, 2009; Hinton et al., 2006). However, the perceived difficulties of developing DBN models for fMRI data include very large number of input variables (e.g., the hundreds of thousands of volumetric image intensities), very large number of training parameters (e.g., millions of DBN weights), the lack of effective software tools (e.g., there is no TensorFlow (Abadi et al., 2015) implementation of DBN), the challenge of results interpretation (e.g., many volumetric brain network maps in multiple layers) and etc. To bridge these technical and knowledge gaps, in this paper, we designed a novel volumetric sparse deep belief network (VS-DBN) model, implemented it based on the popular TensorFlow open source platform, and applied it on the Human Connectome Project (HCP) 900 subjects release. Our extensive experimental results have shown that many meaningful FBNs can be robustly reconstructed from HCP 900 subjects in a hierarchical fashion, and importantly, these reconstructed networks can be well interpreted based on current neuroscience knowledge. Interestingly, these reconstructed brain networks by DBN exhibit reasonably good consistency and correspondence across multiple HCP task-based fMRI (tfMRI) datasets, suggesting a possibly common functional organization architecture of the brain. In general, our works contributed a general DBN deep learning framework for inferring volumetric brain networks and offered new insights into the hierarchical functional organization architecture of the brain. The source codes and models in this paper will be released at: https://github.com/QinglinDong/vsDBN.

2.3 Dataset and Pre-processing

In this paper, we used the Human Connectome Project (HCP) 900 Subjects MR imaging data from Q3 Release as training dataset and all the data is available on https://db.human connectome.org. (Fischer & Igel, 2012). The HCP dataset is a systematic and comprehensive connectome-scale collection over 900 healthy young adults, aging 22-35. In all parts of the dataset, participants were scanned on the same equipment using the same protocol for each subject and the detailed acquisition parameters are shown in TABLE I. The number of brain voxels is either: 228,453 (MNI-152 space of 2 mm spacing), 28,549 (MNI-152 space of 4 mm spacing), or 91,282 (standard co-ordinate system of cortical surface vertices and subcortical voxels). In this paper, our experiments were based on the MNI-152 space of 2 mm spacing due to its superior spatial resolution.

TABLE 2.1 Imaging Protocol of HCP Q3 TFMRI Dataset

Parameter	VALUE	Parameter	Value
Sequence	Gradient-echo EPI	Matrix	104x90
TR	720 ms	Slice thickness	2.0 mm
TE	33.1 ms	Multiband factor	8
flip angle	52 deg	Echo spacing	0.58 ms
FOV	208x180 mm	BW	2290 Hz/Px

Stimuli were projected onto a computer screen behind the subject’s head within the imaging chamber (Barch et al., 2013). The screen was viewed by a mirror positioned approximately 8 cm above the subject’s face. Seven categories of behavioral tasks were involved.

TABLE 2.2 Size of HCP Q3 TFMRI Dataset

Task	VOLUMES	Duration (Min)	Samples
Emotion	176	2:16	152,240
Gambling	253	3:12	218,845
Motor	284	3:34	245,660
Language	316	3:57	273,340
Relational	232	2:56	200,680
Social	274	3:27	237,010
Working Memory	405	5:01	350,325

Before the tfMRI data was modeled with deep learning, standard preprocessing was applied. For consistency and fair comparison, 35 subjects that did not perform all 7 tasks were excluded thus 865 out of 900 subjects’ data was used in this work. For tfMRI images, the preprocessing pipelines included skull removal, motion correction, slice time correction, spatial smoothing etc. These steps are implemented by FSL FEAT (FMRIB's Expert Analysis Tool).

Spatial Resampling

For all 865 subjects in the dataset, all volumes were registered to a same MNI-152(Montreal Neurological Institute) T1-weighted standard template space, hence on a common affine.

Frequency Filtering

A band filter was applied to remove high or low frequency signals.

Detrending

The global drift over the time series was removed. The absolute voxel intensity may be ignored since we are modeling the fMRI volumes.

Normalization

A spatial normalization was applied to all volumes such that they have zero mean and unit variance, yielding a Gaussian distribution. It helps the DBN training to converge better.

Masking

With the MNI152 mask, the backgrounds of all subjects were removed and the original 4D block of fMRI data (three spatial dimensions and a time dimension) was transformed into a 2D array (voxel dimension and time dimension).

The preprocessed dataset consisted of 1,678,100 volumes as training samples in total and the information of each task is shown in TABLE II. The volumes of all subjects in each task were temporally concatenated for a task-specific group-wise volumetric learning scheme and the details are explained in Section E.

2.4 RBMs and DBNs

Restricted Boltzmann Machines (RBMs) are generative models that approximates a closed-form representation of the underlying probability distribution of the training data. RBMs can also be interpreted as deterministic feed-forward neural networks and they are widely used as the building blocks of Deep Belief Nets (DBNs). As shown in Figure.2, RBMs can be viewed as undirected probability graphical models, i.e. Markov Random Fields which are complementary to the directed models, i.e. Bayesian Networks. (Fischer & Igel, 2012)

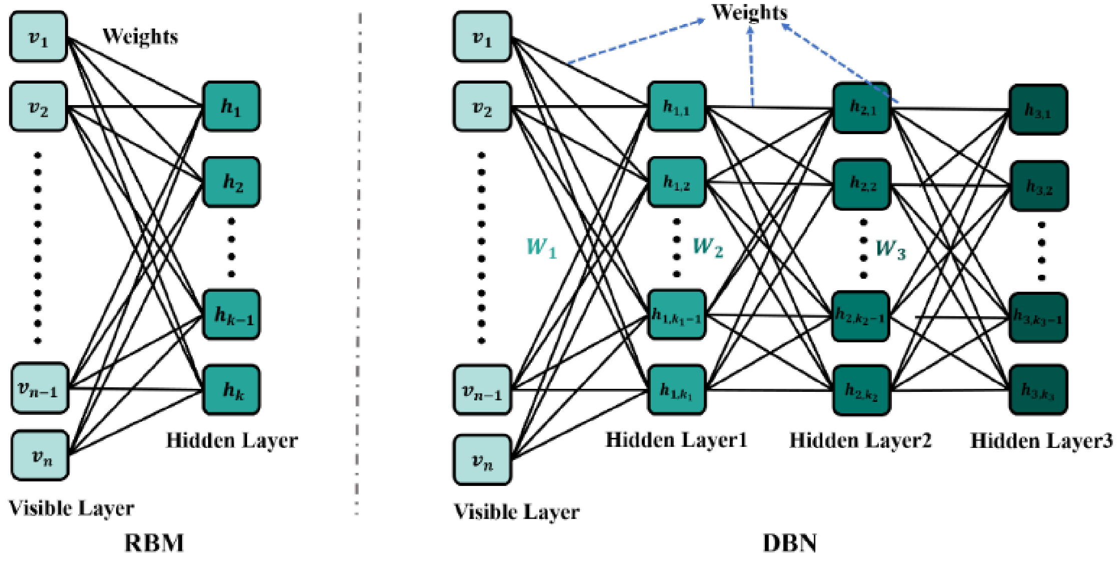


Figure. 2.2 The RBM and DBN structures.

RBMs consist of an input layer of visible variables $v_i \in R$ and a hidden layer of latent variables $h_i \in R$. Given the observed data, RBMs can model the dependencies of a set of visible variables v_i and a set of hidden variables h_i over the set of observed data. For each pair of a visible node v_i and a hidden node h_i, the connection models the joint probability distribution as follows (Bengio et al., 2015; Fischer & Igel, 2012; Hinton et al., 2006):

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{1}$$

where Z is a normalization term and E (v, h) is an energy function defined by the interaction of visible variables and hidden variables. Given a set of observations $\boldsymbol{v}$, the model parameter $\boldsymbol{\theta}$ is optimized when the likelihood of the probability of training data $P(\boldsymbol{v})$ is maximized which is the same as maximizing the log-likelihood given by:

$$\begin{aligned} \ln P(\boldsymbol{v}|\boldsymbol{\theta}) &= \ln \frac{1}{Z} \sum_{h} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \\ &= \ln \sum_{h} e^{-E(\boldsymbol{v}, \boldsymbol{h})} - \ln \sum_{v,h} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \end{aligned} \tag{2}$$

However, computing the likelihood of the undirected models and their gradients for inference in generally computationally intensive, thus RBMs restrict the interactions only those between visible variables and hidden variables to yield simple conditional probabilities:

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{m} P(v_i|\boldsymbol{h}) \tag{3}$$

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{n} P\left(h_j|\boldsymbol{v}\right) \tag{4}$$

To calculate the conditional probabilities in (3) an (4), the energy function is defined for different distribution of training data. For binary images where the RBM was first applied, both visible nodes and hidden nodes are binary variables in Bernoulli distribution. In the context of fMRI data, the activation of each voxel being real-valued and in Gaussian distribution. Thus, in this paper, Gaussian-Bernoulli E (**v**, **h**) is adopted as energy model (Cho, Raiko, & Ilin, 2013). Note the interactions between v_i and h_i as W_{ij}, the visible bias as a_i, the hidden bias as b_j, the standard deviations of visible nodes σ_i, the energy function is defined as:

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i} \frac{(a_i - v_i)^2}{\sigma_i^2} - \sum\nolimits_{j} b_j h_j - \sum_{ij} \frac{v_i}{\sigma_i} W_{ij} h_j \tag{5}$$

With the energy defined in (5), the conditional probabilities in (3) and (4) can be computed as follows, which can also be interpreted as the firing rate of a stochastic neuron with sigmoid activation function.

$$P(h_i \ = 1|\boldsymbol{v}) = sigmoid\left(\sum\nolimits_{j} W_{ij} v_i + b_j\right) \tag{6}$$

$$P(v_i \ = 1|\boldsymbol{h}) = sigmoid\left(\sum\nolimits_{i} W_{ij} h_i + a_j\right) \tag{7}$$

To update the model, the gradient of log-likelihood in (2) is estimated using Gibbs Sampling methods as a Markov Chain Monte Carlo (MCMC) Technique, where the angle brackets denote the expectation with respect to the specified distribution (Fischer & Igel, 2012).

$$\frac{\partial \ln\ (\boldsymbol{v}|\boldsymbol{\theta})}{\partial W_{ij}} \simeq \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{8}$$

$$\frac{\partial \ln\ (\boldsymbol{v}|\boldsymbol{\theta})}{\partial a_i} \simeq \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \tag{9}$$

$$\frac{\partial \ln\ (\boldsymbol{v}|\boldsymbol{\theta})}{\partial b_j} \simeq \langle h_j \rangle_{data} - \langle h_j \rangle_{model} \tag{10}$$

However, the $\langle v_i h_j \rangle_{model}$ is still intractable due to computation burden of Gibbs Sampling where summing over all values of the visible variables has exponential complexity. The contrastive divergence (CD) algorithm is used to approximate the gradient (as shown in TABLE III):

TABLE 2.3 ALGORITHM: Contrastive divergence

```
Input: RBM, training batch X∈ $\mathbb{R}^{t\times n}$
Output: parameter approximation $W_{ij}$, $a_i$, $b_j$, for $i$=1, …, $n$, $j$=1, …, $m$

Initialize $W_{ij}$, $a_i$, $b_j$ to ℕ (0, 1) for $i$=1, …, $n$, $j$=1, …, $m$
for all $\boldsymbol{v}$ in $\boldsymbol{X}$ do
    sample $h_j$~ $p\ (h_j|\boldsymbol{v})$
    sample $v_j$~ $p\ (v_i|\boldsymbol{h})$
    for $i$=1, …, $n$, $j$=1, …, $m$ do
        $W_{ij} \leftarrow W_{ij} + \alpha\big(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon} - \beta \cdot sign(W_{ij})\big)$
        $a_i \leftarrow W_{ij} + \alpha(\langle v_i \rangle_{data} - \langle v_i \rangle_{recon})$
        $b_j \leftarrow W_{ij} + \alpha(\langle h_j \rangle_{data} - \langle h_j \rangle_{recon})$
```

As shown in Figure. 1 and Figure. 2, with 3 RBMs as building blocks, the stacked model forms a DBN, yielding a higher level of features. As similar to an RBM that can be trained by the CD algorithm successfully (TABLE III), a DBN can be trained with the same method in a layer-wise manner. (Hinton et al., 2006)

2.5 VS-DBN for fMRI

Modeling fMRI with temporal features has already been explored in the literature (F. Ge et al., 2015; F. Ge et al., 2018; Hu et al., 2018; Huang, Hu, Dong, et al., 2018; Huang, Hu, Zhao, et al., 2018). However, since the inter-subject variability is relatively more associated with the volatile time courses than with the spatial volumes in different imaging sessions, it appears that taking volumes as input possibly works better than time series in terms of modeling the FBNs for fMRI data in this case (Schmithorst & Holland, 2004). In this paper, a volumetric learning scheme was applied where a volume from the fMRI data was taken as a feature and each time frame was taken as a sample. Thus, after preprocessing, the volumes of all subjects were concatenated along time dimension and shuffled for further group-wise training. As shown in Figure.1(a), each fMRI volume at a time point was used as a training sample and the DBN was trained in an unsupervised fashion DBN.

Moreover, to reduce overfitting and improve generalization, weight regularization was adopted in the proposed model. In each iteration, the weight updates with the estimated derivative and an extra term of weight regularization derivative. In this paper, L1 weight penalty was adopted as the regulation term and it calculated the derivative of the sum of the absolute values of the weights (Bengio et al., 2015; Fischer & Igel, 2012; Hinton et al., 2006). With a weight decay rate β, the overall weight update was:

$$\frac{\partial \ln\ (\boldsymbol{v}|\boldsymbol{\theta})}{\partial W_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon} - \beta \cdot sign(W_{ij}) \tag{11}$$

The sparse weight regularization works by causing most of the weights to become zero while allowing a few of the weights to grow large. As shown in Figure.1(b), a DBN model consisting of 3 RBM with weight regularization was built in a spatial fashion, by taking tfMRI volumes as input. In the context of fMRI data, L1 regularization can denoise the FBNs and improve interpretability by suppressing useless weights and allowing important model parameters to become larger, which is considered as an important methodology contribution of this work and it will be demonstrated in the following results sections.

With respect to interpreting a trained VS-DBN in an fMRI context, each row of weight vector was mapped back into the original 3D brain image space, which was the inverse operation of masking in preprocessing steps and was interpreted as an FBN. As shown in Figure.1(c), after the DBN was trained layer-wisely on a large-scale tfMRI dataset, the trained weights modeled a feature representation with the latent variables in the hidden nodes, thus yielding interpretable FBNs. Each weight showed the dependency of each voxel with a latent variable. For deeper layers, the linear combination approach was used to interpret the connection (Lee, Grosse, Ranganath, & Ng, 2009). With this approach, W_1 was visualized for the first hidden layer as FBNs, $W_2 \times W_1$ for the second layer and $W_3 \times W_2 \times W_1$ for the third layer, respectively. To analyze the internal representation of each layer, the connection to a hidden node will be visualized, which will be illustrated later in Figure. 3(a-c) and Figure. 4(a-c).

Here, a group-wise scheme of VS-DBN was proposed to model fMRI data. Considering the large inter-subject variability among human brains, arbitrary selection of a single individual may not effectively represent the population, thus a group-wise learning scheme was needed to reduce inter-subject variability by jointly registering the volumes to a common reference template corresponding to the group average. Due to the wide individual variation of human brain, the arbitrary selection of a single individual may not truly represent the population. In this paper, a group-wise learning scheme was used to reduce biases by jointly registering the volumes to a common reference corresponding to the group average.

Besides the massive group-wise population size, the VS-DBN was inherently much more computationally expensive despite the merit of the volumetric deep learning model, compared to a temporal fMRI time series model. In the context of fMRI data, the spatial dimension is much higher than the temporal dimension in most cases due to the high cost of single fMRI scan, yielding a much larger trainable parameter space in the model correspondingly. Consider HCP 3D images and one single layer of RBM, there are around 20K trainable parameters for temporal approach is, 20M for volumetric. Moreover, the seriously large population of data will put significantly computational burden on the model training.

To solve this problem, in this paper, TensorFlow (Abadi et al., 2015), which is a popular deep learning framework and provides great convenience coding with GPUs, was adopted with high efficiency GPU computation to fill the gap. Based on TensorFlow, we designed and implemented a fast and flexible DBN to realize the proposed VS-DBN specifically and the source code will be released at https://github.com/QinglinDong/vsDBN. Below are the details of implementation:

Nodes

According to the dataset, the visible layer of DBN model was constructed with 228,453 nodes. Hyperbolic tangent function was chosen as the activation function instead of sigmoid since the fractional nature of the fMRI data.

Initialization

To start with training, the weights and biases are initialized from a Gaussian with zero-mean and a standard deviation of 0.01.

Batch Normalization

To improve the convergence, batch normalization technique was applied to each hidden layer, which explicitly forced the activations to be unit Gaussian distributed.

Converging

With a learning rate of 0.01, batch size of 20 and weight-decay rate of 0.1, the models were trained for 20 epochs and the hyperparameters were selected by grid search to achieve a good convergence and interpretability.

Mni-batches

The volumes were divided into mini batches with a size of 5. Mini batches take the advantage of GPU boards better and accelerate tCraining with a proper size. However, if the batch size was too large, it may end up with less efficiency or even not converging, unless learning rate was decreased even larger.

Progressively Loading

Due to the very large size of fMRI dataset, a progressively loading algorithm was used to tackle the issue that big data won't fit in memory (Ross, Lim, Lin, & Yang, 2008).

Reproducibility

All experiments were repeated 5 times to test the stability of consistency of results.

2.6 Analysis of DBN-FBNs

To explore the representation on task fMRI data, seven task-specific VS-DBNs were trained on fMRI data of 7 HCP tasks independently using the same hyperparameters. For each task-specific VS-DBN, three hidden layers were constructed, each with 100 nodes (empirically set). To fairly compare the FBNs of different layers, the numbers of hidden nodes were set to be the same intentionally. To achieve fair comparison across FBNs from different task datasets and different layers, all maps from different layers were normalized and equally thresholded. Each layer of a task-specific VS-DBN model acquired 100 FBNs and two randomly chosen task of emotion and gambling is illustrated.

The FBNs derived from the emotion task is shown in Figure. 3 (a-c). In the emotion task, the participants are presented with blocks of trials that ask them to decide either which of two faces presented on the bottom of the screen match the face at the top of the screen, or which of two shapes presented at the bottom of the screen match the shape at the top of the screen. The faces have either angry or fearful expressions (Barch et al., 2013). In layer 1, there are visual (no. 30 in Figure. 3(a)), frontal (no. 56 in Figure. 3(a)), motor (no. 99 in Figure. 3(a)), subcortical area (no. 42 in Figure. 3(a)), frontopariatal network (no. 64 in Figure. 3(a)) and sensorimotor (no. 20 in Figure. 3(a)), etc. In layer 2, there are default mode network (no. 71 in Figure. 3(b)), executive control network (no. 38 in Figure. 3(b)) and frontopariatal network (no. 1, 45 in Figure. 3(b)), etc. In layer 3, there are default mode network (no. 35, 64 in Figure. 3(c)) and frontopariatal network (no. 29 in Figure. 3(c)), etc.

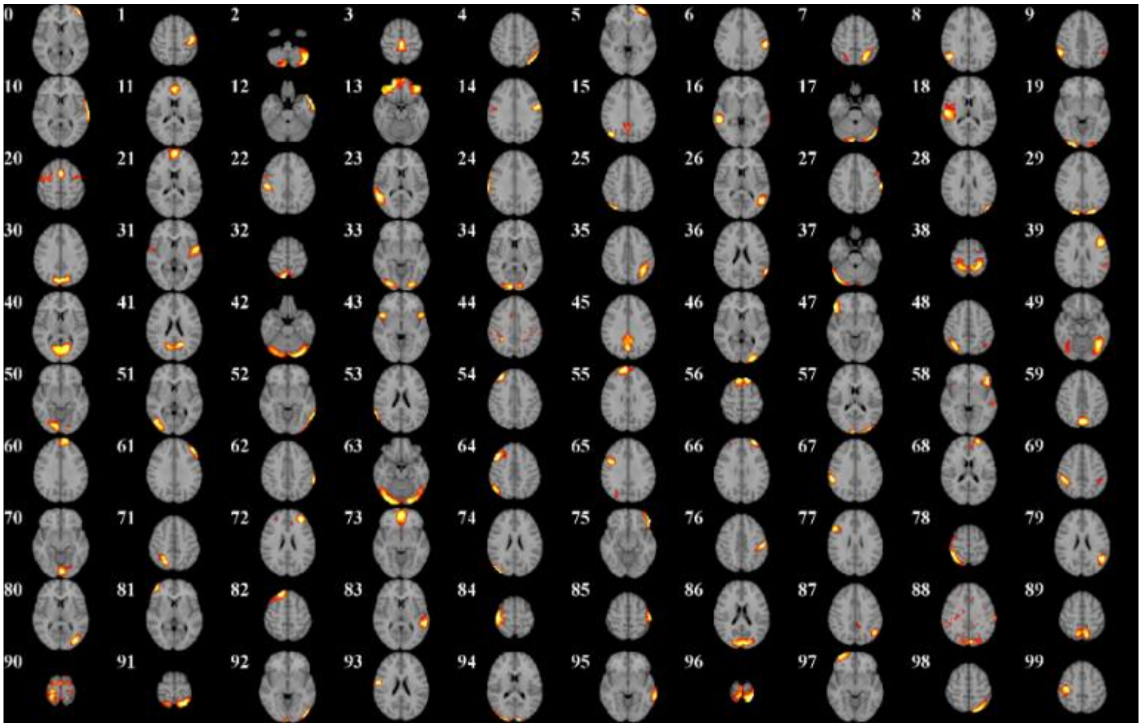


Figure. 2.3(a). Visualization of 100 FBNs learned in layer 1 of the VS-DBN from emotion task. Each network is visualized with the most informative axial slice. The index of each of these 100 networks is on the top left.

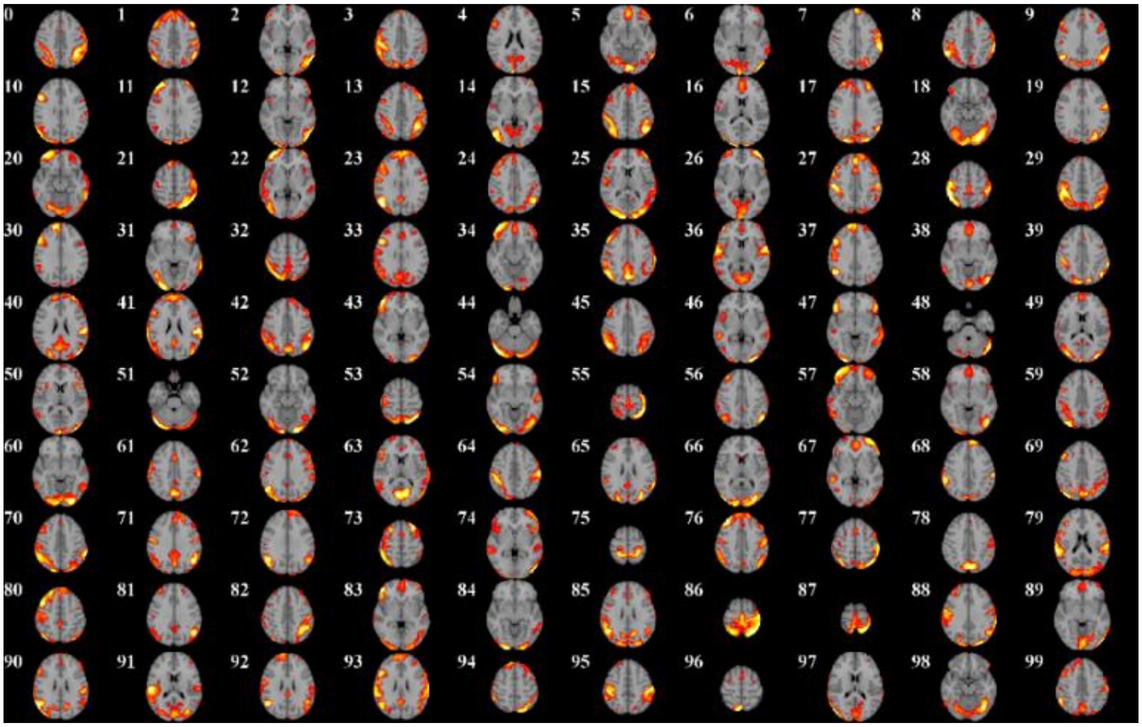


Figure. 2.3(b). Visualization of 100 FBNs learned in layer 2 of the VS-DBN from emotion task. Each network is visualized with one most informative axial slice. The index of each of these 100 networks is on the top left.

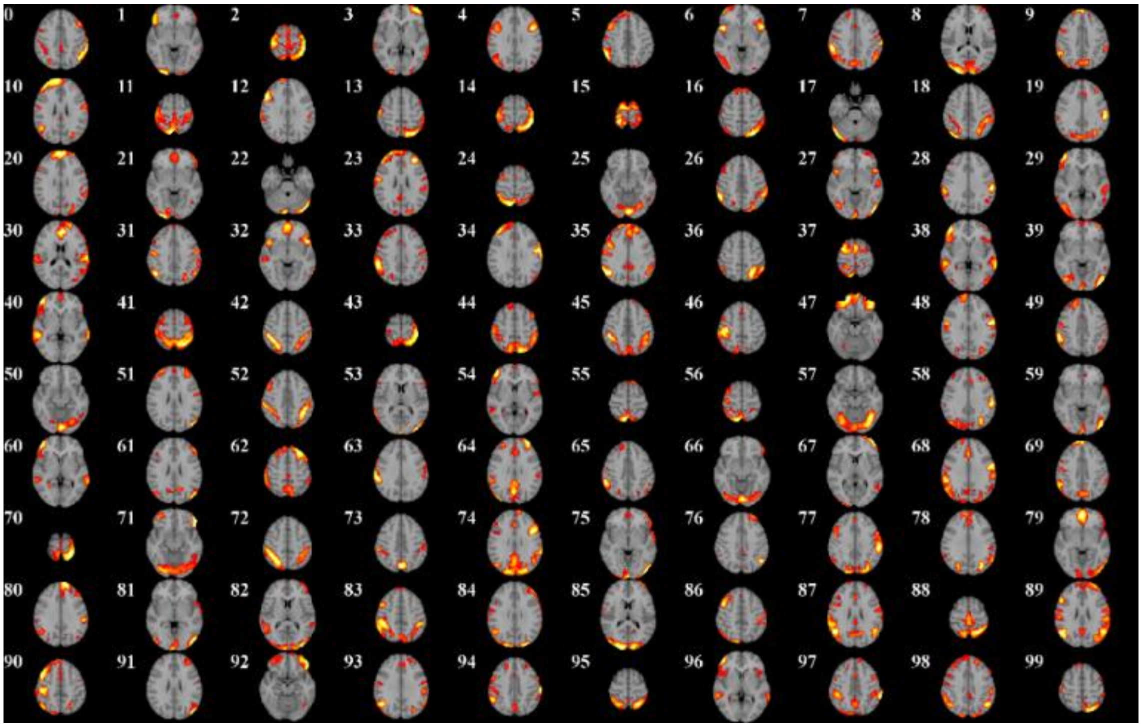


Figure. 2.3(c). Visualization of 100 FBNs learned in layer 3 of the VS-DBN from emotion task. Each network is visualized with one most informative axial slice.

The FBNs derived from the gambling task is shown in Figure. 4(a-c). In the gambling task, the participants play a card guessing game where they are asked to guess the number on a mystery card in order to win or lose money. Feedback is the number on the card the participants are provided with money because of completing the task, though it is a standard amount across subjects. (Barch et al., 2013) In layer 1, there are visual (no. 60 in Figure. 4(a)), frontal (and no. 62 in Figure. 4(a)), motor (no. 15 in Figure. 4(a)) and subcortical area (no. 51 in Figure. 4(a)), etc. In layer 2, there are default mode network (no. 71 in Figure. 4(b)) and frontopariatal network (and no. 2 in Figure. 4(b)), etc. In layer 3, there are default mode network (no. 72 in Figure. 4(c)) and frontopariatal network (no. 38 in Figure. 4(c)), etc.

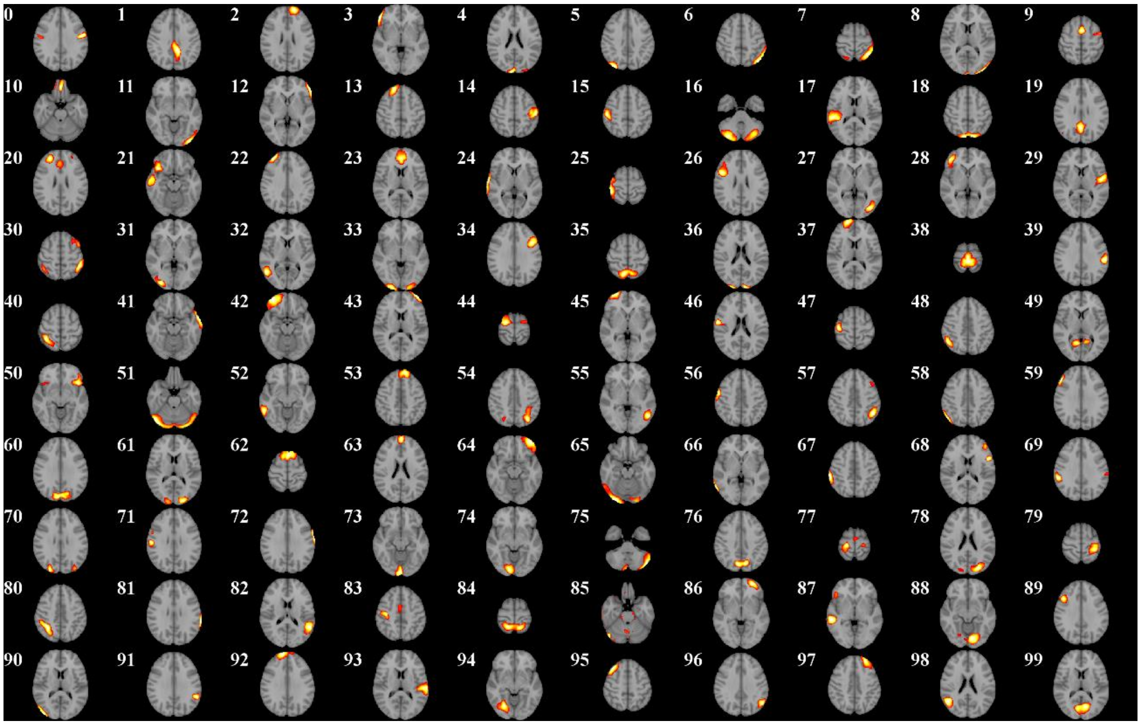


Figure. 2.4(a). Visualization of 100 FBNs learned in layer 1 of the VS-DBN from gambling task. Each network is visualized with one most informative axial slice.

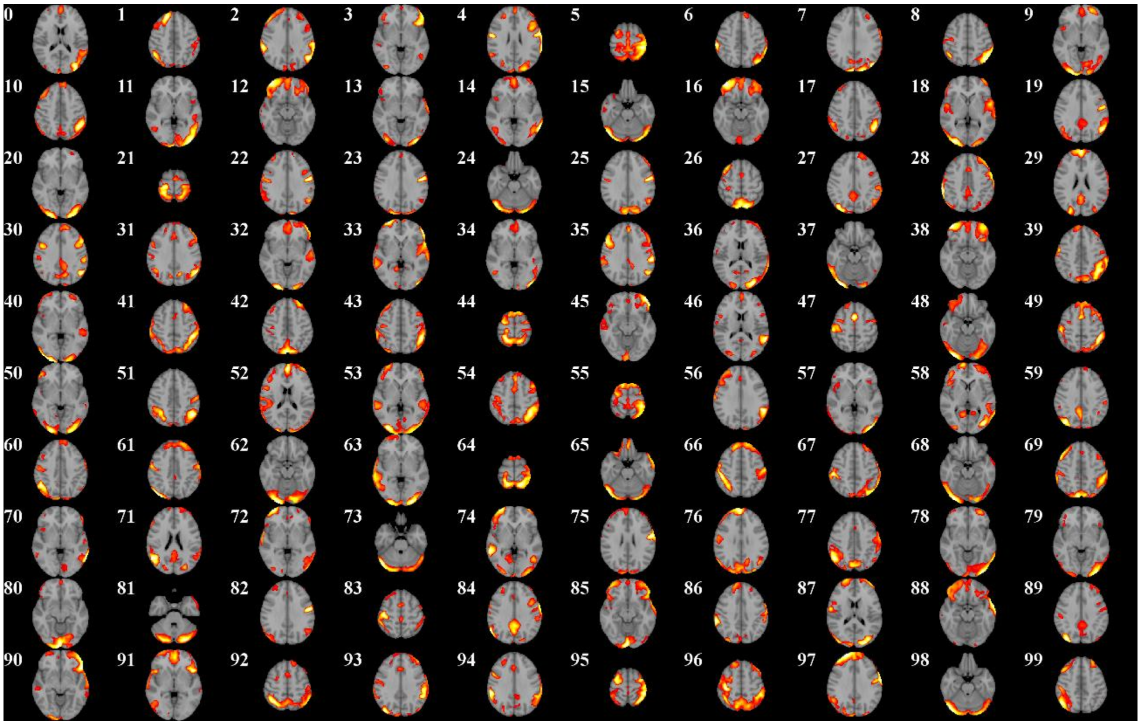


Figure. 4(b). Visualization of 100 FBNs learned in layer 2 of the VS-DBN from gambling task. Each network is visualized with one most informative axial slice.

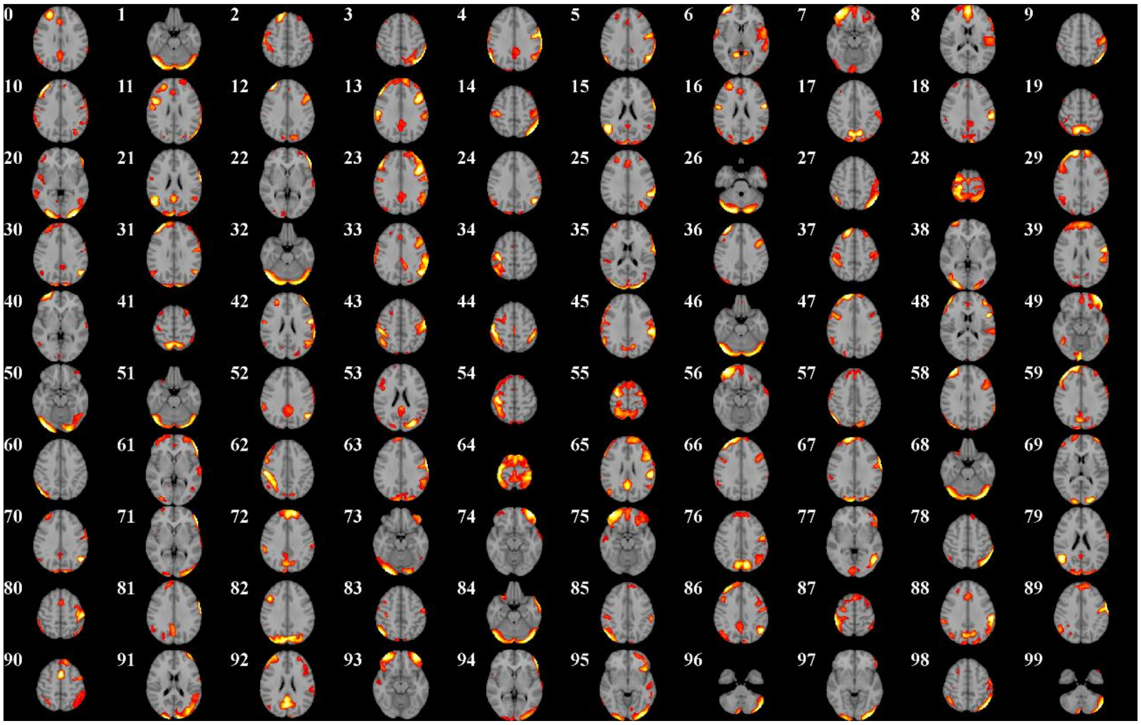


Figure. 2.4(c). Visualization of 100 FBNs learned in layer 3 of the VS-DBN from gambling task. Each network is visualized with one most informative axial slice.

By visual inspection, it is observed that these meaningful group-wise FBNs agree with neuroimaging knowledge and can be well interpreted. The FBNs from other 5 tasks are given in supplementary materials (Figure. 1-5). Notably, we obtained similar promising results in all HCP task fMRI datasets, suggesting the effectiveness of the VS-DBN. In the following sections, the FBNs are compared with and validated by other standard methods, and the hierarchy between layers is quantitatively analyzed.

2.7 Comparison of VS-DBN with GLM

To quantitatively evaluate the performance of DBN in modeling tfMRI data, a comparison study between VS-DBN results and the widely known GLM activation results is investigated in this section. Specifically, the GLM-based activation detection result is performed individually and averaged group-wisely. Task designs are 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 functional networks are thresholded at Z > 2.3 after transformation into “Z-scores” across spatial volumes.

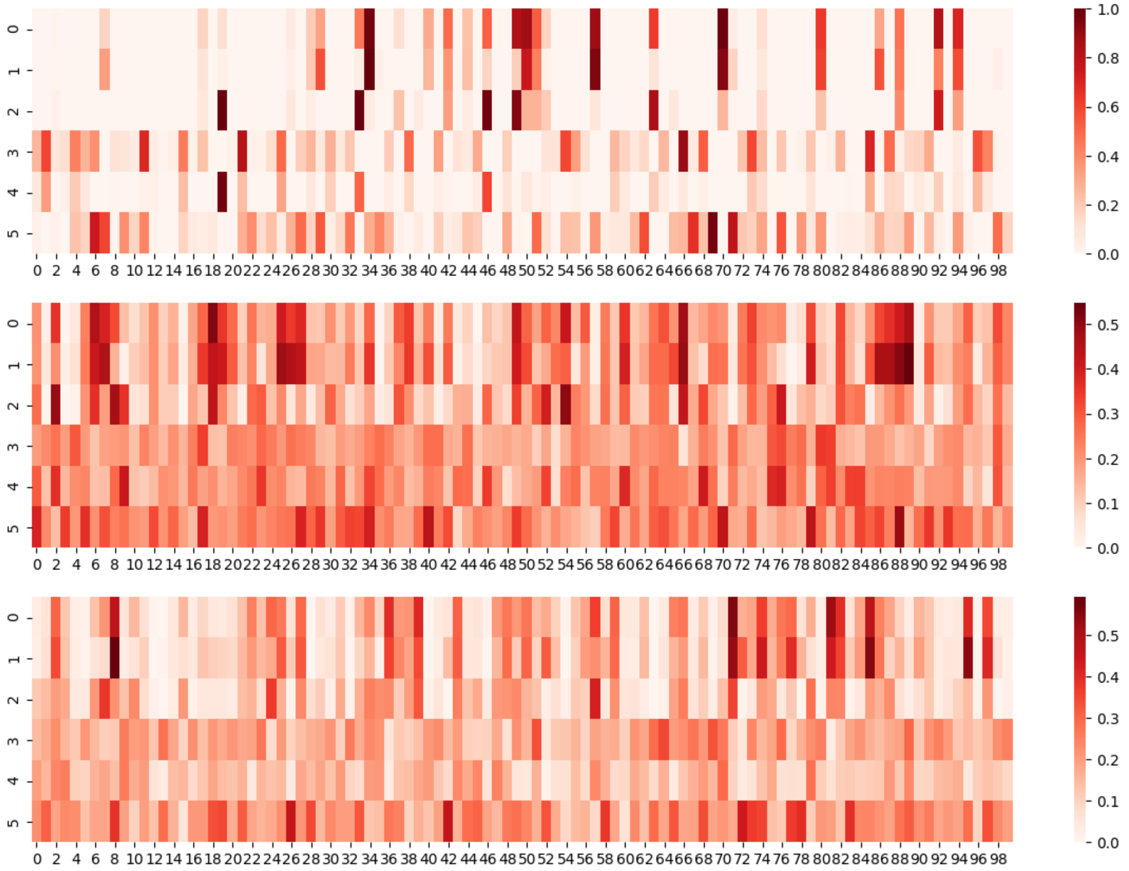


Figure. 2.5. OR between Emotion FBNs from GLM and emotion FBNs from DBN layer 1, layer 2, layer 3

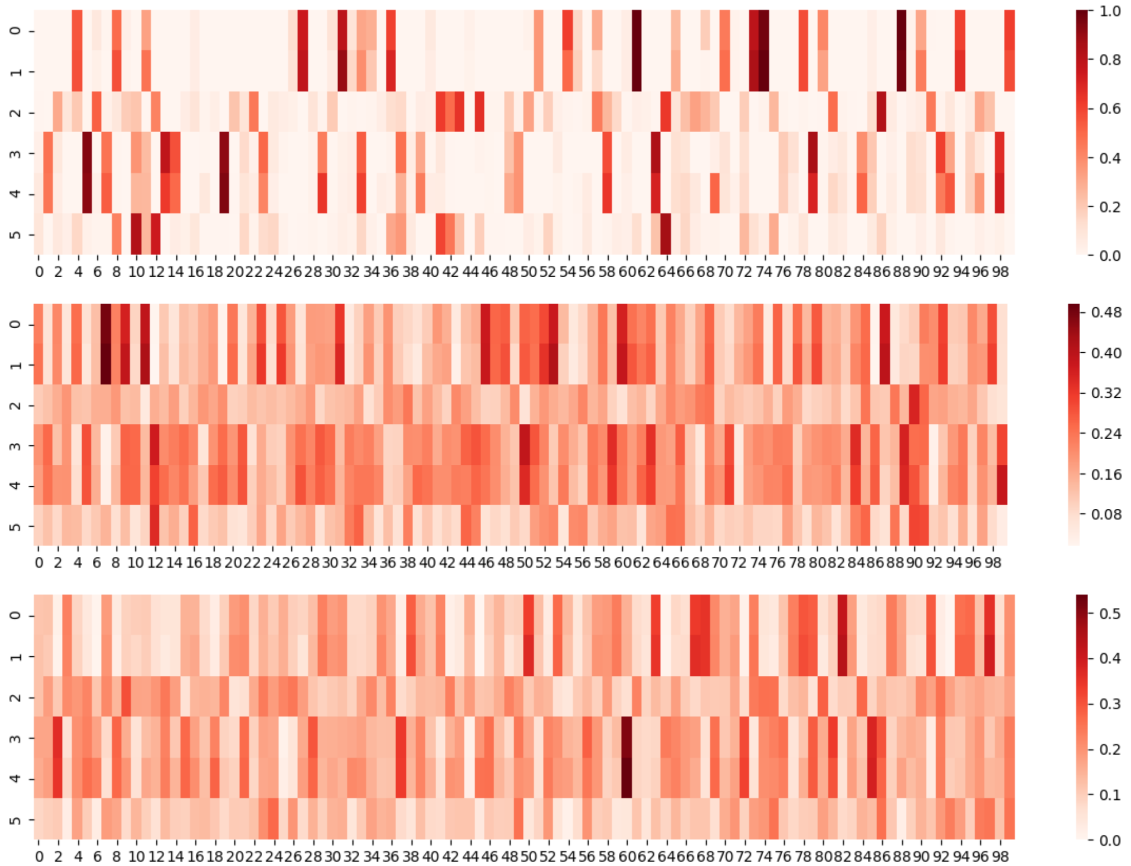


Figure. 2.6. OR between Gambling FBNs from GLM and gambling FBNs from DBN layer 1, layer 2, layer 3

To compare the functional networks derived by these two methods, the spatial overlap rate is defined to measure the similarity of two FBNs in accordance with previous studies (Lv, Jiang, Li, Zhu, Zhang, et al., 2015). Here, the spatial similarity is defined by the overlap rate (OR) between two functional networks $N^{(1)} and\ N^{(2)}$as follows, where N is the volume size:

$$OR\left(N^{(1)}, N^{(2)}\right) = \frac{\sum_{i=1}^{n}\left|N_i^{(1)} \cap N_i^{(2)}\right|}{\sum_{i=1}^{n}\left|N_i^{(1)} \cup N_i^{(2)}\right|} \qquad (12)$$

With the similarity measure defined above, the similarities $OR(N_{DBN}, N_{GLM})$ between the DBN derived functional networks N_{DBN} and the GLM derived functional networks N_{GLM} were quantitatively measured in Figure. 5 and 6. As shown and quantified in Figure.7, the GLM benchmark results are found in VS-DBN FBNs with ORs higher than 0.8. Such results demonstrated that VS-DBN FBNs well include the widely adopted GLM activation results, further suggesting the effectiveness and meaningfulness of the proposed model. The Comparison of all 7 tasks are given in supplementary materials (Figure. 6-10).

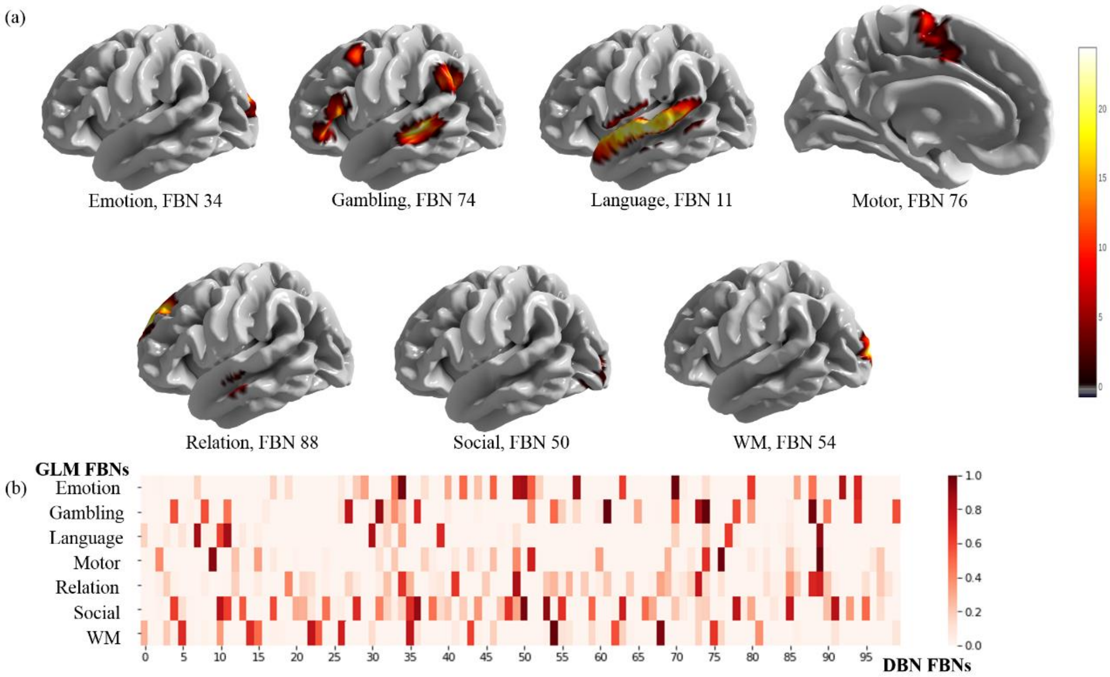


Figure. 2.7. FBNs validation with GLM. (a) Illustrative validated DBN FBNs for each task by GLM. (b)The OR between FBNs from GLM and DBN layer 1 for 7 different tasks.

2.8 Analysis of FBN Hierarchy

In addition to the validation result with GLM, it was also observed that lower layer networks in VS-DBN appear to be more localized, and the higher layer networks more global (as already shown in Figure. 3-4). To demonstrate and measure such association between two hierarchical functional networks, the inheritance similarity rate (ISR) between a lower layer network $N^{(L)}$and a higher layer network $N^{(H)}$ is defined as follows:

$$ISR\left(N^{(L)}, N^{(H)}\right) = \frac{\sum_{i=1}^{n} \left| N_i^{(L)} \cap N_i^{(H)} \right|}{\sum_{i=1}^{n} sign\left(N_i^{(H)}\right)} \tag{13}$$

As an example, the ISR between layer 2 and layer 1 and the ISR between layer 3 and layer 2 are shown in Figure.8-9. The ISR analysis of all 7 tasks are given in supplementary materials (Figure. 11-15). The widespread ISR between the networks across different layers can be clearly appreciated. More interestingly, the phenomenon of several lower layer networks merging into higher layer networks, clearly indicating the existence of traditionally hypothesized hierarchical structure of FBNs. These sparse large ISRs between associated networks in different layers of VS-DBN quantitatively confirmed the hierarchical organization of FBNs networks. It can also be seen that that the inheritance similarity between layer 2 and 3 is more complex than layer 2 and 1.

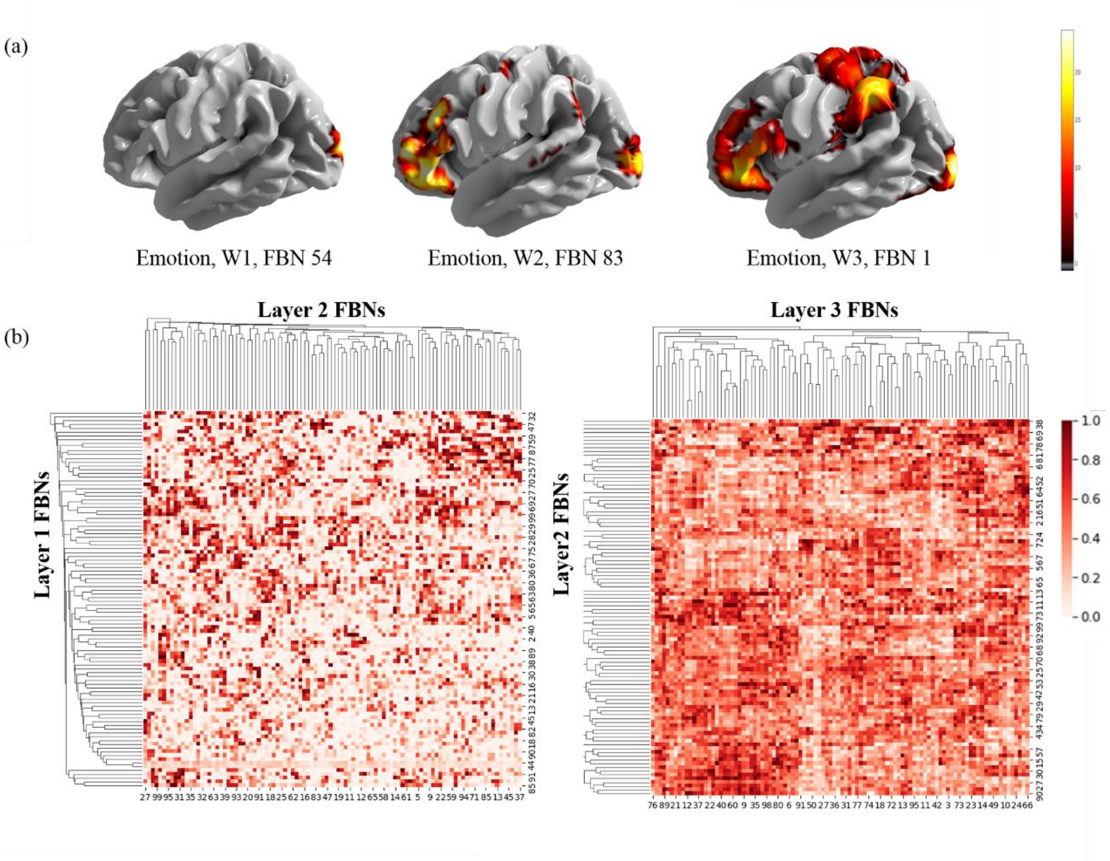


Figure. 2.8. The hierarchy property measured with ISR across VS-DBN layers in emotion task. (a) An illustration of a randomly selected hierarchy with ISR. (b) The left is the ISR between layer 2 and layer 1, and the right is the ISR between layer 3 and layer 2. There are widespread overlaps between the networks across different layers, thus forming a hierarchy of FBN networks.

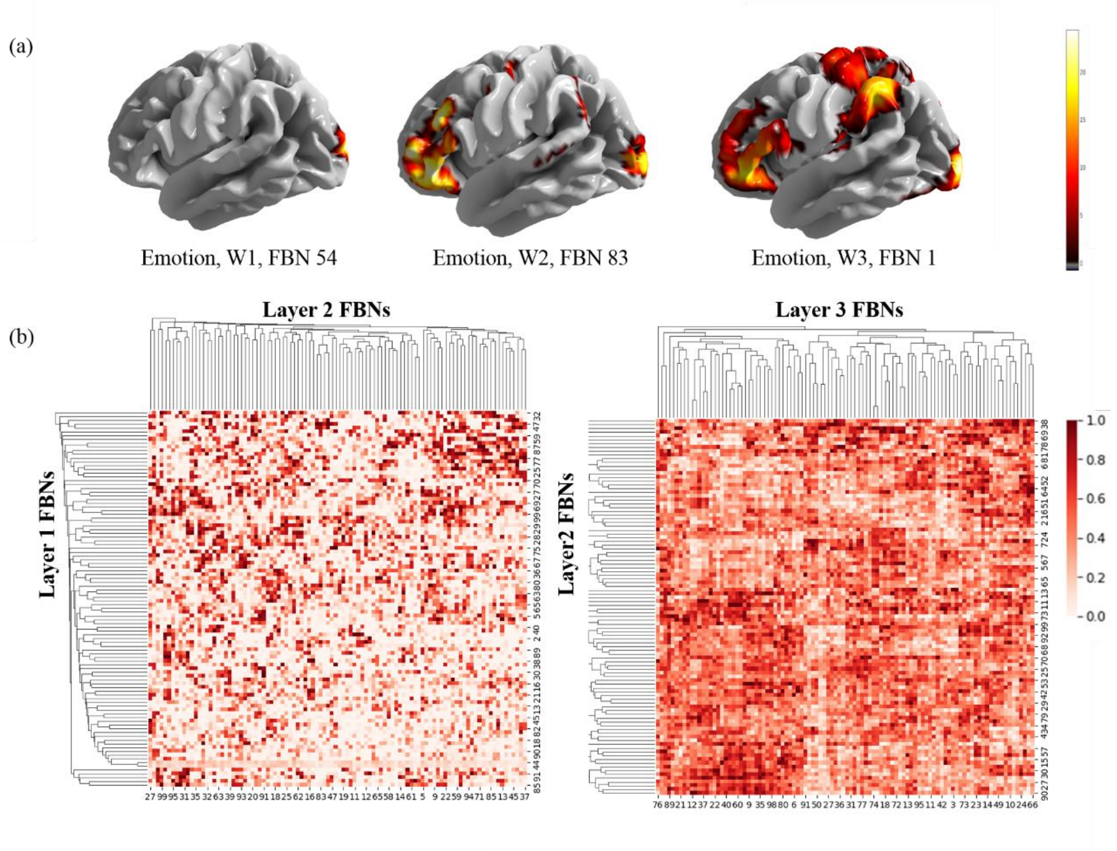


Figure. 2.9. The hierarchy property measured with ISR across VS-DBN layers in gambling task. (a) An illustration of a randomly selected hierarchy with ISR. (b) The left is the ISR between layer 2 and layer 1, and the right is the ISR between layer 3 and layer 2. There are widespread overlaps between the networks across different layers, thus forming a hierarchy of FBN networks.

Interestingly, the proposed VS-DBN can also model task-common FBNs besides task-specific FBNs since it has been observed that some FBNs exist in more than one task fMRI data. As shown in Figure. 10, the illustration is an example low-level network which indicates that VS-DBN can learn consistently task-common FBNs even though that the VS-DBN models were trained on different HCP tasks independently.

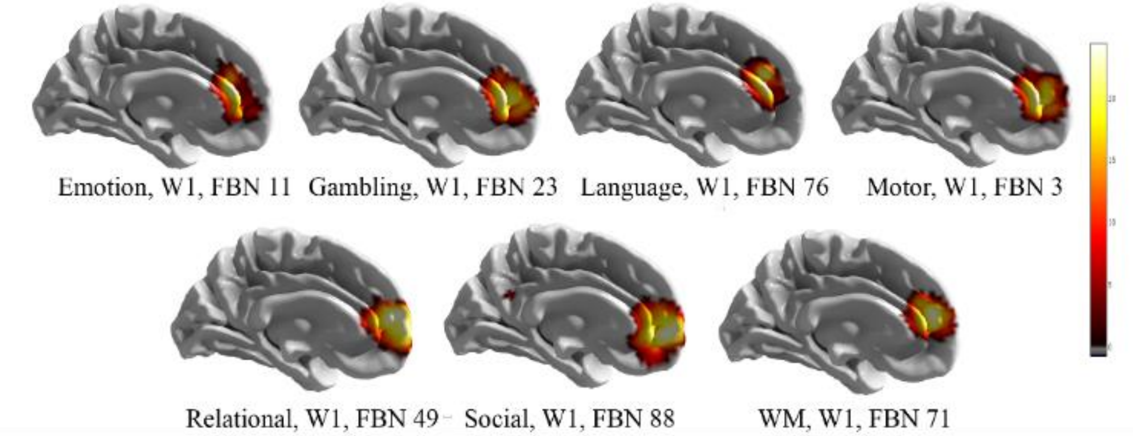
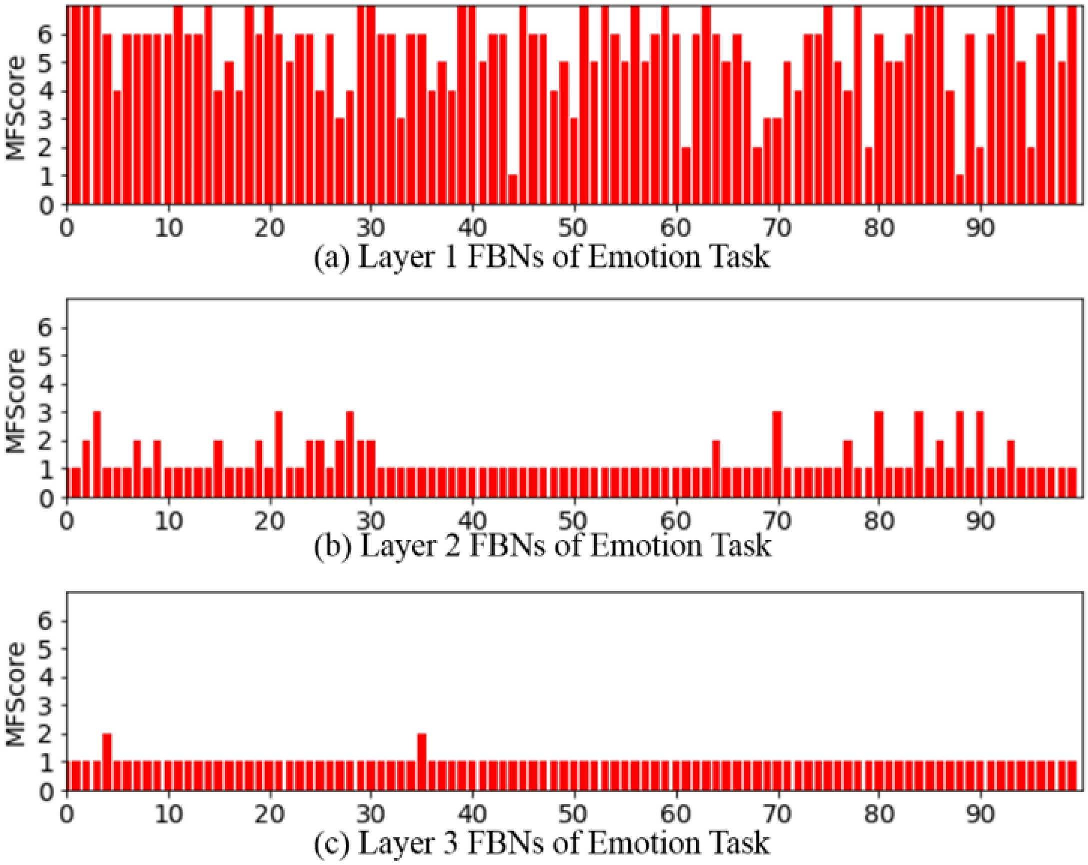


Figure. 2.10. Illustration of a high-MFScore FBN, which is the FBN of emotion layer 1 #11. These low-level FBNs are common in all 7 HCP tasks consistently.

To further investigate whether low-level FBNs tend to be common in more tasks than those learned from higher levels in VS-DBN, a Multifunction Score (MFScore) was defined as the number of tasks in which one FBN has ever existed and the OR was used to determine whether two FBNs are one or not. For unbiased comparison across FBNs from different layers, the OR thresholds were all set to 0.25.

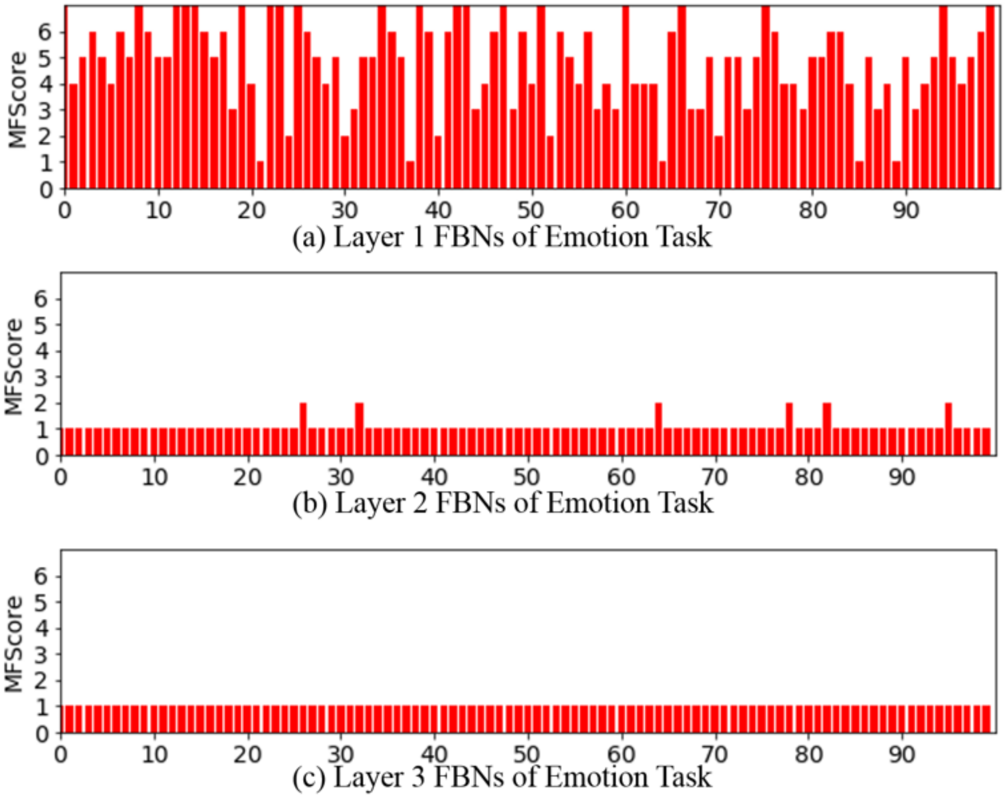


(a) Layer 1 FBNs of Emotion Task

(b) Layer 2 FBNs of Emotion Task

(c) Layer 3 FBNs of Emotion Task

Figure. 2.11. The MFScore of layer 1, 2 and 3 FBNs from emotion tasks. The higher is the MFScore, the FBNs participate in more tasks. Lower level FBNs participate more tasks than higher level FBNs.



(a) Layer 1 FBNs of Emotion Task

(b) Layer 2 FBNs of Emotion Task

(c) Layer 3 FBNs of Emotion Task

Figure. 2.12. The MFScore of layer 1, 2 and 3 FBNs from gambling tasks. The higher is the MFScore, the FBNs participate in more tasks. Lower level FBNs participate more tasks than higher level FBNs.

As shown in Figure. 11 and 12, the illustrated Emotion FBN #11 in Figure. 3(a) was confirmed to have a MFScore of 7, and many more low-level FBNs were common across different tasks. The MFScore analysis of other 5 tasks are given in supplementary materials (Figure. 16-20). Moreover, the MFScores of the FBNs from the first layers were larger than the second and third layers generally, which suggested that different tasks involve a variety of common low-level FBNs to perform some basic core task-common functions, while high level FBNs perform higher level task-specific functions.

2.9 Discussion

In this paper, we designed and implemented the VS-DBN model and exploited its capability of hierarchical representation of tfMRI data. With a group-wise experiment on massive HCP tfMRI data, the VS-DBN model quantitatively and qualitatively showed its promising capability of learning functional networks under a hierarchical structure. A comparison study using GLM validated that the functional networks learned by VS-DBN are meaningful and can be well interpreted. With networks at higher levels in the VS-DBN structure, the activated brain regions in a functional network tend to be larger and the patterns are more global involving both task-common and task-specific regions. It is inspiring that we observed some low-level task-related networks merging into one global task-related network layer, which indeed suggested the hierarchical architecture of FBNs. Also, our results on all 7 HCP task fMRI datasets show that different tasks exhibit a variety of common low-level FBNs to perform some basic core task-common functions, while high level FBNs perform higher level task-specific functions. In the future works, we will explore and optimize more configurations of VS-DBN's parameters, and further investigate the relationship of functional networks between different hierarchies and aim to interpret the corresponding neuroscientific meanings of the hierarchical organization of the brain functions in both healthy and diseased brains.

CHAPTER 3

NEURAL ARCHITECTURE SEARCH FOR OPTIMIZING DEEP BELIEF NETWORK MODELS OF FMRI DATA

3.1 Overview

It has been shown that deep neural networks are powerful and flexible models that can be applied on fMRI data with superb representation ability over traditional methods. However, a new challenge of neural network architecture design has also attracted attention: due to the high dimension of fMRI volume images, the manual process of network model design is very time-consuming and error prone. To tackle this problem, we proposed a Particle Swarm Optimization (PSO) based neural architecture search (NAS) framework for a deep belief network (DBN) that models volumetric fMRI data, named NAS-DBN. The core idea is that the particle swarm in our NAS framework can temporally evolve and finally converge to a feasible optimal solution. Experimental results showed that the proposed NAS-DBN framework can find robust architecture with minimal testing loss. Furthermore, we compared functional brain networks derived by NAS-DBN with general linear model (GLM), and the results demonstrated that the NAS-DBN is effective in modeling volumetric fMRI data.

Figure 1 summarizes our PSO-based NAS framework (Fig.1(A)) and DBN structure (Fig.1(B)) for modeling FBNs. The particle swarm consists of 30 particles, each of which represents a subnet with different initial architecture (Fig.1(A)). We investigated two main hyper-parameters including the number of layer and the number of neurons in each layer. These two parameters are used to construct a mapping between a particle position and a solution of network architecture design. The testing loss of DBN is regarded as the fitness function of PSO, which will be minimized in the searching process. The particle swarm can evolve and converge to an optimal solution. Then we applied this optimal architecture of DBN to model FBNs from task-based fMRI data (Fig.1(B)), and the weights of network are visualized and quantified as FBNs (Fig.1(C)), which will be further compared with GLM-derived network maps.

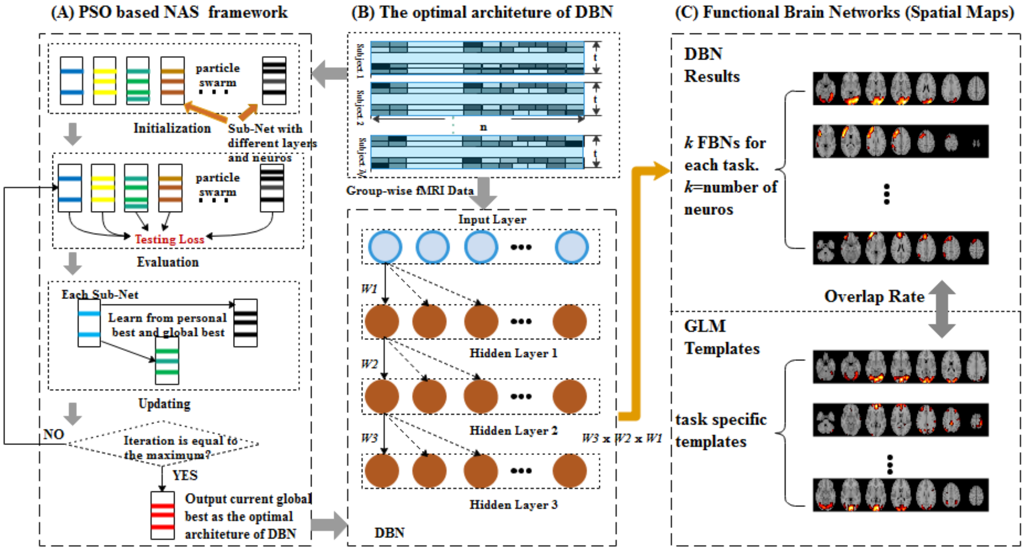


Fig. 3.1. Illustration of proposed NAS-DBN framework for deriving functional brain networks from task-based fMRI data

3.2 Background

Understanding the organizational architecture of functional brain networks has raised intense interest since the inception of neuroscience (Logothetis, 2008; Pessoa, 2014). In recent years, deep learning has attracted much attention in the field of machine learning and data mining, and it has been demonstrated to be a powerful tool for modeling brain networks based on fMRI data, compared to traditional shallow methods such as general linear model (GLM)(Beckmann, Jenkinson, & Smith, 2003), and independent component analysis (ICA)(Beckmann et al., 2005), and sparse dictionary learning (SDL) (Lv, Jiang, Li, Zhu, Zhang, et al., 2015). Although deep learning has enjoyed remarkable progresses over the past few years, most current neural network architectures were developed manually by researchers, which typically is a very time-consuming and error prone process, since all hyper-parameters of neural networks were decided by expert experiences. Fortunately, Neural Architecture Search (NAS), aiming to automatically search for optimal network architecture, is recently considered as a feasible and promising solution to the abovementioned problem. During recent years, several novel NAS methods, e.g., either based on reinforcement learning or evolutionary computation, have been developed and applied in a variety of deep learning tasks (Zoph & Le, 2016). However, due to the high dimension and complexity of volumetric fMRI data, there is still few NAS applications in the field of brain imaging using fMRI.

To fill the above gap, in this work, we firstly proposed a novel multi-layer volumetric deep belief network (DBN) and designed a group-wise scheme that aggregated multiple subjects' fMRI volume data for effective training of the DBN, with the purpose of discovering meaningful functional brain networks (FBN) in task-based fMRI data. Secondly, and more importantly, aiming to find out the optimal network architecture of DBN in modeling fMRI volumes, we developed a novel NAS framework based on particle swarm optimization (PSO). The key idea is that the particle swarm in the NAS framework will temporally evolve and finally converge to a feasible optimal solution. To quantitatively evaluate the performance of the NAS-DBN framework, a series of experiments have been conducted and the results showed the effectiveness of our design. Furthermore, we used the DBN with optimal architecture to extract FBNs from task-based fMRI data of Human Connectome Project (HCP) and compared the results with GLM-derived brain networks. Our results demonstrated that the NAS-DBN is a promising tool for deriving meaningful and interpretable FBNs from fMRI data.

3.3 Dataset and Preprocessing

In this paper, fMRI data from the Human Connectome Project (HCP) 900 Subjects Release was adopted as training dataset. The stimuli were projected onto a computer screen behind the subject's head within the imaging chamber, and 4 out of 7 categories of behavioral tasks are used, including Emotion, Gambling, Language, and Social. The fMRI preprocessing pipelines were implemented by FSL FEAT (FMRIB's Expert Analysis Tool) and Nilearn (Abraham et al., 2014), including spatial resampling to the MNI152 template, frequency filtering, detrending, normalization and masking. The details of acquisition parameters and information of each task can be found in the literature (Lv, Jiang, Li, Zhu, Zhang, et al., 2015).

3.4 PSO based NAS framework

Particle Swarm Optimization (PSO) is a swarm based evolutionary computation algorithm that is originally proposed by Kennedy and Eberhart in 1995 (Kennedy, 2010). Due to its numerous advantages, such as less parameter requirements, simple formula, easy to implement, PSO has become a popular tool for solving various complex optimization problems. In this work, we adopted and designed a PSO based NAS framework to search for the optimal network architecture of DBN. We designed a two-dimensional encoding method to map network architecture of DBN to a particle position.

The dimensions of the particle represent the number of layer and the number of neurons in each layer with the range of (2, 10) and (20, 200), respectively. To reduce computational cost, we assume the number of neurons in each layer is equal.

As shown in Figure 1(A), 30 particles are initialized in the solution space with initial velocities and positions. A particle position represents a solution of network architecture design, and the velocity of particle determines the particle’s next motion, which is affected by three factors: current motion, personal best position and global best position. The whole swarm is attracted by the global best and is exploring in the solution space, and at the same time each particle is exploiting its nearby space because of attraction of personal best. The process of exploring and exploiting also has a randomness, making PSO a stochastic and intellectual searching algorithm, thus the whole swarm can quickly converge to a feasible optimal solution compared to other exhaustive search algorithms.

The evolutionary process of particle swarm mainly consists of two steps: evaluation and updating. First, after initialization, all particles are evaluated by a fitness function which is defined by the testing loss of DBN. To avoid potential overfitting in NAS process, testing loss is adopted instead of training loss as an evaluation index of the model. After training, the trained model is applied to predict testing data (not used in training) and the Mean Squared Error (MSE) between input and output is calculated as testing loss, also the fitness value of corresponding particle. Notably, the input data was normalized to a Gaussian distribution for effective training. Then the personal best solution of each particle and the global best solution of whole swarm are recorded. Second, all particles’ velocities and positions are updated by the following equations:

$$v_{id}^{t+1} = w \cdot v_{id}^{t} + c_1 \cdot r_1 (p_{id}^{t} - x_{id}^{t}) + c_2 \cdot r_2 \left(p_{gd}^{t} - x_{id}^{t}\right) \tag{1}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{2}$$

Equations (1) and (2) are for velocity and position updating, respectively, where x_{id}^{t} and x_{id}^{t+1} are the current and next positions, respectively; v_{id}^{t} and $v_{id}^{t+1}$are the current and next velocities, respectively. The subscripts *t*, *i*, and *d* denote current iteration, subnet, and coding dimension, respectively; *w* is the inertia weight that reflects the inertia of particle motion; c_1 and c_2 are learning rate that affect the ratio of learning towards personal best and global best, making the searching process intelligent; r_1 and r_2 are two uniform random numbers selected from the interval of 0 to 1, which give the searching process a certain randomness. The second and third parts of the right side of Equation (1) reflect that current particle's next motion is affected by personal best position (p_{id}^{t}) and global best position (p_{gd}^{t}), as well as its previous motion. In addition, a uniform mutation strategy with variable mutation probability was introduced to increase the diversity of particle swarm. At the beginning of iteration, greater mutation probability makes the algorithm to have better exploration ability, and smaller mutation probability makes the algorithm to have better exploiting ability in the last stage of iteration. Therefore, the mutation probability was set to a linearly changing value from 0.2 to 0.05 with the increase of iteration number. Notably, we perform convergence check after initialization and updating, since some of particles might be divergent in the training process. These non-convergent subnets will be replaced by re-initiated subnets, and they will also be checked until they converge.

3.5 DBN Model of Volumetric fMRI Data

DBN, constructed by blocks of Restricted Boltzmann Machines (RBM), is widely used for deep generative models and has been proven to be a powerful tool for modeling fMRI data. Here, a group-wise volumetric scheme of DBN is proposed to model fMRI volumes. Considering the large inter-subject variability among human brains, arbitrary selection of a single individual may not effectively represent the population, thus a group-wise learning scheme is needed to reduce inter-subject variability by jointly registering the fMRI volumes to a common reference template corresponding to the group average. Since that the inter-subject variability is relatively more associated with the volatile time courses in different imaging sessions, it appears that taking volumes as input possibly works better than time series in terms of modeling the FBNs from fMRI data in this work. Accordingly, a volume from the fMRI data was taken as a feature, each time frame was taken as a sample, and a group-wise temporal concatenation was applied to all HCP subjects.

To reduce the possibility of overfitting and to improve generalization, a sparse weight regularization was designed and added in the DBN model. In each iteration, the weight was updated with the estimated gradient and an extra term of weight regularization derivative. In this paper, L1 weight penalty served as the regulation term while calculating the derivative of the sum of the absolute values of the weights. With a weight decay rate β, the overall optimization and weight update are formulated as follows:

$$\underset{W_{ij},a_i,b_j}{\text{minimize}} \ \ln \mathcal{L}(P(v)) + \beta \left\| W_{ij} \right\|_1 \qquad (3)$$

$$W_{ij} \leftarrow W_{ij} + \alpha(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon} - \beta \cdot sign(W_{ij})) \qquad (4)$$

The sparse weight regularization works by causing many of the weights to become zero while allowing a few of the weights to grow large. In the context of fMRI data, L1 regularization can denoise the FBNs and improve interpretability by suppressing useless weights and allowing important model parameters to become larger, which is considered as an important methodology/technical contribution of this work and it will be demonstrated in the following results sections. With respect to interpreting a trained DBN in the fMRI context, each row of weight vector was mapped back into the original 3D brain image space, which was the inverse operation of masking in preprocessing steps and was interpreted as an FBN. After the DBN was trained layer-wisely on a large-scale task fMRI dataset, each weight showed the extent of each voxel contributed to a latent variable. For deeper layers, the linear combination approach was used to interpret the connection. With this approach, as an example, $W_3 \times W_2 \times W_1$ was visualized for the first hidden layer as FBNs (Fig.1(C)).

3.6 Implementations

The NAS-DBN is inherently much more computationally expensive, compared to DBN models for temporal fMRI time series. Considering HCP 4D images and one single layer of RBM, there are around 20K trainable parameters for temporal fMRI time series DBN, but 20 million for volumetric fMRI DBN. Moreover, the population size and iteration size will put significant computational burden on the NAS process. To deal with this problem, in this paper, the TensorFlow (Abadi et al., 2015), which is a popular deep learning framework and provides great convenience coding with GPUs, was adopted with high efficiency GPU computation to fill the gap. Based on TensorFlow, we designed and implemented a fast and flexible DBN model. Limited by computing resources, all subnets will be trained one by one and processed collectively. The code was run on a deep learning server with GeForce GTX 1080 TI of GPU and 32Gb of RAM.

3.7 Comparisons between NAS-DBN and DBN

To quantitatively evaluate the effectiveness of our NAS-DBN framework, we ran 10 times of the searching process independently, and analyzed the statistical results. We used 4 shuffled HCP tasks data as input of NAS. After NAS, we used the same optimal architecture of DBN to model each task data independently. As shown in Figure 2, the optimal results show high consistence and robustness in the optimal number of layer and the optimal number of neurons. In most runs (8 out of 10), the result of the optimal number of layer is 3, except only two results are 2 and 4 respectively. The average result in the optimal number of neurons is 80, and all results are in a range from 69 to 112.

These statistical experiments demonstrated that our NAS framework can generate reliable results of architecture design. Furthermore, we compared the testing loss of DBNs with optimal architecture and manually selected architectures. Figure 3 shows comparison of testing loss of 4 DBNs with the same number of neurons and comparison of testing loss of 4 DBNs with the same layers. DBN (3,80) denotes that there are 3 hidden layers and 80 neurons in this DBN structure. DBN with the optimal architecture from NAS has the lowest testing loss of 0.0213 compared to other manually designed DBNs, demonstrating the effectiveness of NAS framework.

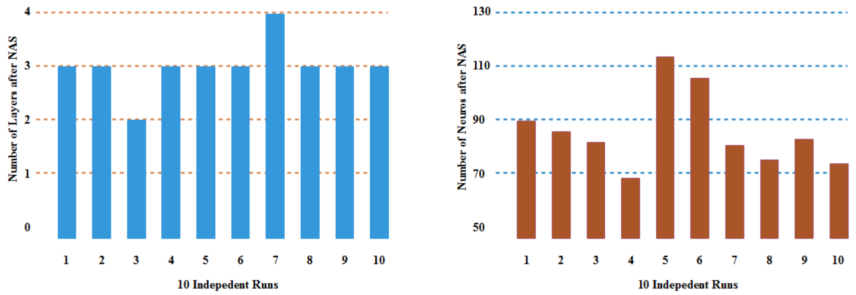


Fig. 2. Statistical results of 10 independent experiments in the optimal number of layers and optimal number of neurons.

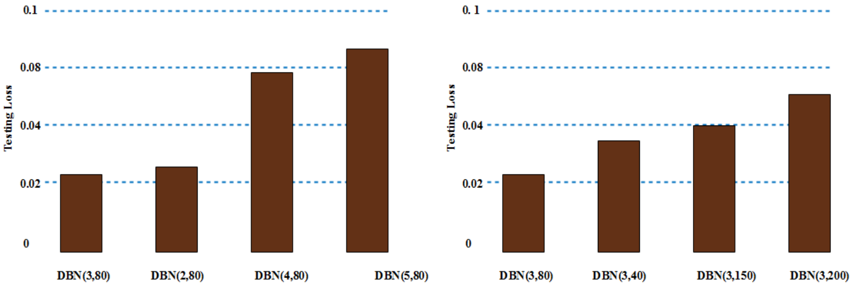


Fig. 3. Testing loss of DBNs with same neurons and Testing loss of DBNs with the same layers

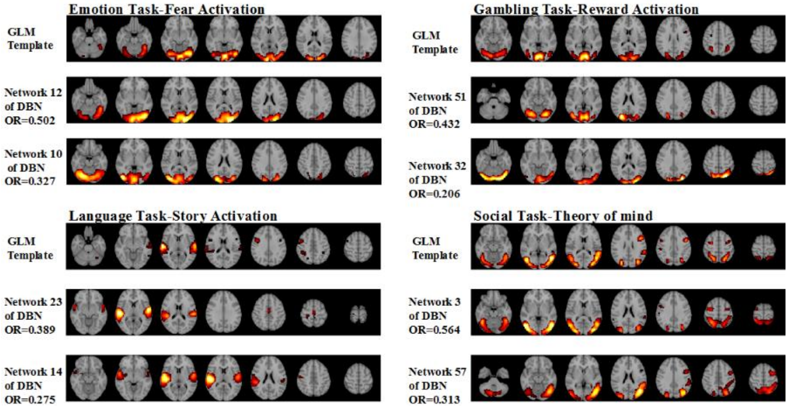
3.8 Comparison of NAS-DBN with GLM

To explore the representation of task-based fMRI data, four task-specific DBNs were trained on fMRI data of 4 HCP tasks independently using the same hyperparameters. To quantitatively evaluate the performance of DBN in modeling tfMRI data, a comparison study between NAS-DBN results and the widely known GLM activation results is investigated in this section. For fair comparison, all the functional networks derived by these two methods are thresholded at Z > 2.3 after transformation into "Z-scores" across spatial volumes. The spatial overlap rate is defined to measure the similarity of two FBNs in accordance with previous literature studies. Here, the spatial similarity is defined by the overlap rate (OR) between two functional networks $N^{(1)} and\ N^{(2)}$as follows, where n is the volume size:

$$OR\left(N^{(1)}, N^{(2)}\right) = \frac{\sum_{i=1}^{n}\left|N_i^{(1)} \cap N_i^{(2)}\right|}{\sum_{i=1}^{n}\left|N_i^{(1)} \cup N_i^{(2)}\right|} \tag{5}$$

With the similarity measure defined above, the similarities $OR(N_{DBN}, N_{GLM})$ between the NAS-DBN derived functional networks N_{DBN} and the GLM derived functional networks N_{GLM} were quantitatively measured. For each of GLM template, we found the most similar FBN derived by NAS-DBN with high OR in all 4 HCP tasks. Notably, we developed in-house GLM codes and obtained our own templates derived by group-wise fMRI data, which are quite like the widely known GLM templates (Barch et al., 2013). Figure 4 shows the comparison of FBNs derived by NAS-DBN and GLM templates in 4 tasks. We selected one specific stimulus for each task and the corresponding GLM templates were all found in NAS-DBN FBNs of these tasks. For emotion task, we can see fear stimulus activated GLM template, and the most similar FBN from NAS-DBN, which is the 12th network out of 80 networks. Comparing this brain network with the benchmark GLM template, the overlap rate is as high as 0.502, and thus it is easy to recognize the close match between them. Since our NAS-DBN is an unsupervised architecture, there are other similar FBNs that can be detected in all 4 tasks. For instance, we found a similar fear activation network in emotion task, and the overlap rate between this network and GLM template is 0.327. For other three tasks, including gambling, language, and social, we also found 2 most similar FBNs compared to GLM templates.

Furthermore, we detected several resting state networks (RSNs) though our NAS-DBN model including the default mode network (in emotion and social tasks), visual network (in gambling and social tasks), auditory network (in gambling and language tasks), and frontoparietal network (in emotion task), demonstrating that our NAS-DBN model can derive not only task activated networks but also resting state networks. As shown in Figure 5, 4 RSNs were found and visualized in our DBN-derived FBNs in different tasks. Here, we used the RSN templates from Nilearn (Y. Zhao, Ge, Zhang, et al.) as benchmark, and details of RSNs can be found in the literature (Smith et al., 2009).



Emotion Task-Fear Activation
GLM Template
Network 12 of DBN OR=0.502
Network 10 of DBN OR=0.327
Gambling Task-Reward Activation
GLM Template
Network 51 of DBN OR=0.432
Network 32 of DBN OR=0.206
Language Task-Story Activation
GLM Template
Network 23 of DBN OR=0.389
Network 14 of DBN OR=0.275
Social Task-Theory of mind
GLM Template
Network 3 of DBN OR=0.564
Network 57 of DBN OR=0.313

Fig. 4. Comparison between GLM templates and similar FBNs derived by NAS-DBN in emotion, gambling, language, and social tasks. Each network is visualized with 7 axial slices.

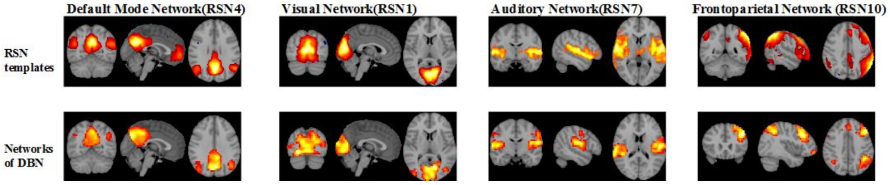


Fig. 5. Comparison between RSN templates and similar FBNs derived by NAS-DBN in different tasks. Each network is visualized with 3 most informative orthogonal slices.

3.9 Discussion

We proposed a novel PSO based NAS-DBN framework for searching optimal architecture of DBN in modeling FBNs from volumetric fMRI data. Based on evolutionary computation, 30 subnets in our framework learn experience of each other, and the whole swarm can evolve and finally converge to a feasible optimal architecture of

DBN. We selected testing loss as fitness function of NAS, instead of training loss, to avoid potential overfitting problems. The statistical experiment of NAS showed high consistence and robustness of our architecture design. Furthermore, we showed that our DBN model can derive both task specific functional networks and resting state networks which are meaningful and can be well interpreted. The promising results by showed the importance of optimizing neural network structures in deep learning.

CHAPTER 4

SPATIAL-TEMPORAL DECOMPOSITION OF CONNECTOME-SCALE BRAIN NETWORKS BY DEEP SPARSE RECURRENT AUTOENCODERS

4.1 Overview

Exploring the spatial patterns and temporal dynamics of human brain activities has been of intense interest to better understand connectome-scale brain networks. Though modeling spatial and temporal patterns of FBNs has long been a research topic, the development of a unified and simultaneous spatial-temporal model to realize such a purpose is challenging. For instance, although some deep learning methods have been proposed recently to model FBNs, most of them can only represent either spatial or temporal perspective of functional Magnetic Resonance Imaging (fMRI) data and rarely model both domains simultaneously. Inspired by the recent success in applying sequential auto-encoders for brain decoding, in this paper, we proposed a novel deep sparse recurrent auto-encoder (DSRAE) in an unsupervised way to learn spatial patterns and temporal fluctuations of brain networks jointly. The proposed DSRAE were evaluated and validated on three tasks of the publicly available human connectome project (HCP) fMRI dataset with promising results. To our best knowledge, the proposed DSRAE is among the early efforts in developing unified models that can extract connectome-scale spatial-temporal networks from 4D fMRI data simultaneously.

Figure.1 summarizes the architecture of our proposed DSRAE model, as well as the two validation experiments. After data preprocessing (Sec. B), the fMRI data in each subject is extracted and normalized. Based on RNN (Sec. C), we trained the DSRAE model in an unsupervised manner (Sec. D) with the fMRI data. Particularly, the activation of each hidden node represents a typical functional brain state, and its hidden response to specific stimulus represents the temporal activities of the brain states.

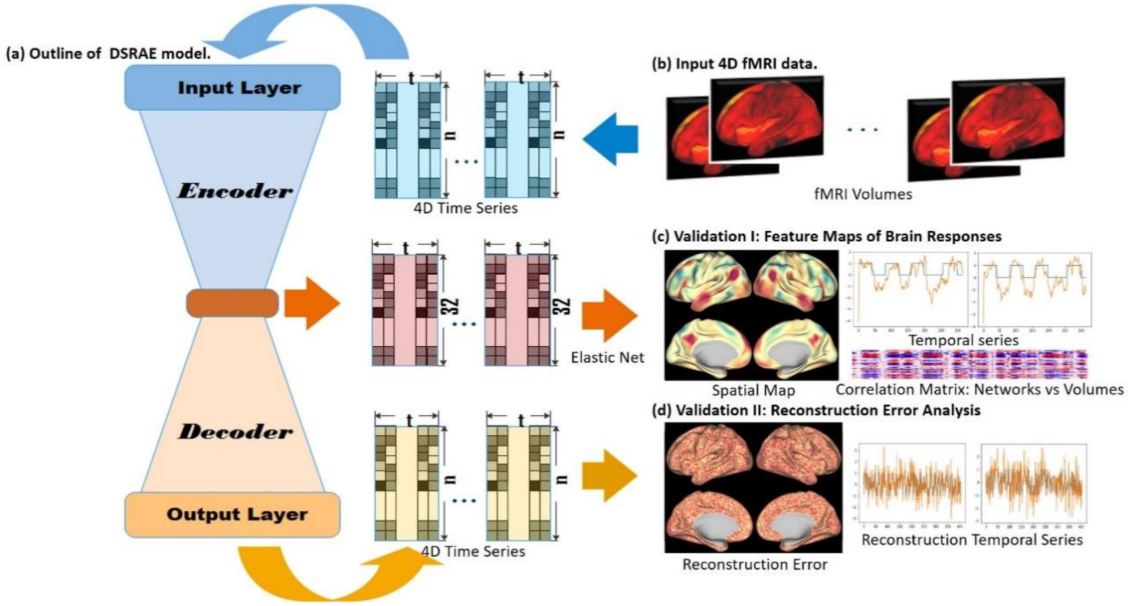

(a) Outline of DSRAE model.
Input Layer
Encoder
Decoder
Output Layer
t
n
32
4D Time Series
Elastic Net
(b) Input 4D fMRI data.
fMRI Volumes
(c) Validation I: Feature Maps of Brain Responses
Temporal series
Spatial Map
Correlation Matrix: Networks vs Volumes
(d) Validation II: Reconstruction Error Analysis
Reconstruction Temporal Series
Reconstruction Error

Figure. 4.1. Illustration of DSRAE model and the two validation experiments. (a) The outline of DSRAE model, in which the input and output layers are both fMRI time series, and the latent layer features are the networks with temporal information. (b) The input of DSRAE is the 4D fMRI data, which is a series of 3D brain volumes acquired in each task session. (c) Validation I: spatial maps, derived from latent layer time series via Elastic Net, are compared with benchmark maps via GLM; the time series are validated by the task stimulus curve; and the correlation matrix is calculated between the obtained networks and original volumes to detect the brain states. (d) Validation II: The signal reconstruction error analysis based on Pearson correlation between reconstructed output and original input.

4.2 Background

Exploring human brain function and its dynamics has been one of the important topics in modern neuroscience (Logothetis, 2008; Pessoa, 2014). Mounting evidence shows that the human brain undergoes noisy, massive, and complex neural processes that are highly correlated both spatially and temporally(Friston, 1997; Shimony et al., 2009; Smith et al., 2012), which suggests that it is desirable to model spatial and temporal information at the same time to account for the relationship between a stimulus such as cognitive task and the brain response measured with fMRI(Derado, Bowman, & Kilts, 2010; Shen, Mayhew, Kourtzi, & Tiňo, 2014; Woolrich, Jenkinson, Brady, & Smith, 2004). To better explore brain networks based on fMRI data, inspired by the successful applications of machine learning methods, a variety of data-driven methods have been proposed on fMRI data, such as sparse dictionary learning (SDL) (Lv, Jiang, Li, Zhu, Zhang, et al., 2015; W. Zhang et al., 2019), deep belief network (DBN)(Hu et al., 2018; Plis et al., 2014), convolutional neural network (CNN)(Huang, Hu, Zhao, et al., 2018), and recurrent neural network (RNN) (Cui et al., 2018; Han Wang; Wang et al., 2018). These computational methods can be generally classified into two categories: spatial approaches and temporal approaches. Studies based on spatial approaches usually focused on the spatially decomposed components derived from fMRI data and typically ignored temporal dynamics information(Jiang et al., 2015; Lv, Jiang, Li, Zhu, Zhang, et al., 2015; Plis et al., 2014). For examples, Lv *et al.* (Lv, Jiang, Li, Zhu, Zhang, et al., 2015)used SDL to investigate the brain's spatial functional networks from fMRI data; Jiang *et al. (Jiang et al., 2015)* used the sparse representation to characterize the spatial functional regions with task-based fMRI data. On the other hand, studies based on temporal approaches mostly focused on temporal features modeling while spatial information is overlooked (Hu et al., 2018; Huang, Hu, Zhao, et al., 2018; Wang et al., 2018). For examples, Hu *et al. (Hu et al., 2018)* used restricted Boltzmann machine (RBM) to interpret fMRI temporal courses; Huang *et al. (Huang, Hu, Zhao, et al., 2018)* used deep convolutional auto-encoder (DCAE) to derive the features from task-based fMRI time series. Notably, a recent study based on deep spatial-temporal convolutional neural network (ST-CNN) tried to take the advantages of both spatial and temporal domains(Y. Zhao, Li, et al., 2018), yet the temporal features were derived from the spatial features inherently. In general, these previous studies focused on either spatial or temporal perspective of fMRI data and rarely modeled both domains simultaneously, thus, few of them has the ability to model the spatial-temporal variation patterns of FBNs. Therefore, a comprehensive and systematic framework is still in great need to recognize dynamic, temporal brain states at connectome-scale and model the brain's spatial-temporal dynamic activities simultaneously. However, development of such a comprehensive framework faces major challenges including the lack of ground truth of underlying neural activities and the inherent complexity associated with those spatial-temporal patterns of connectome-scale functional networks (Huang, Hu, Zhao, et al., 2018; Wang et al., 2018).

Recently, RNN has been widely employed in many research areas, such as language modeling and handwriting recognition. Due to its great promise and performance to capture the temporal dependences in the data sequences effectively, RNN has been also introduced to model dynamic biological signals(Schuster & Paliwal, 1997). What's more, RNN can preserve the information of the past sequences and restore the memories with the memory cells to make predictions, rather than predicting based only on current samples in time serie. Therefore, RNN can effectively and easily acquire the temporal dependence of the sequential data to model temporal dynamic activities that can help to recognize the brain states from fMRI data. Recently, Wang *et al.* (Wang et al., 2018)proposed a deep sparse RNN (DSRNN) framework based on the traditional RNN to recognize functional brain states with task-based fMRI data and achieved promising results. However, the work in (Wang et al., 2018)still focused on the temporal information analysis and did not analyze spatial-temporal variation patterns of FBNs simultaneously and its training process relies on the label/annotation of each frame of the neuroimages.

Inspired by the previous studies of using RNN for video frames reconstruction (Srivastava, Mansimov, & Salakhudinov, 2015), neuros dynamic analysis (Pandarinath et al., 2018), speech recognition(Pei & Tax, 2018), and recent success of using RNN for fMRI data modeling (Wang et al., 2018), in this paper, we designed a novel deep sparse recurrent auto-encoder (DSRAE) in order to bridge the above-mentioned gaps and better understand the spatial-temporal patterns of connectome-scale FBNs. To our best knowledge, our proposed model is among the early efforts of developing a unified spatial-temporal model for 4D fMRI data in the literatures, and notably our model can be applied on both task-based fMRI (tfMRI) and resting fMRI (rfMRI) due to the entirely unsupervised training process, which can effectively deal with the above-mentioned two challenges. That is, our DSRAE model naturally represents the complex spatial-temporal patterns of connectome-scale functional networks and the training process of DSRAE does not need the spatial or temporal labels/annotations of neural activities. Conceptually, the DSRAE is a unified spatial-temporal data-driven framework that can jointly characterize and recover the embedded spatial and temporal information from 4D fMRI data. Specifically, the proposed DSRAE model contains four layers for both encoder part and decoder part, respectively. The encoder part includes one input layer, one fully connected layer and two recurrent layers; the decoder part includes two recurrent layers, a fully connected layer and an output layer. To evaluate the performance of our DSRAE model, two validation experiments of reconstruction error calculation and spatial-temporal feature interpretation were performed on three tfMRI datasets (Language, Working Memory and Gambling) of 791 subjects of Human Connectome Project (HCP)(Barch et al., 2013). Our experimental results demonstrated that the proposed DSRAE framework can learn the representations of both task-evoked spatial networks and their temporal time series fluctuations across all subjects and tasks effectively and robustly. The FBNs derived in the spatial domain showed great consistency with those derived by the traditional General Linear Model (GLM), and the temporal time series fluctuations of those networks also exhibited relatively high correlations with the task stimulus curves correspondingly. In general, the meaningful and interpretable spatial-temporal functional network patterns derived from our proposed DSRAE framework offer a new approach to represent the human brain function based on fMRI data, contributing to the understanding of functional mechanisms of the human brain. The source codes of our DSRAE model and associated sample datasets will be released at: https://github.com/ChloeLeeBnu/DSRAE.

4.3 Dataset and Pre-processing

In this paper, we used 791 healthy adult participants that executed all seven tasks in HCP 900 Subjects Data Release (Barch et al., 2013) with grayordinate tfMRI dataset and it is publicly available on https://db.humanconnectome.org. Here, three behavioral tasks were used in this work, including Language, Working Memory and Gambling tasks. The detailed acquisition parameters were as follows: TR=720 ms, TE=33.1 ms, flip angle=52°, in-plane FOV=208 mm×180 mm, 104×90 matrix, slice thickness=2 mm, 72 slices, multiband factor=8, echo spacing=0.58 ms, BW=2290 Hz/Px. The HCP grayordinate data model the gray matter as combined cortical surface vertices and subcortical voxels across subjects in the standard MNI space. The preprocessing steps of the fMRI dataset include spatial smoothing, temporal filtering, nuisance regression, and motion censoring, which were all implemented with FreeSurfer Software (Glasser et al., 2013).

In our experiments in this paper, we used the fMRI data of subjects who performed all three tasks. In language task, there were two stimulations: story and math, in which the math blocks were designed to provide a comparison task that was attentionally demanding(Binder et al., 2011). Story blocks presented subjects with auditory stories, which transcribed from Aesop's fables, and two options related to the topic asked subjects to choose subsequently. Math blocks presented participants addition and subtraction problems with two alternative options, and participants needed to choose the right answer. The lengths of math stimulations were corresponding to the story stimulations. In working memory task, a version of the N-back task was used to assess working memory capability (Drobyshevsky, Baumann, & Schneider, 2006). In the design, 2-back task (respond ‘target’ whenever the current stimulus was the same as the one 2-back) and 0-back task (respond ‘target’ only if the fixation was presented) fMRI data were collected to study the brain states for working memory. It was reported that the associated brain activations were reliable across subjects (Drobyshevsky et al., 2006) and time (Caceres, Hall, Zelaya, Williams, & Mehta, 2009). In gambling task, participants played card guessing games in which they needed to guess the number on a card, and won money if hit, or lost money if missed (Caceres et al., 2009). Prior studies demonstrated that the task elicited activations in the striatum and some other reward related regions which were robust and reliable across the subjects (Caceres et al., 2009; Forbes et al., 2009; Tricomi, Delgado, & Fiez, 2004). More experiment details are available in the supplemental materials.

4.4 Basics of Recurrent Neural Network (RNN)

RNN is a kind of feedforward neural network, which uses a internal state (memory) to process sequences of inputs. Due to this advantage, RNN has been applied in diverse tasks for sequence modeling (Sak et al., 2014; Wang et al., 2018; Yamins & DiCarlo, 2016). As illustrated in Figure.2a, each recurrent neural network layer can be unfolded as a feedforward network along temporal series. However, because of the vanishing gradient, the traditional basic RNN is difficult to maintain expertise learned on earlier data. To overcome this problem, Long Short-Term Memory (LSTM) unit (Hochreiter & Schmidhuber, 1997), with the “forget gate”, was specifically designed and have become one of the most widely-used RNN architectures.

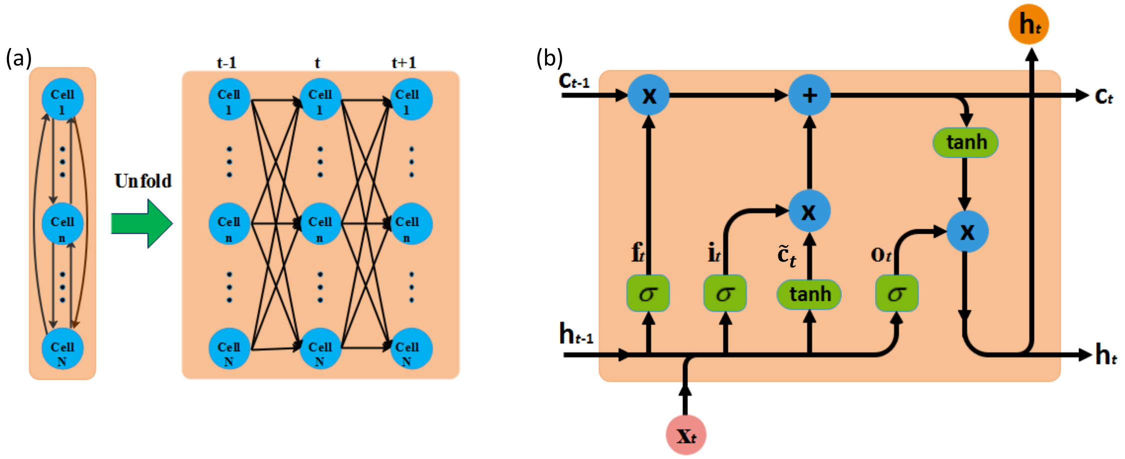


Figure. 4.2. A schematic illustration of LSTM cell model. (a) The interconnections in a common recurrent hidden layer. (b) LSTM memory unit.

The internal memory, which stores information from previous time points, is in a cell state of each LSTM unit. As shown in Figure.2b, there are gates in each unit to control the contents of the cell states $\boldsymbol{c}_t$ and determine the outputs $\boldsymbol{h}_t$ based on the inputs $\boldsymbol{x}_t$. These gates regulate the ability to remove or add information to the cell state of LSTM. The cell state of an LSTM unit is defined as follows:

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{c}}_t \qquad 1)$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{b}_f\right) \qquad 2)$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{b}_i) \qquad 3)$$

$$\tilde{\boldsymbol{c}}_t = tanh(\boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{b}_c) \qquad 4)$$

where $\boldsymbol{f}_t$ and $\boldsymbol{i}_t$ are the forget gate and input gate activities respectively, $\tilde{\boldsymbol{c}}_t$ are auxiliary variables, $\boldsymbol{U}_f$, $\boldsymbol{U}_i$, $\boldsymbol{U}_c$ and $\boldsymbol{W}_f$, $\boldsymbol{W}_i$, $\boldsymbol{W}_c$ are the corresponding weights, and $\boldsymbol{b}_f$, $\boldsymbol{b}_i$, $\boldsymbol{b}_c$ are the biases. The cell states $\boldsymbol{c}_t$ maintain information about the previous time points, the forget gates control what or if the previous information will be discarded from the cell states, and the input gates control what new information will be stored in the cell states. Then, based on the cell state definition, the states of an LSTM unit are defined as follows:

$$\boldsymbol{h}_t = \boldsymbol{o}_t * tanh(\boldsymbol{c}_t) \qquad 5)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{U}_o\boldsymbol{h}_{t-1} + \boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{b}_o) \qquad 6)$$

where $\boldsymbol{o}_t$ are the output of gate activities. After the output gate, the final outputs $\boldsymbol{h}_t$ could be derived. As a result, after a *sigmoid* function, which decides what parts of the cell state will be outputted, the cell state through *tanh* (pushing the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, the output parts could be derived.

4.5 RAE for fMRI

Previous literature study has demonstrated the efficiency of RNN in modeling tfMRI data while preserving memories of previous state information and capturing the temporal dependences of input fMRI volumes (Wang et al., 2018). However, the fMRI data is weak supervised in nature, due to the common resting state fMRI not providing any frame-wise labels, letting alone the noise, inter-subject variations and intrinsic brain activities. Thus, the previous study in (Wang et al., 2018) is conceived to have two possible limitations when modeling fMRI data. First, RNN heavily relies on strong supervision; the lack of volume-wise labels would simply restrain RNN’s training and convergence. Second, since the RNN must be trained with labels, which are mostly task-related, the features learned are limited to task-related, possibly leaving out intrinsic component networks. Therefore, learning the intrinsic spatial-temporal structure in fMRI data without using explicitly provided labels is in great need and much desired, which partly motivated our work in this paper.

To overcome these limitations, we proposed a novel LSTM-based deep sparse recurrent auto-encoder (DSRAE), which is a deep unsupervised sequential neural network framework to model connectome-scale FBNs based on fMRI data. As shown in Figure.3, the proposed DSRAE is an eight-layer deep neural network model and it is composed of two parts: encoder and decoder. The encoder part encodes the input into high-level features and is comprised of the first four layers: one input layer, one fully connected layer, and two recurrent layers. The decoder part reconstructs the input and is comprised of the last four layers: two recurrent layers, one fully connected layer, and one output layer. Specifically, the node number decreases from 59,421 (the number of voxels for each volume) to 128 in the fully connected layer, then down to 64 in the first recurrent layer, and then down to 32 in the second recurrent layer. For the decoding component, there is a reverse way for node numbers, eventually increasing to 59,421 (the length of input) by layers.

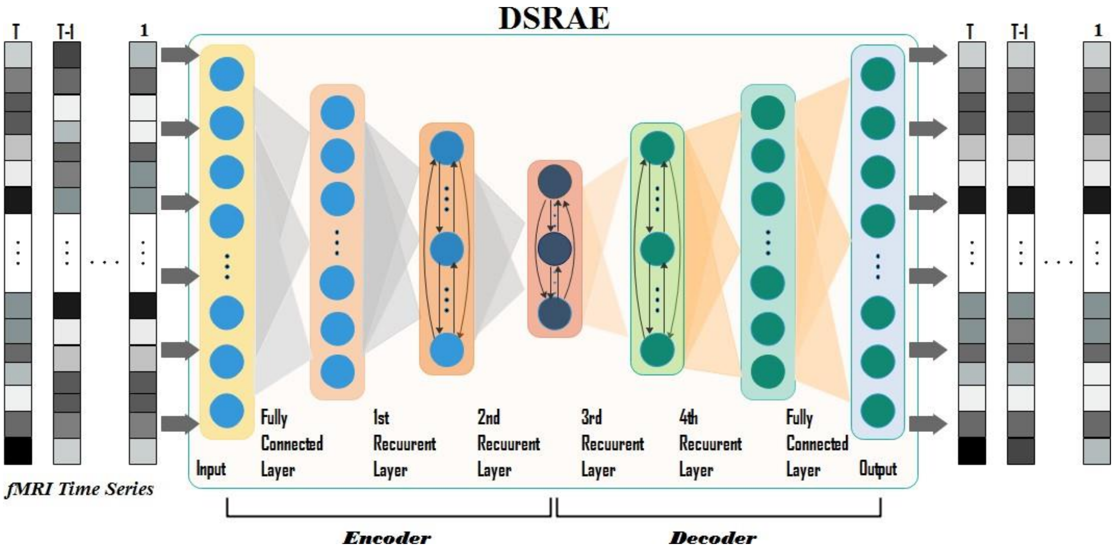


Figure. 4.3. Architecture of DSRAE. The DSRAE contains eight layers, in which the first four layers are encoding layers, and the last four layers are decoding layers. One input layer, one fully-connected layer, two recurrent layers included in encoding layers; two recurrent layers, one fully-connected layer, and one output layer correspondingly included in decoding layers.

To be more specific, for example, in language task, the DSRAE model received the 4D fMRI data as input, which represented the 3D volumes in the corresponding time series as shown in Figs.1a and b. Each 3D volume was transformed into a 1D vector (59,421×1). In this way, the whole input for one subject was a data matrix with dimension of 59,421×316. With the fully connected layer, the original large dimension recordings were converted into small size (128×316). After that, two recurrent layers were employed with the neuron unit size 64 and 32. With the Elastic Net(Zou & Hastie, 2005), 32 groups of activation networks were derived, as shown in Figure.1c. To illustrate the brain networks clearly, we regarded GLM-derived maps and task design curve as the spatial and temporal "ground truth", respectively. According to the Pearson correlation coefficients between the task design and the obtained temporal fluctuation, we found the most task-related spatial activated networks. To predict and decode the brain states during the tasks in time series, we calculated the Pearson correlation coefficients of the spatial activation networks and the original fMRI volume data as the representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008). After the decoder part (two recurrent layers increased the unit from 32 to 64 and then to 128, and one fully connected layer), the reconstructed layer (output layer) was derived. As shown in Figure.1d, the reconstruction errors of both the spatial and temporal dimensions were calculated to illustrate the matching relations between them.

The DSRAE models were trained on 791 subjects on tfMRI dataset of language task, working memory task and gambling task. For each task, we set the LSTM see-back step (the length of considered time point samples for predicting) as the length of the scan volumes, 316, 405 and 253, respectively. And all the other hyperparameters are same for all three tasks: the learning rate is set as 0.01, the epoch is set as 10, the batch size is set as 1 based on convergence. Notably, during the training of DSRAE, we use the L1-regulation (10e-7) and L2-regulation (10e-4) empirically at the same time to make our model more robust. In this paper, we implemented the DSRAE model with Keras, and ran on a GPU of Quadro K600.

4.6 Analysis of Reconstruction Error

We investigated the proposed DSRAE models on three tasks to assess the effectiveness and robustness across different tasks. The results have shown consistently good performance in terms of inferring and characterizing both task-evoked and spontaneous brain networks at the connectome scale, demonstrating that the DSRAE model is effective and robust.

To quantitatively evaluate and validate the signal reconstruction by DSRAE, we used one randomly selected subject in gambling task as an example to calculate the Pearson correlation coefficients between original tfMRI signals and reconstructed signals. The reconstruction result of DSRAE in spatial dimension is illustrated in Figure.4a, in which DSRAE has good reconstruction performance. That is, all the correlations are larger than 0.9. Specifically, we randomly selected 4 voxels to show the reconstruction error in temporal dimension, in which the reconstruction time series (blue curve) have remarkably high correlation with the original fMRI data time series (orange curve) (as shown in Figure.4b). We achieved the similar results and conclusion in other validation tasks (language task and working memory task), which are summarized in supplemental Figs.1 and 2. These high correlations between the original signals and reconstructed signals indicate that our DSRAE model can encode good spatial-temporal representations and reconstruct the input data.

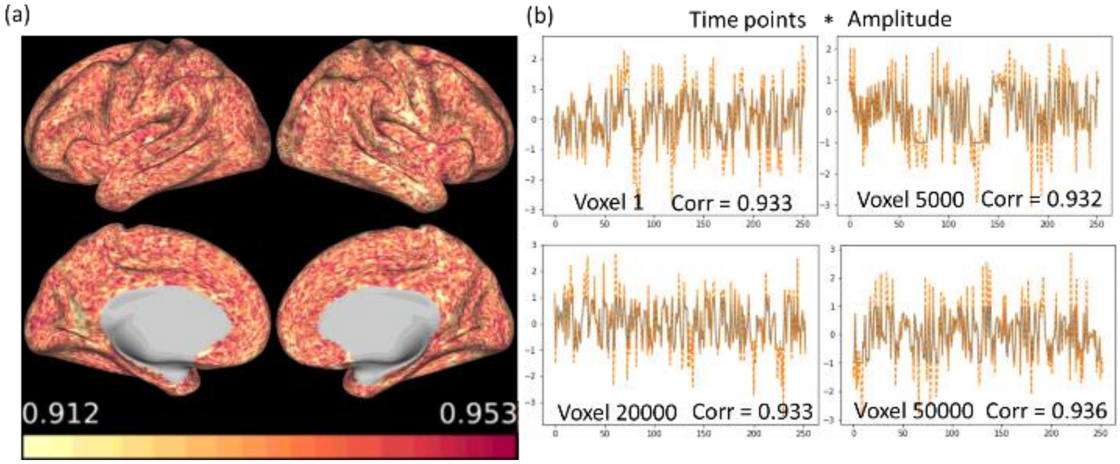


Fig. 4.4. Performance of DSRAE in reconstructing the gambling task fMRI signals. (a) The Pearson correlation coefficients between the reconstructed signals by DSRAE and the original signals in spatial dimension. (b) Temporal fluctuation comparison of the reconstructed voxels and original voxels. Blue curves are the reconstruction time series. Orange curves are the time series of original fMRI data.

4.7 Analysis of DSRAE-FBNs

To interpret the complex feature maps obtained by DSRAE, we treated the task paradigm via GLM as the spatial benchmark event block maps, which is a common practice in the fMRI field. For each task, we totally obtained 32 networks via DSRAE. Figs.5-7 show the estimated outputs activated by language task, and Figs.8-9 show the estimated outputs activated by working memory task. Figs.10-11 show the estimated outputs activated by gambling task and their comparison results.

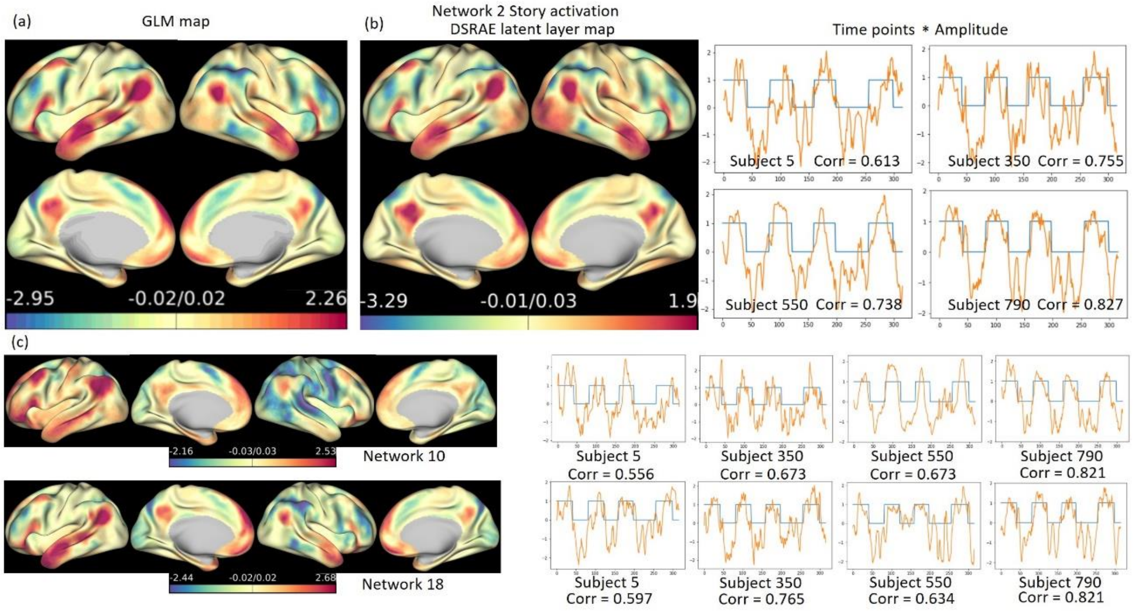


Fig. 4.5. Comparison between DSRAE and GLM-derived benchmark outputs of story stimulation specific maps in language task. (a) The spatial benchmark activation map predicated by GLM. (b) The spatial activation map predicted by DSRAE, the 2nd network of the total 32 networks. And the corresponding temporal fluctuation of network #2 and ground truth. Here, we showed the comparison results of four randomly selected subjects. Blue curves are ground truth. Orange curves are DSRAE temporal outputs. (c) Several other similar spatial maps of story stimulus activated maps and their temporal fluctuation.

As shown in Figure.5, the resulted spatial and temporal patterns predicted by DSRAE model are consistent with the GLM-derived benchmark patterns and task stimulus curve. Figure.5a is the brain network predicted with the benchmark event block map via GLM. Figure.5b is the estimated most similar output map with GLM by story stimulation, which is the 2nd network out of 32 networks. Comparing the networks derived by DSRAE with GLM, it is easy to recognize the close spatial match between them. The correlation coefficient between the maps via GLM and DSRAE is 0.931, which proves that the outstanding and effective performance of encoding ability of the DSRAE. The temporal fluctuation tendency of network #2 with the time series is also shown in Figure.5b, in which the predicted time series (orange curve) has a very high similarity (the highest correlation coefficient is 0.827 among the 4 randomly selected subjects) with the ground truth of story stimulation (blue curve). Although all the subjects we showed here have similar high correlation coefficients, the temporal time series vary among different subjects, which suggests the unique activations of different individuals. Since our DSRAE model is an unsupervised architecture, it can detect some other similar spatial maps during the activation of story stimulation (shown in Figure.5c). Based on the time series compared with ground truth and the corresponding correlation coefficients (the highest correlation coefficient is 0.821), network #10 and #18 are both story stimulation activated maps, which indicates that the brain states would vary during time sessions even under the same stimulation, and the dynamic fluctuations would not change in a transient way.

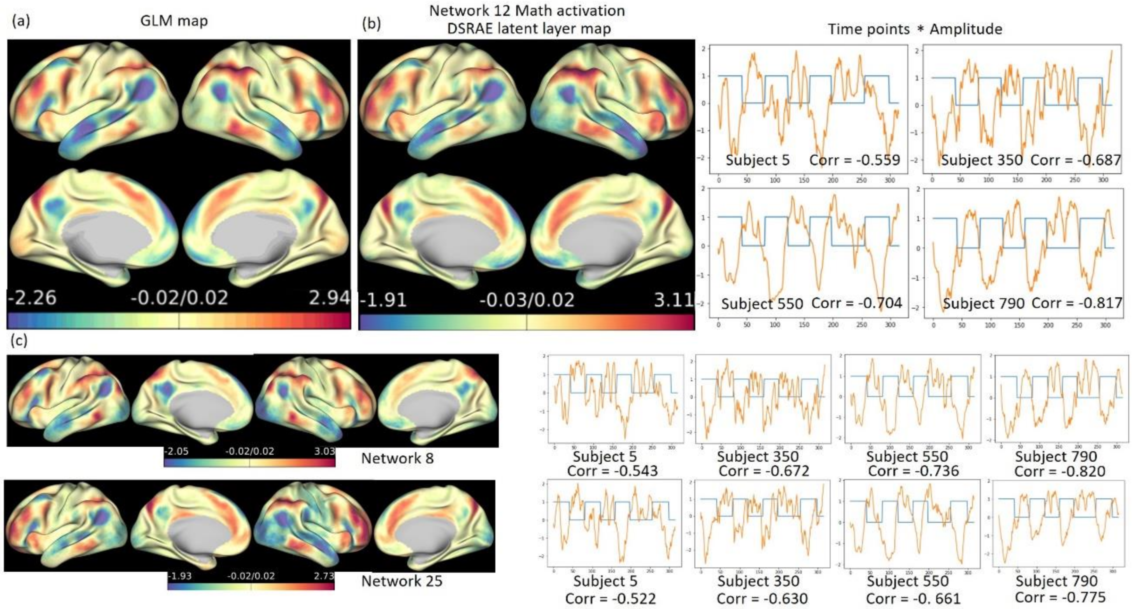


Fig. 4.6. The comparison between DSRAE and GLM-derived benchmark outputs o f math stimulation specific maps in language task. (a) The spatial benchmark activation map predicated by GLM. (b) The spatial activation map predicted by DSRAE, the 12th

network of the total 32 networks. And the corresponding temporal fluctuation of network #12 and ground truth. Here, we showed the comparison results of four randomly selected subjects. Blue curves are ground truth. Orange curves are DSRAE temporal outputs. (c) Several other similar spatial maps of story stimulus activated map and their temporal fluctuation.

Figure.6 shows the specific maps of math stimulation, which is the control experiment of story stimulation, the comparison with the GLM-derived benchmark map, and task curve ground truth. Though the math stimulation is similar in auditory and phonological input, the spatial activation maps are quite different from that of story stimulation (as shown in Figs.6a and b). The correlation coefficient between the maps predicted by DSRAE and GLM is 0.91, which proves the high correlation of the hidden layer of DSRAE with the spatial ground truth. In Figure.6b, there is a quite high negative correlation of network #12 (the math specific state) with the story stimulation (versus with math stimulation). Here, the greater the negative correlation coefficient is, the more relevant the network is with the math stimulation. For subject #790, the absolute value of the correlation coefficient can even reach up to 0.817, which indicates that the network predicted by DSRAE is quite meaningful. Besides, in Figure.6c, several similar networks also have good correlations with the task-designed ground truth, which can be seen as the math stimulus activated maps that varied in different time points.

Since the DSRAE models were trained in an unsupervised fashion, the features are not limited to task-related networks. In this way, the Dorsal Attention Network (DAN) can be detected via DSRAE model, which is a common distinct control network when subjects direct their attention (Fox, Corbetta, Snyder, Vincent, & Raichle, 2006; Giesbrecht, Woldorff, Song, & Mangun, 2003). Figure.7 shows the spatial map of activation of DAN and its time series, where the DAN is uncorrelated with the task stimulation (the absolution of correlation coefficient is less than 0.1). This result suggested that DAN should be a spontaneous network whenever the subjects are paying attention, and that would not be affected by the specific task.

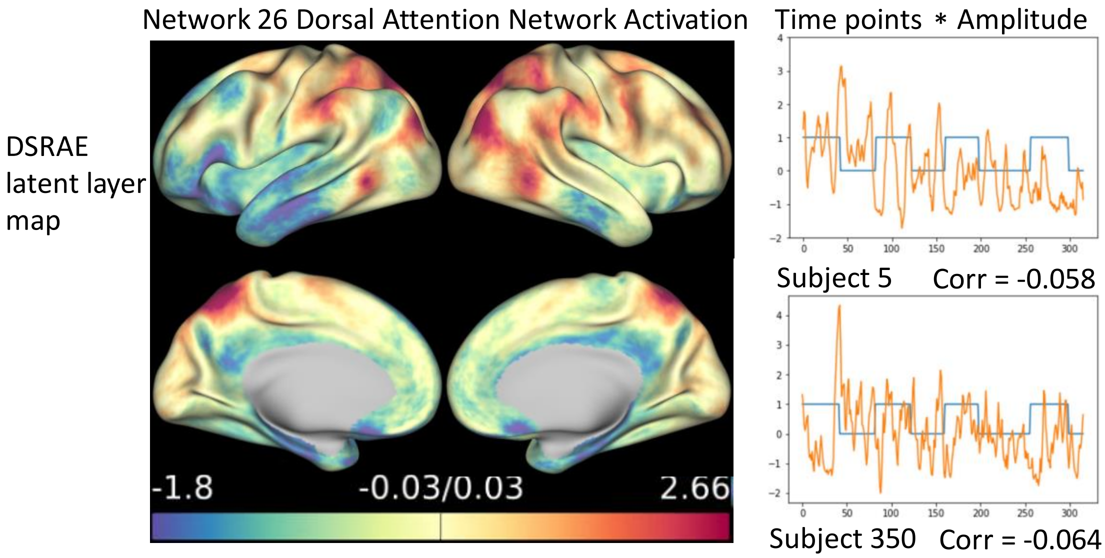


Fig.4.7. New map found by DSRAE in language task. The Dorsal Attention Network activation (network #26), and its temporal fluctuation. Here, we showed the comparison results of two randomly selected subjects. Blue curves are ground truth. Orange curves are DSRAE temporal outputs.

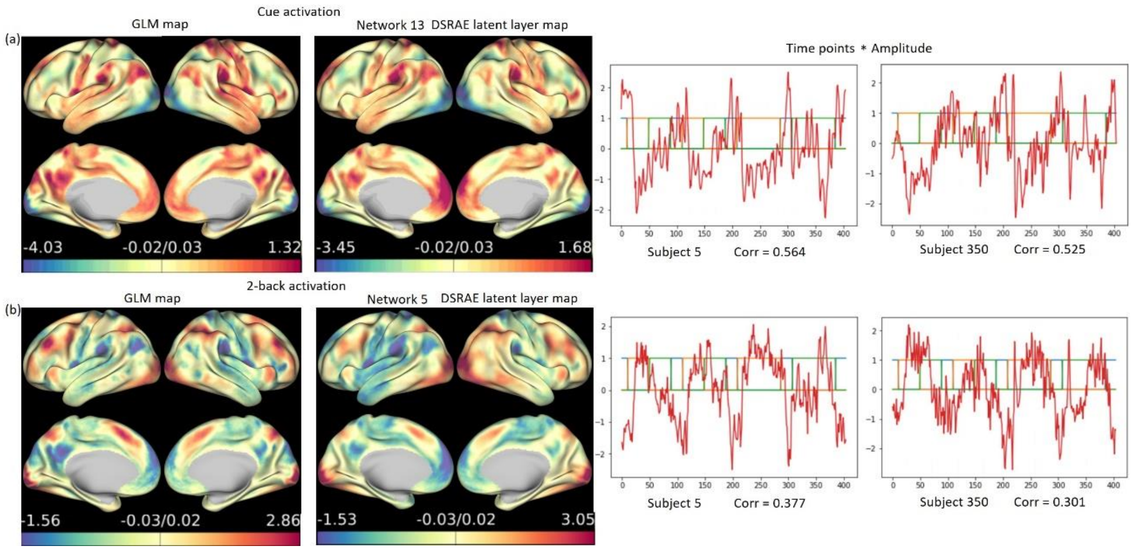


Fig. 4.8. Comparison between DSRAE and GLM-derived benchmark outputs of cue and 2-back stimulation specific maps in working memory task. (a) The spatial benchmark activation map predicated by GLM and DSRAE (the 13th network of the total 32 networks) of cue stimulation. And the corresponding temporal fluctuation of network #13 and ground truth. Here, we showed the comparison results of two randomly selected subjects. Blue curves are the cue-stimulus ground truth. Orange curves are 2-back stimulus ground truth. Green curves are 0-back ground truth. Red curves are DSRAE temporal outputs. (b) The spatial benchmark activation map predicated by GLM and DSRAE (the 5th network of the total 32 networks) of 2-back stimulation. The corresponding temporal fluctuation of network #5 and ground truth are shown here too.

To evaluate the robustness and efficiency of our DSRAE model, we also tested the architecture on working memory task and gambling fMRI datasets. The working memory task embedded two stimulations, including 0-back memory test, 2-back memory test, and cue period. First, as in shown in Figure.8a, the maps of the cue period relevant network derived by GLM and DSRAE have a high correlation coefficient of 0.936. No matter for subject #5 or #350, the correlation coefficients between network #13 and the cue-period ground truth are higher than 0.5, which proves the meaning of hidden layer in DSRAE. For the task-relevant maps, though the two kinds of working memory tests have very similar activation maps, we still obtain the 2-back task-specific spatial maps, as shown in Figure.8b. In which, the 2-back relevant spatial map via DSRAE (network #5) is similar to that benchmark derived from GLM with the correlation coefficient of 0.892. And the time series of network #5 is also positively correlated with that of 2-back test ground truth (correlation coefficient is 0.377). The 0-back test always has less memory loaded on the brain, and is hard to classify its activation patterns from 2-back stimulation (Archbold, Borghesani, Mahurin, Kapur, & Landis, 2009), so it is difficult to detect the specific network.

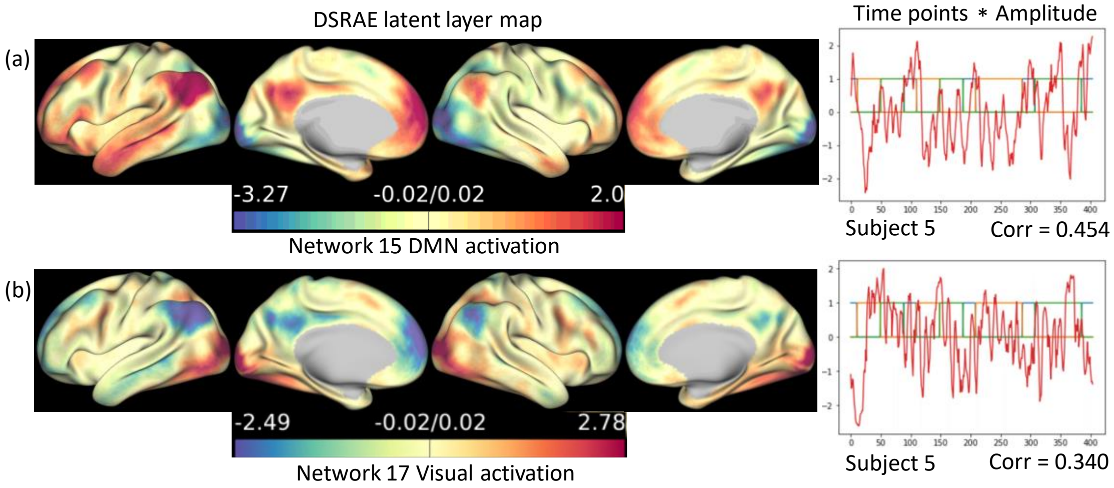


Fig. 4.9. New maps found by DSRAE in working memory task. (a) The Default Mode Network activation (network #15), and its temporal fluctuation. Here, we showed the comparison results of one randomly selected subject. Blue curves are the cue stimulus ground truth. Orange curves are 2-back stimulus ground truth. Green curves are 0-back ground truth. Red curves are DSRAE temporal outputs. (b) The Visual Network activation (network #17), and its temporal fluctuation.

For the spontaneous networks, the task-negative network of Default Mode Network (DMN) has been detected via DSRAE. Figure.9a shows the DMN (network #15 out of 32 networks) from DSRAE and that from GLM. Here, one subject's time series of network #15 is shown in Figure.9b as an example, where the time series have high positive correlation with the cue period (correlation coefficient is 0.454), i.e., anti-correlation with the working memory task positive activation (Fox et al., 2005; Prilipko et al., 2011). Interestingly, via the unsupervised DSRAE model, the Visual Network (network #17) has also been detected, as shown in Figure.9b. Because the working memory task needed the subjects to look at the screen to finish the tests during the task, the Visual Network could be activated (Woodman, Luck, & Schall, 2007). Moreover, network #17 has a positive correlation (correlation coefficient is 0.34) with the working memory tasks, which may indicate that the subjects paid stronger attention to the screen during the task performances.

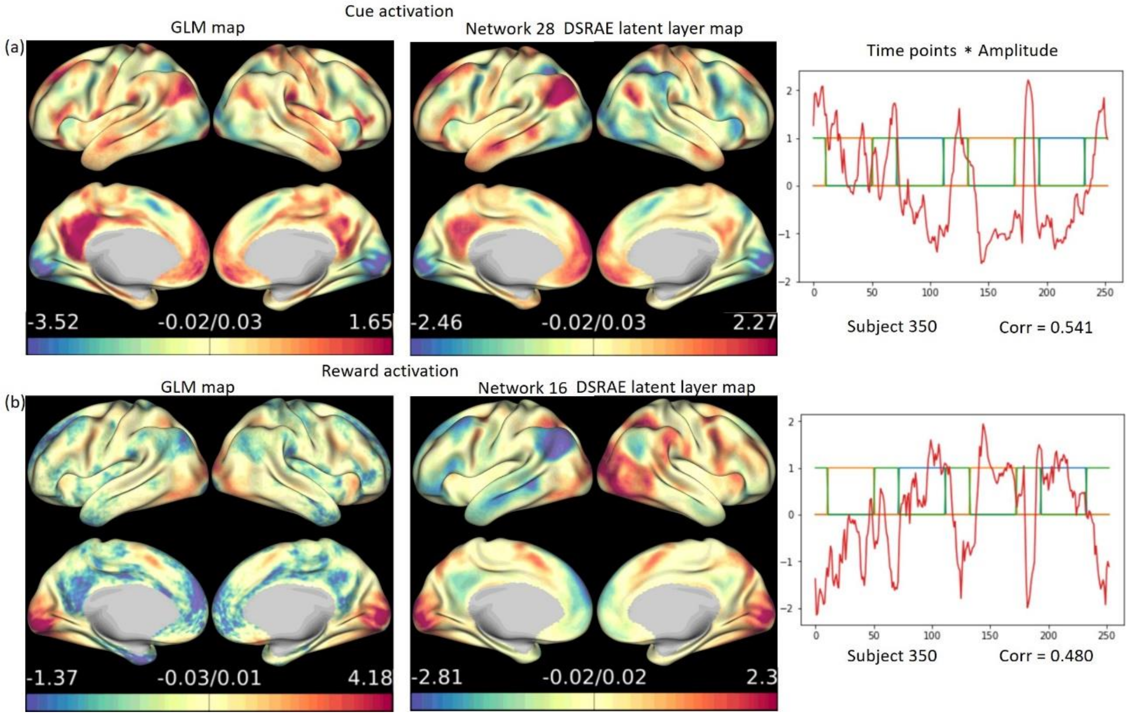


Fig. 4.10. Comparison between DSRAE and GLM-derived benchmark outputs of cue and reward stimulation specific maps in gambling task. (a) The spatial benchmark activation map predicated by GLM and DSRAE (the 28th network of the total 32 networks) of cue stimulation. And the corresponding temporal fluctuation of network #28 and ground truth. Here, we showed the comparison results of one randomly selected subjects. Blue curves are the reward-stimulus ground truth. Orange curves are loss-stimulus ground truth. Green curves are cue stimulus ground truth. Red curves are DSRAE temporal outputs. (b) The spatial benchmark activation map predicated by GLM and DSRAE (the 16th network of the total 32 networks) of reward stimulation. The corresponding temporal fluctuation of network #16 and ground truth are also shown here.

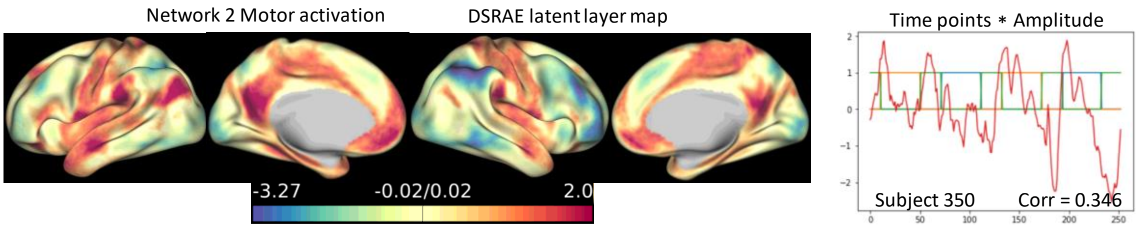


Fig. 4.11. New map found by DSRAE in gambling task. The Motor Network activation (network #2), and its temporal fluctuation. Here, we showed the comparison results of one randomly selected subject. Blue curves are the reward-stimulus ground truth. Orange curves are loss-stimulus ground truth. Green curves are cue stimulus ground truth. Red curves are DSRAE temporal outputs.

The gambling task dataset includes cue, reward, and loss stimulations. Figure.10a shows the cue stimulation activation (network #28), which is similar with the DMN, and the correlation coefficient of the activated spatial maps between the GLM and DSRAE is 0.848. Some studies found that the core regions of the DMN would be activated during rest and are dampened in response to task performance (task suppression) (Fox et al., 2005; Raichle et al., 2001). Here, the temporal fluctuation shows high positive correlation (correlation coefficient is 0.541) with the cue stimulus curve that consistent with the characteristics of DMN, which is negatively correlated with the task stimuli curve. The reward and loss stimulation activations have highly overlapped brain regions and are hard to be differentiated (C.-H. Lin, Chiu, Cheng, & Hsieh, 2008), so here we just demonstrated the main reward (6 reward trials pseudo randomly interleaved with either 1 neutral and 1 loss trial, 2 neutral trials, or 2 loss trials) stimulation specific activation map and the corresponding temporal fluctuation. As shown in Figure.10b, the reward stimulation activated map predicted by DSRAE (network #16) is similar with GLM-derived map with the correlation coefficient 0.66. For the temporal fluctuations, the network #16 has a high correlation coefficient of 0.48 with the reward-stimulation ground truth.

With our unsupervised DSRAE, the Motor Network activation has been detected during the gambling task (network #2 out of 32 networks), where striatal areas are involved in instrumental behaviors (Van Den Bos, Koot, & de Visser, 2014). As shown in Figure.11, the motor activation has positive correlation with the task stimulations (reward and loss are included) that are consistent with the task design requiring subjects to press button.

4.8 Analysis of Task Decoding

To verify the stability and robustness of our DSRAE model, we take the language task as an example to employ a repeated experiment and put other tasks' result into the supplemental materials (please see supplemental Figs.3 and 4). In order to eliminate the effect of the number and variation of subjects, a total of 791 subjects have been used in the original and repeated experiments. All the hyperparameters of the DSRAE framework are same for the both experiments. As shown in Figure.12, the networks yielded by the repeated experiments look very similar with those in the original experiment (Figs.5b and 6b). The correlation coefficients between the networks derived based on the repeated experiment and original experiment are 0.925 and 0.902 for story and math stimulation, respectively. Moreover, as the same as in the original experiment, no matter it is the story or the math stimulation specific network, both of their temporal fluctuations showed high correlation coefficients with the ground truth (the highest absolution correlation coefficients are 0.843 and 0.837). Given the lack of ground truth data of FBNs, these high robustness and reproducibility of our results validated the effectiveness of our DSRAE model.

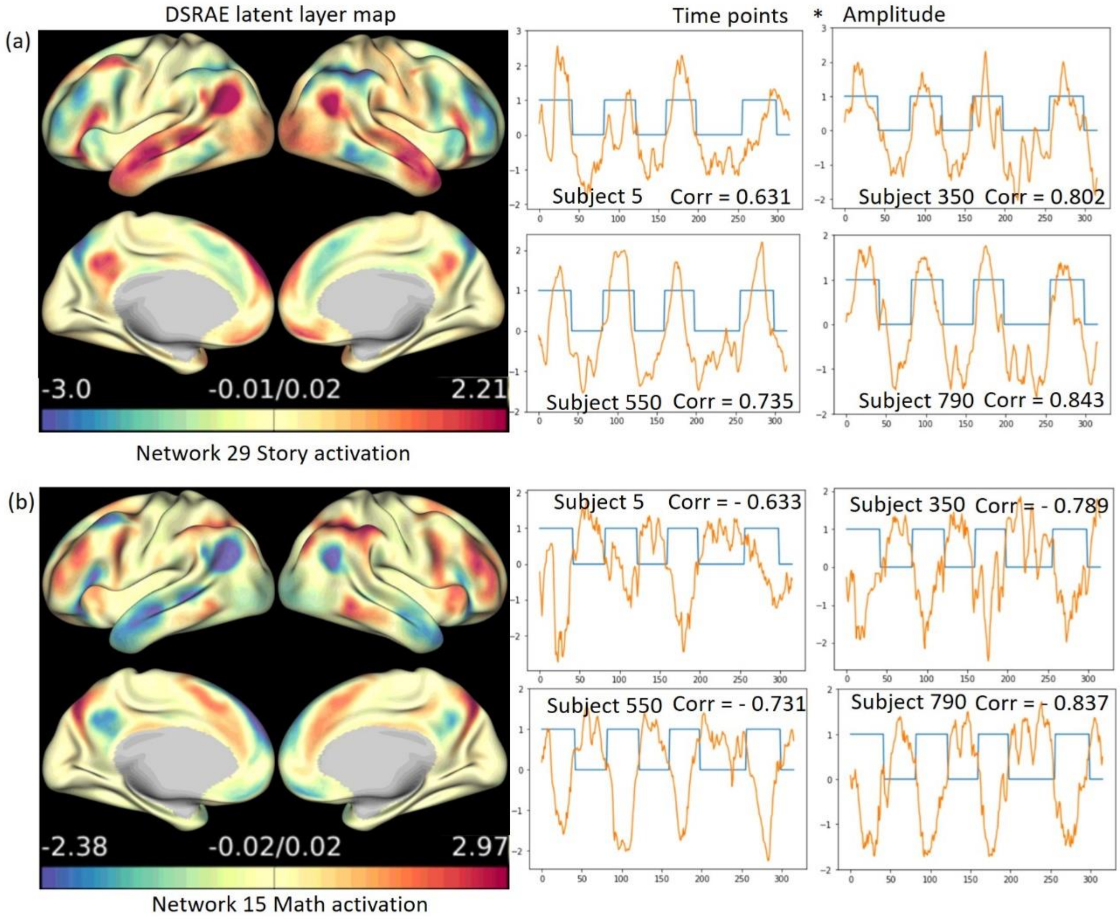


Fig. 4.12. The repeat experiment results in language task. (a) The story activation (network #29), and its temporal fluctuation. Here, we showed the comparison results of four randomly selected subjects. Blue curves are the stimulus ground truth. Orange curves are DSRAE temporal outputs. (b) The math activation (network #15), and its temporal fluctuation.

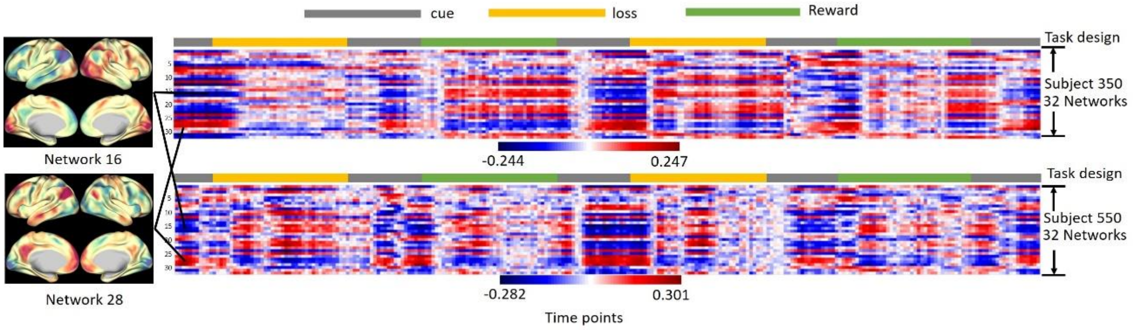


Fig. 4.13. The correlation matrix of the networks derived by DSRAE and the fMRI volumes of gambling task. Here, 2 randomly selected subjects are shown.

To examine what functional state the brain is in during the task period more specifically, we investigated the correlation of the input fMRI volumetric data and the learned feature networks via Pearson correlation, which is a kind of representational similarity analysis method (Dimsdale-Zucker & Ranganath, 2019) to measure how good the learned network map represents the original fMRI volume image (Y. Zhao, Dong, et al., 2018). In the correlation matrix, each row or column represents each network or original volume separately, which can show the relationship between each network and each original volume. In Figure.13, taking gambling task as an example, the network #16 and #28 have the negative or positive correlation with the cue stimulations, respectively. Especially, the network #28, which is the DMN, has high positive relevance with the cue period, which is an evidence that DMN could be suppressed during task period. Some volumes that have high positive correlation or negative correlation with some networks, which suggests that the brain is in such certain states at those time points. However, some volumes have low correlation with all networks, mainly during the task transition periods, which might indicate that the brain stays in multiple states at the same time. We achieved the similar results in other tasks, which are summarized in supplemental Figs.5 and 6. In general, our DSRAE model has revealed interesting patterns of spatial-temporal brain network dynamics that warrants future interpretation studies.

4.9 Discussion

In this work, we proposed a novel deep sparse recurrent auto-encoder (DSRAE) model to identify and characterize connectome-scale functional networks from spatial-temporal 4D fMRI data in an unsupervised way. We used the HCP language, working memory, and gambling tasks as our experiment testbed in this study and obtained promising results. To our best knowledge, the proposed DSRAE is among the early unified spatial-temporal deep learning models that can infer large-scale brain networks from 4D fMRI data directly in an unsupervised way. By visualizing and analyzing the feature maps of the brain responses, we confirmed that the feature maps are robust and meaningful. Besides, some networks have been identified and compared with the benchmark from classic GLM method, and the results confirmed the effectiveness of the DSRAE model. Meanwhile, the DSRAE model has also revealed the complex patterns of temporal brain network dynamics. For instance, DSRAE can detect some volumes that are highly correlated with one of the well-characterized networks. However, for some volumes with low correlation, there might be some more subtle states that need to be further decoded and interpreted in the future.

Though the current DSRAE model achieved promising performances in both spatial patterns and temporal dynamics understanding, it can still be enhanced and improved in a few aspects in the future. First, in this paper, although the length of the scan volumes for each task is set as the LSTM see-back step, it can be further revisited by introducing the attention mechanism to find out more task specific networks by the flexible see-back steps. The attention mechanism could make the neural networks choose which features they pay attention to (Xu et al., 2015), which could have the potential to use our DSRAE to map more hidden networks. Second, here we only used the simple and direct representational similarity analysis method to decode the brain states during different tasks. In our future work, we will conduct more interpretation into the feature networks learned by DSRAE to provide more in-depth neuroscientific insights into the DSRAE. Based on the previous improvement, the DSRAE model could be potentially applied as a real-time tool in the future to detect the brain behaviors in a shorter time, which can be potentially used as a brain states identifier.

In addition, there are several possible applications with our DSRAE model that can be considered in basic neuroscience and clinical research. For the neuroscience field, given that the brain activities and states are under dynamical changes, the DSRAE model could provide a useful tool to recognize brain states at fast time-scale using either task-based fMRI or resting fMRI data. For the clinical research field, dynamic performance could be sensitive to psychiatric or neurologic disorders, thus we could consider applying the DSRAE framework on resting fMRI datasets to potentially investigate the altered brain network states in brain disorders.

In general, our work contributes a novel deep sparse recurrent auto-encoder framework for spatial-temporal fMRI data modeling with future significant applications in cognitive and clinical neuroscience. Finally, the source codes of our DSRAE model and its associated sample datasets will be released at: https://github.com/ChloeLeeBnu/DSRAE.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This dissertation addressed methodological progress in the estimation of FBNs from fMRI data. From the perspective of machine learning, representation learning is interesting because it provides one way to perform unsupervised learning. Specifically, we can learn good representations for the unlabeled data, and then use these representations to solve the supervised learning task. The contribution of this work was three-fold since three major challenges of applying deep representation learning on fMRI volumes were conquered.

First, the enormous feature dimension. 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 was evident when the DNN with tens of thousands of input nodes (i.e., number of voxels; more than approximately 70,000 voxels within a whole brain with a 3-mm isotropic voxel size from a single fMRI volume) was

Second, the insufficient samples. A recurring theme in machine learning is the limit imposed by the lack of labelled datasets, which hampers training and task performance. Considering the enormous feature dimension of fMRI data, the dataset size is usually significantly smaller than the number of voxels per volume which may yield severe overfitting problem.

Third, the weak supervision. In medical image analysis, the lack of data is two-fold and more acute: there is general lack of publicly available data, and high-quality labelled data is even more scarce. The fMRI data are given with only coarse-grained labels or even no labels at all. What's more, due to the complexity of human brain activity, many intrinsic FBNs could be activated at the same time and they cannot be labeled.

This dissertation proposed VS-DBN, NAS-DBN, DSRAE models to identify and characterize connectome-scale functional networks from spatial-temporal 4D fMRI data in an unsupervised way. The HCP dataset was used as our experiment testbed in this study and obtained promising results. To our best knowledge, the proposed models are among the early unified spatial-temporal deep learning models that can infer large-scale brain networks from 4D fMRI data directly in an unsupervised way. By visualizing and analyzing the feature maps of the brain responses, we confirmed that the feature maps are robust and meaningful. Besides, some networks have been identified and compared with the benchmark from classic GLM method, and the results confirmed the effectiveness of the proposed models. Meanwhile, the proposed models have also revealed the complex patterns of temporal brain network dynamics. For instance, the proposed models can detect some volumes that are highly correlated with one of the well-characterized networks. However, for some volumes with low correlation, there might be some more subtle states that need to be further decoded and interpreted in the future.

In addition, there are several possible applications with the proposed models that can be considered in basic neuroscience and clinical research. For the neuroscience field, given that the brain activities and states are under dynamical changes, the proposed models could provide a useful tool to recognize brain states at fast time-scale using either task-based fMRI or resting fMRI data. For the clinical research field, dynamic performance could be sensitive to psychiatric or neurologic disorders, thus we could consider applying the proposed frameworks on resting fMRI datasets to potentially investigate the altered brain network states in brain disorders.

REFERENCE

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Devin, M. (2015). TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow. org. 2015. *URL https://www. tensorflow. org*.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., . . . Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics, 8*, 14.

Archbold, K. H., Borghesani, P. R., Mahurin, R. K., Kapur, V. K., & Landis, C. A. (2009). Neural activation patterns during working memory tasks and OSA disease severity: preliminary findings. *Journal of Clinical Sleep Medicine, 5*(01), 21-27.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., . . . Feldt, C. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage, 80*, 169-189.

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*(1457), 1001-1013.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage, 20*(2), 1052-1063.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning, 2*(1), 1-127.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., . . . Lowe, M. J. (2011). Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *NeuroImage, 54*(2), 1465-1475.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10*(3), 186.

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C., & Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage, 45*(3), 758-768.

Calhoun, V. D., & Adali, T. (2012). Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE reviews in biomedical engineering, 5,* 60-73.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping, 14*(3), 140-151.

Calhoun, V. D., Liu, J., & Adalı, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage, 45*(1), S163-S172.

Cho, K. H., Raiko, T., & Ilin, A. (2013). *Gaussian-bernoulli deep boltzmann machine.* Paper presented at the Neural Networks (IJCNN), The 2013 International Joint Conference on.

Cui, Y., Zhao, S., Wang, H., Xie, L., Chen, Y., Han, J., . . . Liu, T. (2018). *Identifying Brain Networks of Multiple Time Scales via Deep Recurrent Neural Network.* Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Derado, G., Bowman, F. D., & Kilts, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics, 66*(3), 949-957.

Dimsdale-Zucker, H. R., & Ranganath, C. (2019). Representational Similarity Analyses: A Practical Guide for Functional MRI Applications *Handbook of Behavioral Neuroscience* (Vol. 28, pp. 509-525): Elsevier.

Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron, 50*(5), 799-812.

Drobyshevsky, A., Baumann, S. B., & Schneider, W. (2006). A rapid fMRI task battery for mapping of visual, motor, cognitive, and emotional function. *NeuroImage, 31*(2), 732-744.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences, 14*(4), 172-179.

Ferrarini, L., Veer, I. M., Baerends, E., van Tol, M. J., Renken, R. J., van der Wee, N. J., . . . Penninx, B. W. (2009). Hierarchical functional modularity in the resting-state human brain. *Human brain mapping, 30*(7), 2220-2231.

Fischer, A., & Igel, C. (2012). *An introduction to restricted Boltzmann machines.* Paper presented at the Iberoamerican Congress on Pattern Recognition.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*(47), S102.

Forbes, E. E., Hariri, A. R., Martin, S. L., Silk, J. S., Moyles, D. L., Fisher, P. M., . . . Axelson, D. A. (2009). Altered striatal activation predicting real-world positive affect in adolescent major depressive disorder. *American Journal of Psychiatry, 166*(1), 64-73.

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences*, *103*(26), 10046-10051.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences, 102*(27), 9673-9678.

Friston, K. J. (1997). Transients, metastability, and neuronal dynamics. *NeuroImage, 5*(2), 164-171.

Ge, B., Makkie, M., Wang, J., Zhao, S., Jiang, X., Li, X., . . . Han, J. (2016). Signal sampling for efficient sparse representation of resting state FMRI data. *Brain imaging and behavior, 10*(4), 1206-1222.

Ge, F., Lv, J., Hu, X., Ge, B., Guo, L., Han, J., & Liu, T. (2015). *Deriving ADHD biomarkers with sparse coding based network analysis.* Paper presented at the Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on.

Ge, F., Lv, J., Hu, X., Guo, L., Han, J., Zhao, S., & Liu, T. (2018). *Exploring intrinsic networks and their interactions using group wise temporal sparse coding.* Paper presented at the Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on.

Giesbrecht, B., Woldorff, M., Song, A., & Mangun, G. (2003). Neural mechanisms of top-down control during spatial and feature attention. *NeuroImage, 19*(3), 496-512.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., . . . Polimeni, J. R. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage, 80,* 105-124.

Han Wang, K. X., Zhichao Lian, Yan Cui, Yaowu Chen, Jing Zhang, Li Xie, Joe Tsien, Tianming Liu. Large-scale Circuitry Interactions upon Earthquake Experiences Revealed by Recurrent Neural Networks. *IEEE Transactions on Neural Systems & Rehabilitation Engineering, in press*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation, 14*(8), 1771-1800.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.

Hjelm, R. D., Calhoun, V. D., Salakhutdinov, R., Allen, E. A., Adali, T., & Plis, S. M. (2014). Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. *NeuroImage, 96*, 245-260.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hu, X., Huang, H., Peng, B., Han, J., Liu, N., Lv, J., . . . Liu, T. (2018). Latent source mining in FMRI via restricted Boltzmann machine. *Human brain mapping, 39*(6), 2368-2380.

Hu, X., Lv, C., Cheng, G., Lv, J., Guo, L., Han, J., & Liu, T. (2015). Sparsity-constrained fMRI decoding of visual saliency in naturalistic video streams. *IEEE Transactions on Autonomous Mental Development, 7*(2), 65-75.

Huang, H., Hu, X., Dong, Q., Zhao, S., Zhang, S., Zhao, Y., . . . Liu, T. (2018). *Modeling task fMRI data via mixture of deep expert networks.* Paper presented at the Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., . . . Liu, T. (2018). Modeling task fMRI data via deep convolutional autoencoder. *IEEE transactions on medical imaging, 37*(7), 1551-1561.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1): Sinauer Associates Sunderland, MA.

Jiang, X., Li, X., Lv, J., Zhang, T., Zhang, S., Guo, L., & Liu, T. (2015). Sparse representation of HCP grayordinate data reveals novel functional architecture of cerebral cortex. *Human brain mapping, 36*(12), 5301-5319.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, 107*(25), 11163-11170.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). *Large-scale video classification with convolutional neural networks.* Paper presented at the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

Kennedy, J. (2010). Particle swarm optimization. *Encyclopedia of machine learning*, 760-766.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks, 8*(1), 98-113.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.* Paper presented at the Proceedings of the 26th annual international conference on machine learning.

Li, Y., Huang, H., Chen, H., & Liu, T. (2018). Deep Neural Networks for Exploration of Transcriptome of Adult Mouse Brain. *IEEE/ACM transactions on computational biology and bioinformatics*.

Lin, C.-H., Chiu, Y.-C., Cheng, C.-M., & Hsieh, J.-C. (2008). Brain maps of Iowa gambling task. *BMC neuroscience, 9*(1), 72.

Lin, L., Osan, R., & Tsien, J. Z. (2006). Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes. *TRENDS in Neurosciences, 29*(1), 48-57.

Liu, N., Han, J., Liu, T., & Li, X. (2018). Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE transactions on neural networks and learning systems, 29*(2), 392-404.

Liu, N., Han, J., Zhang, D., Wen, S., & Liu, T. (2015). *Predicting eye fixations using convolutional neural networks.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), 869.

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., . . . Huang, H. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Medical image analysis, 20*(1), 112-134.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., . . . Han, J. (2015). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Transactions on Biomedical Engineering, 62*(4), 1120-1131.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhao, S., Zhang, T., . . . Li, Z. (2015). Assessing effects of prenatal alcohol exposure using group-wise sparse representation of fMRI data. *Psychiatry Research: Neuroimaging, 233*(2), 254-268.

Lv, J., Lin, B., Li, Q., Zhang, W., Zhao, Y., Jiang, X., . . . Guo, C. (2017). Task fMRI data analysis based on supervised stochastic coordinate coding. *Medical image analysis, 38,* 1-16.

McKeown, M. J. (2000). Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage, 11*(1), 24-35.

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience, 4*, 200.

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K., & Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics, 3*, 37.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., . . . Hochberg, L. R. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 1.

Pei, W., & Tax, D. M. (2018). Unsupervised Learning of Sequence Representations by Autoencoders. *arXiv preprint arXiv:1804.00946.*

Pessoa, L. (2012). Beyond brain regions: Network perspective of cognition–emotion interactions. *Behavioral and Brain Sciences, 35*(3), 158-159.

Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of life reviews, 11*(3), 400-435.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., . . . Calhoun, V. D. (2014). Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience, 8*, 229.

Prilipko, O., Huynh, N., Schwartz, S., Tantrakul, V., Kim, J. H., Peralta, A. R., . . . Guilleminault, C. (2011). Task positive and default mode networks during a parametric working memory task in obstructive sleep apnea patients and healthy controls. *Sleep, 34*(3), 293-301.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676-682.

Ross, D. A., Lim, J., Lin, R.-S., & Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International journal of computer vision, 77*(1-3), 125-141.

Sak, H., Senior, A., & Beaufays, F. (2014). *Long short-term memory recurrent neural network architectures for large scale acoustic modeling.* Paper presented at the Fifteenth annual conference of the international speech communication association.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schmithorst, V. J., & Holland, S. K. (2004). Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 19*(3), 365-368.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673-2681.

Shen, Y., Mayhew, S. D., Kourtzi, Z., & Tiňo, P. (2014). Spatial–temporal modelling of fMRI data through spatially regularized mixture of hidden process models. *NeuroImage, 84*, 657-671.

Shimony, J. S., Zhang, D., Johnston, J. M., Fox, M. D., Roy, A., & Leuthardt, E. C. (2009). Resting-state spontaneous fluctuations in brain activity: a new paradigm for presurgical planning using fMRI. *Academic radiology, 16*(5), 578-583.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., . . . Laird, A. R. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences, 106*(31), 13040-13045.

Smith, S. M., Miller, K. L., Moeller, S., Xu, J., Auerbach, E. J., Woolrich, M. W., . . . Glasser, M. F. (2012). Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences, 109*(8), 3131-3136.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). *Unsupervised learning of video representations using lstms.* Paper presented at the International conference on machine learning.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., & Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage, 129*, 292-307.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41*(2), 281-292.

Van Den Bos, R., Koot, S., & de Visser, L. (2014). A rodent version of the Iowa Gambling Task: 7 years of progress. *Frontiers in psychology, 5,* 203.

Wang, H., Zhao, S., Dong, Q., Cui, Y., Chen, Y., Han, J., . . . Liu, T. (2018). Recognizing Brain States Using Deep Sparse Recurrent Neural Network. *IEEE transactions on medical imaging*.

Woodman, G. F., Luck, S. J., & Schall, J. D. (2007). The role of working memory representations in the control of attention. *Cerebral Cortex, 17*(suppl_1), i118-i124.

Woolrich, M. W., Jenkinson, M., Brady, J. M., & Smith, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE transactions on medical imaging, 23*(2), 213-231.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., . . . Bengio, Y. (2015). *Show, attend and tell: Neural image caption generation with visual attention.* Paper presented at the International conference on machine learning.

Yamashita, T., Tanaka, M., Yoshida, E., Yamauchi, Y., & Fujiyoshii, H. (2014). *To be bernoulli or to be gaussian, for a restricted boltzmann machine.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience, 19*(3), 356.

Zhang, S., Li, X., Lv, J., Jiang, X., Guo, L., & Liu, T. (2016). Characterizing and differentiating task-based and resting state fMRI signals via two-stage sparse representations. *Brain imaging and behavior, 10*(1), 21-32.

Zhang, W., Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, S., . . . Hu, D. (2018). Experimental Comparisons of Sparse Dictionary Learning and Independent Component Analysis for Brain Network Inference from fMRI Data. *IEEE Transactions on Biomedical Engineering*.

Zhang, W., Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, S., . . . Hu, D. (2019). Experimental comparisons of sparse dictionary learning and independent component analysis for brain network inference from fMRI data. *IEEE Transactions on Biomedical Engineering, 66*(1), 289-299.

Zhao, S., Han, J., Lv, J., Jiang, X., Hu, X., Zhao, Y., . . . Liu, T. (2015). Supervised dictionary learning for inferring concurrent brain networks. *IEEE transactions on medical imaging, 34*(10), 2036-2045.

Zhao, Y., Chen, H., Li, Y., Lv, J., Jiang, X., Ge, F., . . . Lyu, C. (2016). Connectome-scale group-wise consistent resting-state network analysis in autism spectrum disorder. *NeuroImage: Clinical, 12*, 23-33.

Zhao, Y., Dong, Q., Chen, H., Iraji, A., Li, Y., Makkie, M., . . . Liu, T. (2017). Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder. *Medical image analysis, 42*, 200-211.

Zhao, Y., Dong, Q., Zhang, S., Zhang, W., Chen, H., Jiang, X., . . . Liu, T. (2018). Automatic Recognition of fMRI-Derived Functional Networks Using 3-D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering, 65*(9), 1975-1984.

Zhao, Y., Ge, F., & Liu, T. (2018). Automatic recognition of holistic functional brain networks using iteratively optimized convolutional neural networks (IO-CNN) with weak label initialization. *Medical image analysis, 47*, 111-126.

Zhao, Y., Ge, F., Zhang, S., & Liu, T. (2018). *3D Deep Convolutional Neural Network Revealed the Value of Brain Network Overlap in Differentiating Autism Spectrum Disorder from Healthy Controls.* Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Zhao, Y., Li, X., Zhang, W., Zhao, S., Makkie, M., Zhang, M., . . . Liu, T. (2018). Modeling 4D fMRI Data via Spatio-Temporal Convolutional Neural Networks (ST-CNN). *arXiv preprint arXiv:1805.12564*.

Zhou, C., Zemanová, L., Zamora, G., Hilgetag, C. C., & Kurths, J. (2006). Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical review letters, 97*(23), 238103.

Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67*(2), 301-320.