Cerebral Cortex, 2020;00: 1–11

doi: 10.1093/cercor/bhaa292
Original Article

ORIGINAL ARTICLE

DS-GCNs: Connectome Classification using Dynamic Spectral Graph Convolution Networks with Assistant Task Training

Xiaodan Xing[1,2], Qingfeng Li[3], Mengya Yuan[3], Hao Wei[1,4], Zhong Xue[1], Tao Wang[3], Feng Shi[1] and Dinggang Shen[1,5,6]

[1]United Imaging Intelligence Co., Ltd., Shanghai 201210, China, [2]Shanghai Advanced Research Institute, Shanghai 201210, China, [3]Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 201108, China, [4]School of Computer Science and Engineering, Central South University, Hunan 410083, China, [5]School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China and [6]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea

Address correspondence to Tao Wang. Email: wtshhwy@163.com; Feng Shi. Email: feng.shi@united-imaging.com; Dinggang Shen. Email: dinggang.Shen@gmail.com

Xiaodan Xing and Qingfeng Li have contributed equally to this work.

Abstract

Functional connectivity (FC) matrices measure the regional interactions in the brain and have been widely used in neurological brain disease classification. A brain network, also named as connectome, could form a graph structure naturally, the nodes of which are brain regions and the edges are interregional connectivity. Thus, in this study, we proposed novel graph convolutional networks (GCNs) to extract efficient disease-related features from FC matrices. Considering the time-dependent nature of brain activity, we computed dynamic FC matrices with sliding windows and implemented a graph convolution–based LSTM (long short–term memory) layer to process dynamic graphs. Moreover, the demographics of patients were also used as additional outputs to guide the classification. In this paper, we proposed to utilize the demographic information as extra outputs and to share parameters among three networks predicting subject status, gender, and age, which serve as assistant tasks. We tested the performance of the proposed architecture in ADNI II dataset to classify Alzheimer's disease patients from normal controls. The classification accuracy, sensitivity, and specificity reach 90.0%, 91.7%, and 88.6%, respectively, on ADNI II dataset.

Key words: Connectome, fMRI, GCN

Introduction

Neurological diseases, such as Alzheimer's disease (AD) and major depression disease (MDD), cause abnormalities in brain functioning and affects patients' daily lives. Functional MRI, which evaluates brain activity by measuring the blood oxygenation level-dependent (BOLD) over time, is thus a perfect tool to investigate possible brain functional changes in many neurological disorders. Considering the nature of functional integration and segregation in the brain, researchers assess the correlations among neuronal activities in order to analyze brain function. Resting state functional MR images are first segmented into several brain regions of interest (ROIs) according to a brain atlas, and the correlation between each pair of brain ROIs could be computed and summarized in a matrix, known as functional connectivity matrix.

Note that unlike natural images that contain shape and texture information, the spatial locality of the entries of FC does not directly correspond to the locality of brain networks.

Thus, a difficult but important challenge in FC analysis is to extract efficient disease-related features. Reshaping FCs into vectors of features (Plis et al. 2014; Sen et al. 2016) and then sending it into off-the-shelf classifiers is a widely used operation. However, a vectorized FC will lose spatial information by discarding the topological structure of the matrix. Other feature extraction methods include BrainNetCNN (Kawahara et al. 2017), which proposed a cross-shape convolutional kernel to process connectivity matrix, as well as matrix clustering (Rajpoot et al. 2015).

A brain network can be represented as a graph structure naturally. It contains a set of brain regions of interest (ROIs), known as nodes, and describes their connectivity, known as edges. In the context of fMRI, the edges of brain functional graphs are derived from the correlations between each pair of brain ROIs. The nodes of brain functional graphs are the brain ROIs.

In brain network, graphical algorithms can be applied on FC data without loss of information. Graph kernel, which measures the inherent information in the graph structure (Vishwanathan et al. 2010; Shervashidze et al. 2011), is extensively used in many graph-based algorithms. Recently, researchers also applied graph kernel for neuro-imaging studies. For example, Jie et al. (Jie et al. 2014) used a graph kernel–based approach to measure directly the topological similarity between connectivity networks. However, the computational complexity of graph kernel is intensive.

Graph convolution neural networks (GCNs) allow an implementation of neural networks on graph structures. Neuroimaging pattern recognition applications include node classification and graph classification. Node classification assigns a predefined demographic graph for all subjects accompanied by a set of features. In the work of Parisot (Parisot et al. 2018), the feature of every individual, that is, every graph node, was a feature vector extracted from images, while the edges were calculated from the similarities between corresponding subjects. Graph classification treats each individual as an independent graph. For example, Ktena et al. (Ktena et al. 2018) proposed a Siamese graph convolution network to learn the similarity between brain networks. However, their works neglected the dynamic nature of brain activity, which can possibly improve the performance of classifiers.

Demographic information has been proved to be useful in many clinical studies. Researchers have reported a gender difference on cognitive functioning in brain (Halpern 2012). Besides, the incidence of many neurological diseases correlates with the years of age (Braak and Braak 1997; Butwicka and Gmitrowicz 2010). Thus, gender and age information was used in many neurological disease classification studies. However, in small datasets, which is common in medical imaging scenario, we argue that the status of subjects and demographic information of subjects are not always strongly correlated because of sampling preference. For example, a major guideline of collecting data for diagnosis is to balance the demographic distribution in both patient group and healthy control group, making gender and age weakly correlates with diagnosis. In this case, involving gender and age directly in the prediction model may not help models to learn diagnosis better.

In our previous work presented at a conference (Xing et al. 2019), a novel graph convolution recurrent network for fMRI pattern recognition was proposed. Graphs are defined with time-varying edges, that is, the dynamic functional connections, and fixed nodes, which are characterized by the structural information retrieved from average fMRI. The proposed method can improve diagnosis accuracy on fMRI through three aspects: 1) a graph classification algorithm which is dedicated for functional connectivity data; 2) an LSTM architecture that could further extract temporal information relevant to diagnosis; and 3) multitask settings that utilize demographic information as extra outputs for guiding the extraction of effective features. Details could be found in the following sections.

Materials and Method

We introduce below the proposed dynamic spectral GCNs with assistant task training using dynamic functional connectivity matrices, as in Figure 1. **First**, the dynamic connectivity matrices based on correlations of BOLD signals from sliding windows are defined as time-varying edges. Each node of the connectivity graph reflects an anatomical ROI. The feature on each node is defined by the volume of the corresponding ROI. **Second**, after defining the graph structure, a spectral graph convolution–based LSTM network is employed to extract information from the dynamic connectivity graphs. **Finally**, we make use of the demographic information as extra outputs, by adding two assistant networks with similar structure but different parameters, predicting gender and age. The feature maps from assistant networks were then weighted and combined with feature maps from the main network, guiding the parameter training and finally optimizing the results.

Graph Construction

A graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ was defined by two matrices, that is, node feature matrix and adjacency matrix. Node feature matrix denoted as$X \in R^{M\times N}$, where M is the number of nodes in the graph and N denotes the number of features, describes the property of every node in the graph. Node feature matrix describes the unique role of every node in the graph. Node feature matrix is required for graph convolutional operations and can be used as a check for graph isomorphism. If there are no node features provided, node feature matrix could be simply defined by identity matrix, where the numerical order of all nodes is binary encoded. Adjacency matrix $A \in R^{M\times M}$ represents the graph structure in a matrix form.

As we do not have predefined node features for brain ROIs from fMRI data, we first choose the identity matrix as node feature matrix. To further differentiate the brain regions, we replaced the diagonal entry of identity matrix by the corresponding volume of each ROI. Thus, the feature matrix in our method is then denoted as$\{X|X \in R^{M\times M}, x_{ij} = \begin{Bmatrix} v_i, i = j \\ 0, i \neq j \end{Bmatrix}\}$, where v_i is the brain volume of i-th ROI. Dynamic adjacency matrices are computed by a sliding window over the entire time series and can be represented as $\{A_t|A_t \in R^{M\times M}, t = 1, 2, \ldots, T\}$. Here T is the overall time points of dynamic adjacency matrices.

By aforementioned method, we could thus define a dynamic brain graph with static node features and dynamic connection pattern.

Graph Convolution LSTM

Graph Convolution

Conventional CNN performs spatial convolution on 2D or 3D images, which is reasonable because of the natural adjacent properties of neighboring pixels/voxels. However, when the input of neural network is a graph, such convolution operations

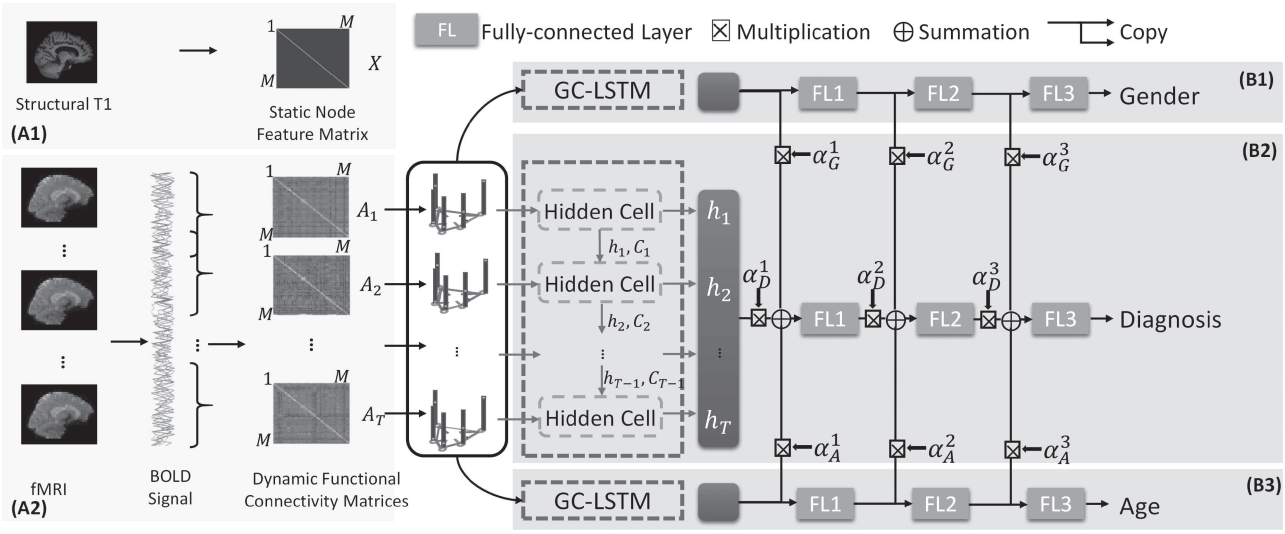

Figure 1. The schematic representation of the proposed DS-GCN with assistant task training for disease diagnosis using fMRI data. The static features of each node are defined by structural images (A1), and time-varying edges are defined by dynamic functional connectivity (A2). Dynamic graphs are then processed by diagnosis network (B2) and two subnetworks (B1 and B3) for assistance.

should be designed specifically. In our study, we chose spectral graph convolution (Defferrard et al. 2016; Kipf and Welling 2017), because of 1) the convenience of avoiding the matching of spatial local neighborhoods for a node and 2) the complete mathematical definition of spectral graph convolution. Consider a graph adjacency matrixA, the normalized graph Laplacian of A is

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}. \tag{1}$$

Here $D = \mathrm{diag}\left(\sum_j a_{i,j}\right)$ is the degree matrix of A. The eigenvectors $U$of graph Laplacian form the Fourier bases of the graph. Thus, the Fourier transform of feature maps is $\hat{x} = U^T x$. Since the convolution operation on spatial domain is the multiplication on spectral domain, the graph convolution operation is defined as

$$g_\theta * x = U\left(\left(U^T g_\theta\right) \odot \left(U^T x\right)\right), \tag{2}$$

where g_θ represents the learnable parameters of the graph convolutional kernel. In our study, we used Chebyshev polynomial to approximate $U^T g_\theta$. This approximation enables spectral convolution to be spatial localized and decreases the learning and computational complexity. Using Chebyshev polynomial, equation (2) can be approximated as

$$g_\theta * x = \sum_{j=0}^{K} \theta_j T_j\left(\tilde{L}\right) x. \tag{3}$$

Here, Chebyshev polynomial is $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$. $\tilde{L} = \frac{2}{\lambda_{\max}} L - I$ is the scaled graph Laplacian, $\lambda_{\max}$ denotes the largest eigenvalue of L, and I is the identity matrix. θ_j is one of the learnable parameters. And this expression is K localized because the K-th polynomial of the graph Laplacian only depends on maximum Kth nearest neighbor of the central node. For brain networks, every brain region densely connected with others, either weakly or strongly. Thus, we only need to consider the situation where $K = 1$. We further approximate $\lambda_{\max} \approx 2$, and thus equation (3) is simplified as

$$g_\theta * x = \theta_0 x + \theta_1 (L - I) x = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x \approx \theta'\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}\right) x, \tag{4}$$

with $\tilde{A} = A + I$ and$\tilde{D} = \mathrm{diag}\left(\sum_j \tilde{a}_{i,j}\right)$. With equation (4), for input signal $X \in R^{M \times F_{in}}$, with F_{in} input features for N nodes and F_{out} expected output features, we could define graph convolutional filter as follow:

$$Z = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta, \tag{5}$$

where $\Theta \in R^{F_{in} \times F_{out}}$ is a matrix of graph convolutional parameters and $Z \in R^{M \times F_{out}}$ is the output feature matrix.

GC-LSTM

Dynamic graphs require a recurrent structure to handle temporal information. Recurrent structure is composed of a series of identical components, known as hidden cells. The output of each hidden cell is known as hidden representations. Hidden cells are aligned one after the other. By receiving the hidden representations learned in the last cell and outputting representations for following cell, dynamic graphs are processed in order. Hidden representations from all hidden cells are then sent into fully connected layers for the diagnosis result.

In our paper, LSTM (long short–term memory) network is used, which could avoid the long-term dependency problem by recording the cell state in every time step. By replacing matrix multiplication in conventional LSTM with graph convolution, we could obtain below mathematical formulas, which graph convolutional LSTM follows:

$$\text{Forget gate}: f_t = \sigma_f\left(\omega_{xf} * x_t + \omega_{hf} * h_{t-1} + \omega_{Cf} \odot C_{t-1} + b_f\right)$$

$$\text{Input gate}: i_t = \sigma_i\left(\omega_{xi} * x_t + \omega_{hi} * h_{t-1} + \omega_{Ci} \odot C_{t-1} + b_i\right)$$

$$\text{Cell state}: C_t = f_t \odot C_{t-1} + i_t \odot \tan h\left(\omega_{xc} * x_t + \omega_{hc} * h_{t-1} + b_c\right)$$

$$\text{Output gate}: o_t = \sigma_o\left(\omega_{xo} * x_t + \omega_{ho} * h_{t-1} + \omega_{Co} \odot C_t + b_o\right)$$

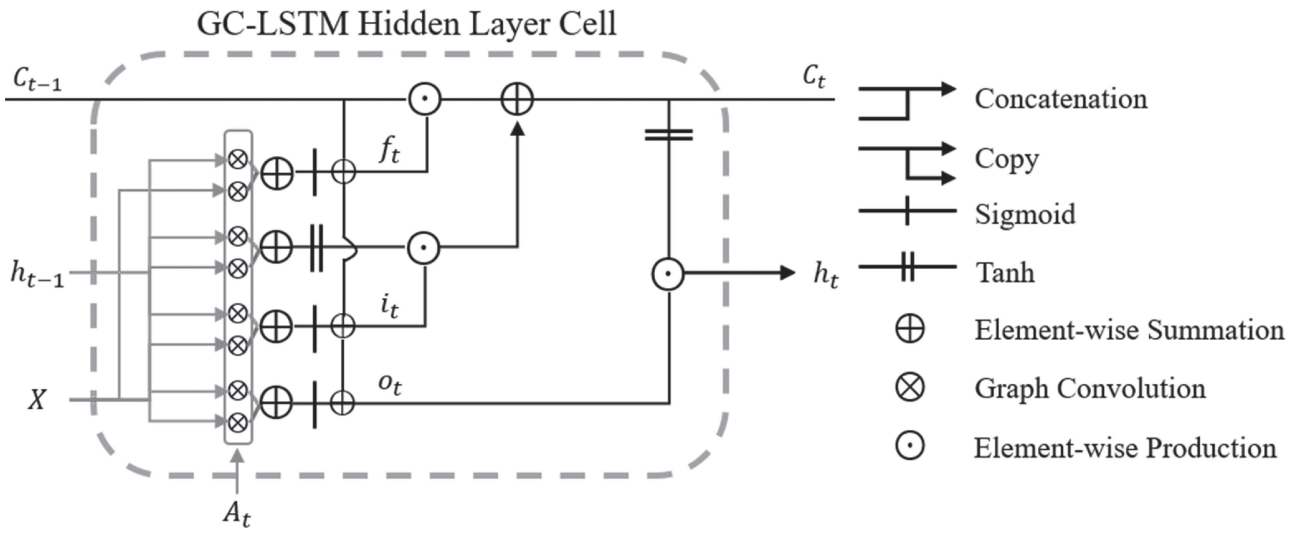

Figure 2. GC-LSTM hidden cell in detail. Each cell has two sources of input, the present (A_t and$X$) and the recent past (h_{t-1} andC_{t-1}). Output sequence $\{h_t|t = 1, 2, \ldots T\}$ was then treated as feature maps for following layers.

$$\text{Hidden state}: h_t = o \odot \tan h\,(C_t)\,. \tag{6}$$

Here, * denotes the graph convolution operator. Forget gate decides what information is going to be discarded from the cell state, while input gate decides what new information to store. Cell state is updated after forget gate and input gate, and output gate calculates the output, which is then filtered by cell state and send to next time step. Details of the GC-LSTM cell are shown in Figure 2.

Assistant Task Training

Gender and age provide important demographic information for disease prediction. Conventional deep learning classifiers added gender and age as additional features into the last fully connected layer. However, when adding demographic information as input features, the status and demographic information of subjects correlates in a statistical manner. For example, in ADNI datasets, gender and age are balanced among all groups, which makes these features correlate weakly with diagnosis labeling. In this situation, adding additional demographic features may not improve the classification performance. In addition, a balanced distribution in dataset does not imply a balanced distribution in real life: gender and age do affect the incidence of many neurological diseases (Hebert et al. 1995; Nebel et al. 2018). Thus, we propose to use demographic information as extra outputs in our networks. This strategy could not only improve the classification performance of weakly correlated demographic features but also guide the parameter optimizing in diagnosis task.

However, conventional multitask settings adopt a hard representation-sharing manner, by which different tasks share same parameters in beginning several convolutional layers of the network. In fact, it is difficult to decide which layers to be shared and which layers to be split among different tasks. Thus, our networks learn a linear combination of feature maps from different tasks to determine the shared representations by itself. We adopted this architecture from cross stitch networks (Misra et al. 2016), but our networks are task centralized, that is, only the main network, diagnosis network, receives the weighted combination of feature maps from assistant networks.

Considering the feature maps x_D^l from diagnosis network, x_G^l from gender prediction network, and x_A^l from age prediction network in layer$l$, the linear combination of these feature maps are learnt and sent into next layer in diagnosis network

$$\tilde{x}_D = \alpha_D^l x_D^l + \alpha_G^l x_G^l + \alpha_A^l x_A^l. \tag{7}$$

Here, α^l is a learnable parameter in layer$l$, which determines the contribution of demographic assistance. The loss for these networks is the weighted combination of losses from three tasks. For diagnosis and gender prediction, cross entropy loss was used, while for age prediction, mean square error loss was used.

$$\mathcal{L} = \omega_D \mathcal{L}_D + \omega_G \mathcal{L}_G + \omega_A \mathcal{L}_A. \tag{8}$$

Experiments

We have compared the proposed method with a number of state-of-the-art classification approaches to demonstrate the improvements brought by graph convolution LSTM and assistant task training.

Data

We evaluated the proposed method on public dataset ADNI (http://adni.loni.usc.edu/) (Alzheimer's Disease Neuroimaging Initiative). Under a series of criteria (http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI-2_Protocol.pdf), ADNI-2 dataset is split into 177 healthy controls (HC), 191 early MCI (eMCI) patients, 158 late MCI (lMCI) patients, and 115 AD patients with 1.5 T T1-weighted structural images and functional images (subjects had their eyes open). From ADNI-2 dataset, we randomly chose 329 for training, 138 for validating, and 174 for testing. The model that performs best on validation dataset is chosen as the final model for testing.

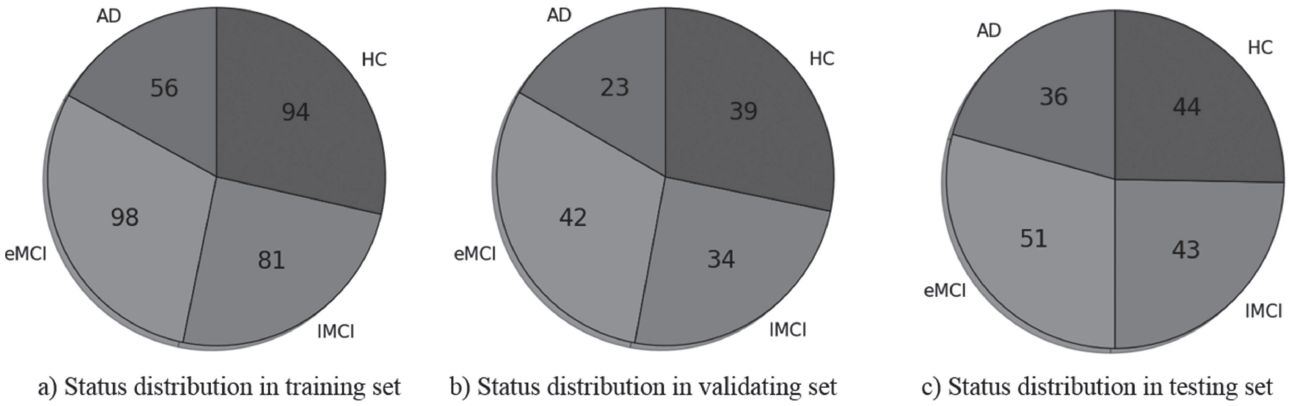

Figure 3. Demographic information of ADNI II dataset. The average age in training set is 73.6 years, average age in validating set is 73.7 years and 73.4 in testing set.

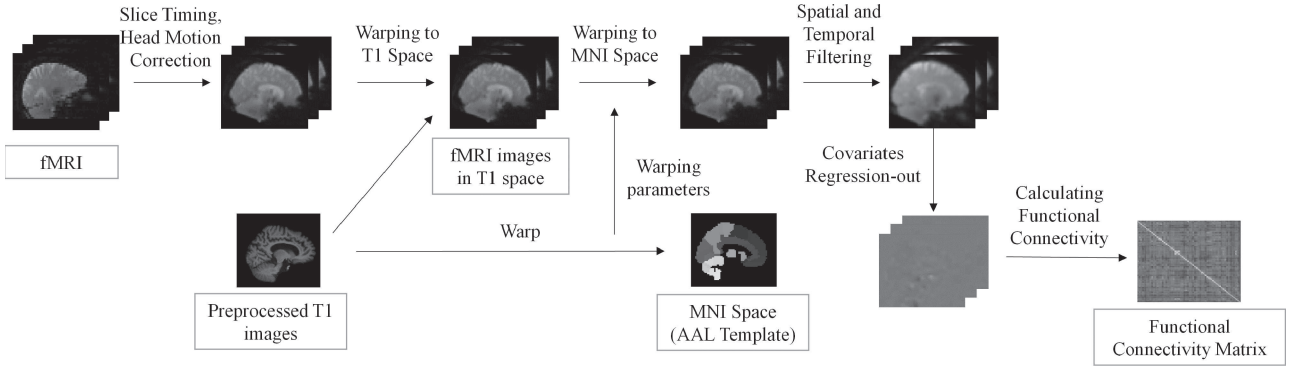

Figure 4. Preprocessing steps of structural and functional data. Preprocessions of T1 images include AC-PC correction, resampling, intensity correction, skull stripping, and registration. Preprocessions of fMRI include slice time correction, motion correction, registration, spatial and temporal filtering, and nuisance covariates regression.

Data distributions of training, validating, and testing datasets are shown in Figure 3.

Preprocessing

Preprocessions of both structural and functional images are under a standard pipeline. The preprocessing steps of data is shown in Figure 4. For structural images, we performed anterior commissure (AC)—posterior commissure (PC) correction and then resampled them into size $256 \times 256 \times 256$ with a resolution of$1 \times 1 \times 1$ mm^3. Then, after intensity inhomogeneity correction by N3 algorithm, structural images were skull stripped and registered into MNI space. Functional images were first slice time corrected by interpolation and motion corrected by a rigid body transformation on volumes, where mutual information was used as the cost function. Then, functional images were rigidly registered to the corresponding T1 MR images and further aligned to MNI space using the warping parameters of T1 MR to MNI space. Thus, functional images were demarcated into 116 ROIs by AAL (Automated Anatomical Labeling) template. Spatial filtering was applied with a Gaussian kernel with 4-mm FWHM (full width at half maximum). BOLD signals were then temporally filtered using a band-pass filter between 0.01 and 0.1 Hz. Dynamic functional connectivity matrices were computed by a sliding window after four nuisance covariates were regressed out, including head motion parameters, global mean signal, white matter signal, and cerebrospinal fluid (CSF) signal. Fisher-Z transformation was applied on all FC matrices.

Ablation Study on Dynamic Graphs

In order to demonstrate the improvements brought by graph convolution LSTM, we compared the proposed method with several baseline models. We included 5 comparison methods in total, which compared two types of inputs, static FCs and dynamic FCs, as detailed below as well as shown in Figure 5. To ensure a fair comparison, all methods were trained under super parameters (i.e., learning rate, batch size, etc.), which could optimize their performances on validation set. We employed these methods on two different tasks: 1) HC versus AD, and 2)HC versus eMCI.

Static FCs. 1) SVM: Static FCs and brain ROI volumes were reshaped as vectors of features and were then put into SVM with Gaussian kernel. It should be noted that only entries in the upper triangle static FC matrices were used to reduce feature dimension. 2) CNN: Static FCs and node feature matrices were treated as two channel images and put into VGG-16. 3) GCN: Static graphs were constructed by static FCs and node feature matrix and then put into a graph convolution network with one graph convolution layer and three fully connected layers.

Dynamic FCs. 4) LSTM: Dynamic FCs were reshaped as a time series of features and then concatenated with static vectorized node feature matrices at every time point. 5) DS-GCN: No demographic information is used in this method. Dynamic graphs were constructed from functional connectivity matrices and structural information. The proposed network has one GC-LSTM layer followed by three fully connected

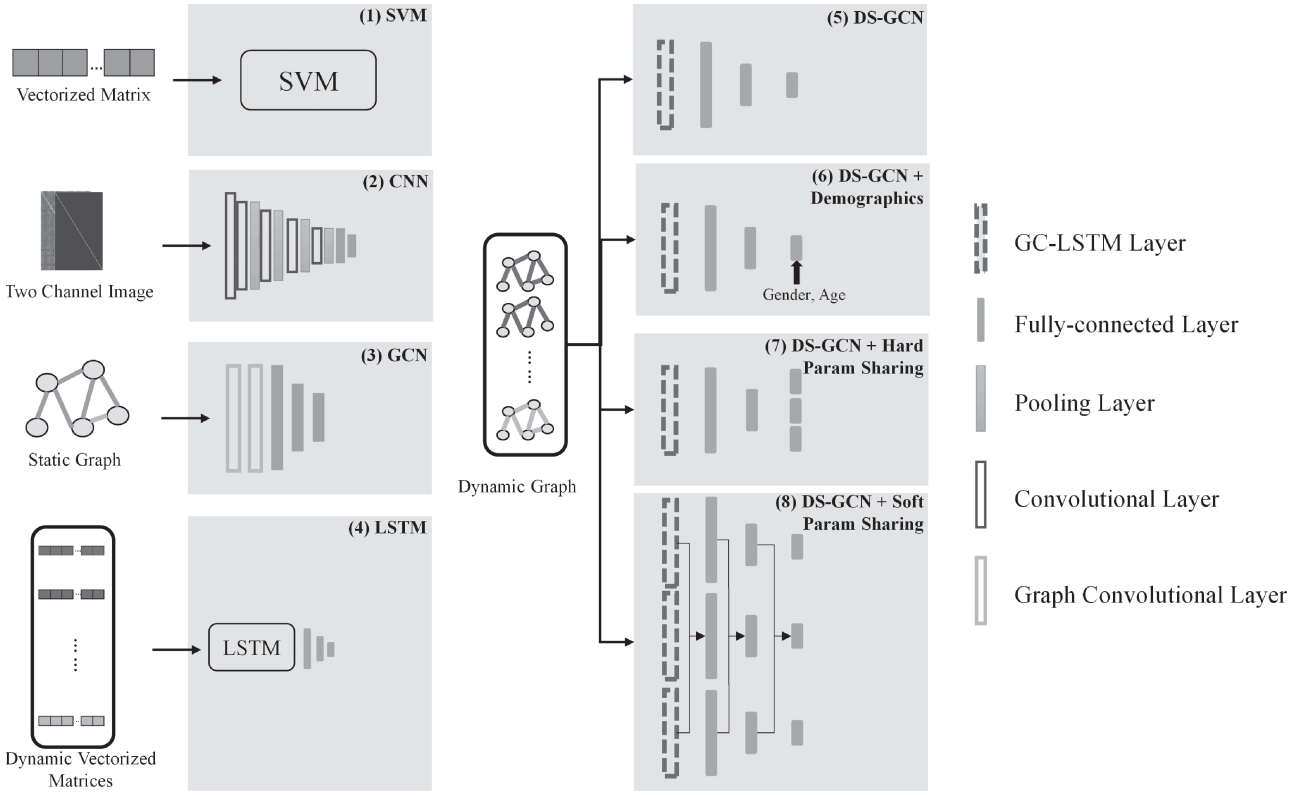

Figure 5. Schematic representations of all methods for ablation study. It should be noted that for every layer in our comparison, batch normalizations and *ReLU* activations were contained.

Table 1 Detailed network architecture of DS-GCN

Layer	Input size	Output size
GC-LSTM	39 × 116 × 116	39 × 116 × 1
FL1	39 × 116 × 1	116 × 1
FL2	116 × 1	64 × 1
FL3	64 × 1	2 × 1

layers. The detailed network layer dimensions are shown in Table 1.

Results. Classification results on ADNI II dataset are shown in Tables 2 and 3. From the tables, one can observe that overall graph-based networks performed better than no graph–based methods. In addition, dynamic connectivity outperformed static connectivity methods. The results indicate that dynamic connectivity with graph-based neural networks could fully exploit the information in fMRI connectivity analysis.

Ablation Study on Assistant Task Training

To demonstrate the efficacy of assistant task training, we employed the strategy on several deep learning–based models, including GCN, LSTM, and DS-GCN. The results can be seen in Table 4.

We compared assistant task learning strategy with other multitask settings. 1) Demographic input: In this method, demographic information, that is, gender and age, was used as extra inputs in the last fully connected layer. 2) Hard parameter sharing: We compared hard parameter sharing, where different tasks share the same parameters in front layers of the network with the proposed algorithm. Networks were split for three tasks at the last fully connected layer. 3) Soft parameter sharing: Soft parameter sharing is the proposed method and the feature maps from assistant networks were weighted and combined with feature maps from main network.

When we compared different settings of our algorithm, the performance by using assistant task training outperformed

Table 2 Classification results of different algorithms on ADNI II dataset. The classification task is to classify HC and AD

Inputs	Methods	Accuracy	Sensitivity	Specificity
Static FCs	SVM (linear kernel)	68.8%	72.2%	65.9%
	Spectral GCN	81.3%	88.9%	75.0%
	CNN	/	/	/
Dynamic FCs	LSTM	78.8%	86.1%	72.7%
	DS-GCN	83.8%	80.6%	86.4%

Table 3 Classification results of different algorithms on ADNI II dataset. The classification task is to classify HC and eMCI

Inputs	Methods	Accuracy	Sensitivity	Specificity
Static FCs	SVM (linear kernel)	60.0%	80.4%	38.6%
	GCN	70.5%	76.4%	63.6%
	CNN	/	/	/
Dynamic FCs	LSTM	67.3%	64.7%	70.5%
	DS-GCN	71.6%	68.6%	75.0%

Table 4 Classification results under different settings on ADNI II dataset. The classification task is to classify HC and AD

Models	Settings	Accuracy	Sensitivity	Specificity
GCN	Naive	81.3%	88.9%	75.0%
	Extra input	80.0%	83.3%	77.3%
	Hard parameter sharing	**83.8%**	83.3%	**84.1%**
	Soft parameter sharing	**83.8%**	**86.1%**	81.2%
LSTM	Naive	78.8%	**86.1%**	72.7%
	Extra input	80.0%	80.6%	79.5%
	Hard parameter sharing	81.3%	77.8%	84.1%
	Soft parameter sharing	**85.0%**	**80.6%**	**88.6%**
DS-GCN	Naive	83.8%	80.6%	86.4%
	Extra input	86.3%	88.9%	84.1%
	Hard parameter sharing	87.5%	**91.7%**	81.8%
	Soft parameter sharing	**90.0%**	**91.7%**	**88.6%**

Table 5 Classification results of multilabel classification tasks. The classification task is to classify HC, eMCI, lMCI, and AD

Class	Sensitivity (recall)	Precision	*F*1-score
HC	65.9%	60.4%	63.0%
eMCI	70.6%	65.4%	67.9%
lMCI	63.9%	52.3%	57.5%
AD	62.8%	100.0%	77.1%
Avg	65.8%	69.5%	67.6%

those without such a training strategy. Assistant task training may help improve the performance of classification.

Multilabel Classification Tasks

Despite of binary classification tasks, we also tested the proposed method under multilabel setting, as in Table 5. The outputs of the model are a 4-channel one-hot encoded labels representing HC, eMCI, lMCI, and AD, respectively. Recall (sensitivity), precision, and *F*1-score are the mostly used evaluation metrics for multilabel classification tasks. Recall is the fraction of the total amount of instances, which were correctly selected. Precision is the fraction of correctly selected cases of one class among all cases that were classified as this class. *F*1-score is the harmonic means of recall and precision.

Discussion

To further demonstrate the improvements in our model and to evaluate the assistant task training strategy, we compared the computational cost of all methods for comparison, computed the class activation map of our model, and plotted the contribution from assistant tasks during training. The proposed model consumes less storage and operations than other methods and located bilateral hippocampus, right precuneus, right frontal middle cortex, and left precentral cortex as class activated brain regions. Feature maps from gender and age prediction tasks have shown increasing contributions to diagnosis task through training.

Computational Cost

We compared the FLOPs (floating point operations) (Hunger 2005) and the physical size of aforementioned models. Comparing to the baseline models without using demographic information, the learnable parameters are tripled in our methods with soft parameter sharing. Besides, to handle dynamic FC, we implemented LSTM structure into our network, also causing a considerable increase of parameter amount.

However, as shown in Table 6, even with three times larger than baseline DS-GCN, the proposed method still consume less storage and operations than state-of-the-art methods, such as SVM, CNN, and LSTM. We attribute it to the computational simplicity of GCN. Considering an adjacency matrix$A^{116\times116}$ and a node feature matrix$X^{116\times116}$ and vectorizing both matrices as input, SVM requires at least 6670 (upper triangle entries in connectivity matrix) + 116 (diagonal features) parameters to assign a prediction. However, if we adopt graph convolution as described in equation (5), the dimensionality of parameters needed declines into 116×1. Graph convolution allows matrix

Table 6 Parameter sizes of different models. For ADNI-2 dataset, $T = 39$

Methods	Input resolution	Size/kiB	FLOPs/kMAC
SVM (linear kernel)	$6786 \times 1 \times 1$	18.4×10^2	13.2
GCN	$116 \times 116 \times 1$	64.4	22.2
CNN	$116 \times 116 \times 2$	12.2×10^3	14.5×10^6
LSTM	$13572 \times 1 \times T$	52.5×10^5	9.3×10^4
DS-GCN	$116 \times 116 \times T$	39.5	21.0
DS-GCN with demographics	$116 \times 116 \times T + 2$	39.5	21.0
DS-GCN with multi-task	$116 \times 116 \times T$	41.1	21.1
DS-GCN with assistant task training	$116 \times 116 \times T$	121.7	72.1

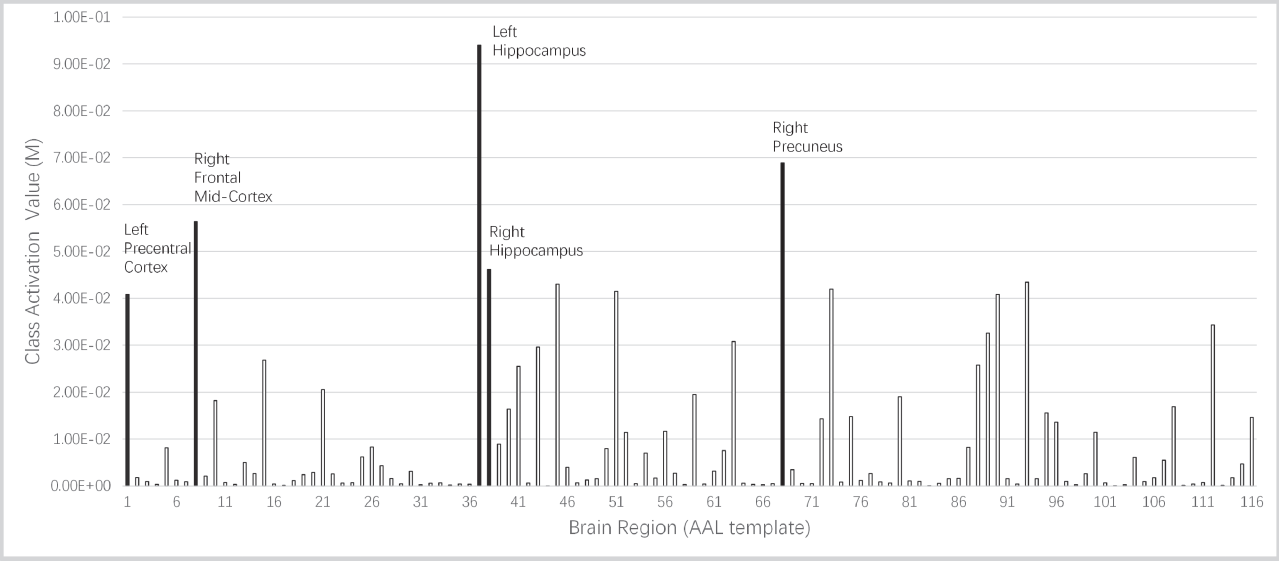

Figure 6. A bar chart showing the average class activation value of all test subjects. Top 5 activated regions are shown in different colors. Left hippocampus is reported as the most active region when our model diagnosis AD.

calculation on functional connectivity matrices and thus considerably reduce the number of parameters. It is also worth noting that in GCN model, we implemented two graph convolution layers to ensure a stable and acceptable performance, making GCN model more complex than DS-GCN.

Network Explainability

Despite of the superior performance of graph convolutional models, the explainability of the model is also helpful, because graph structures cannot be classified easily by human intuition. Recently, an increasing number of researches have been proposed to study the inner working of graph convolutional networks (Baldassarre and Azizpour 2019; Pope et al. 2019). Gradient-guided Class Activation Mapping (Grad-CAM) is a widely used algorithm to interpret the decision-making procedure of neural networks (Selvaraju et al. 2016). First, Grad-CAM calculates the gradient of y_c (probability for class c, in this case, the probability for AD) with respect to the feature maps x_D^1 output from GC-LSTM layer. Then, the importance weight $\omega_{c,k}$ of the kth node in this feature map is then

$$\omega_{c,k} = \frac{\partial y_c}{\partial x^1_{D,k}}.$$

The class activation map M is the feature map x_D^1 filtered by importance weight:

$M_c = ReLU(\omega_c x_D^1)$. Here, activation function ReLU makes heat map focus on brain regions, which only play positive parts in classification. The class activation value of all testing subjects is averaged and shown in Figures 6 and 7.

Analysis on Assistant Task Training

Synthetic Experiment

To illustrate the effectiveness of assistant task learning, we reproduced the result of a synthetic experiment from (Caruana and Sa 1997). Here, we used two fully connected layer with activation function to approximate$(A + B)^2$. A and$B$ are uniformly chosen from range $[-5, 5]$ and are encoded into 2^{10} bins. Besides the binary codes of A and$B$, another feature $(A - B)^2$ is also provided. However, $(A - B)^2$ weakly correlates with our target (the correlation reaches zero if A and$B$ are "random" enough). This does not mean we should discard$(A - B)^2$. By using $(A - B)^2$ as extra outputs, our simple network could easily generate the best prediction performance by learning to model subfeatures A and$B$. Figures 8 and 9 show the network design and the result of this synthetic experiment.

Task Contribution

The numeric values of αs as well as the values of losses during training are plotted in Figure 10. To present the stable

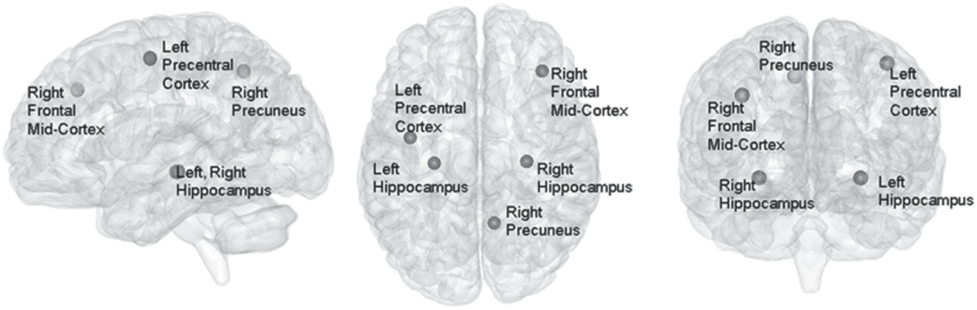

Figure 7. The Alzheimer's disease activation map shown on a smoothed MNI 152 template. This image is shown by BrainNetViewer (Xia et al. 2013).

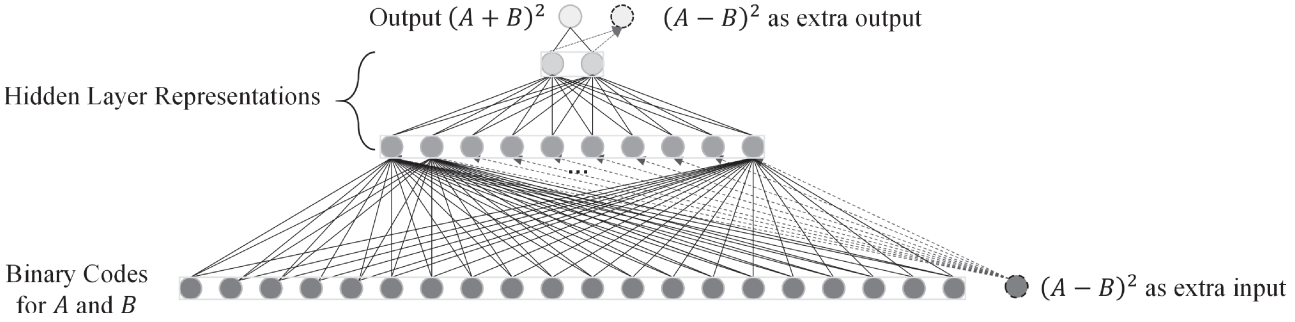

Figure 8. Network design for synthetic experiment to illustrate the effectiveness of assistant task learning. The classification model is composed of three fully connected layers. Number of node here represents the dimension of features.

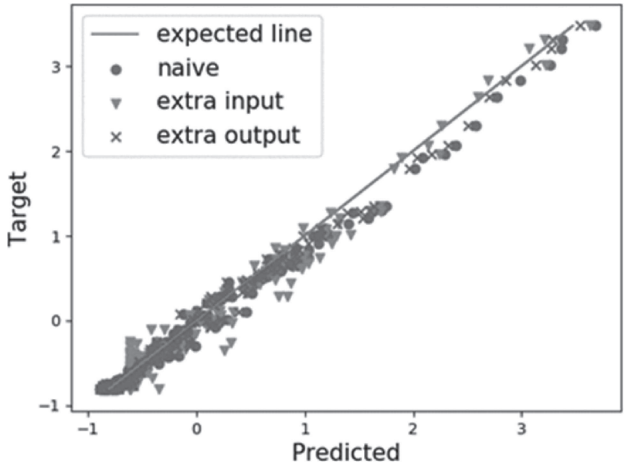

Figure 9. The result for synthetic experiments. Both $(A-B)^2$ and $(A+B)^2$ are normalized. The mean squared error for prediction without $(A-B)^2$ (naive) is 0.016. The mean squared error for prediction with $(A-B)^2$ as an extra input is 0.036 and is 0.009 for prediction with $(A-B)^2$ as an additional output.

contribution from different tasks, we compared the updating values of alphas under different initializations. Experiment results show that using the training part of ADNI dataset, from which 43 subjects (20% of the training dataset) was selected for validation. Training procedure in all experiments used Adam (Kingma and Ba 2015) with initial learning rate10^{-3}. For random initialization, the cross entropy loss of diagnosis and gender classification were between [0.6, 0.8], while the MSE loss of age prediction were between [0.07, 0.08]. Thus, to weight the losses from different tasks into the same scale, the weights of losses were then $\omega_D = \omega_G = 1$, and $\omega_A = 10$. As seen in equation (6), α_D refers to the weight of the main tasks, α_G refers to the contribution from gender, and α_A refers to the contribution from age. These values reflect the influences of each factor to the final classification performance.

The above results confirmed our hypothesis that the training of gender classification and age prediction may help generate stable and robust network feature maps for diagnosis task. Alphas in the last convolutional layers did not change as dramatically as the first two layers, indicating a weaker correlation among different tasks for deeper layers.

Conclusion

In this paper, we have proposed a model specially to deal with dynamic functional connectivity. The novelty of this paper can be shown in three aspects. First, we proposed a specially designed graph construction algorithm, which could utilize both functional and structural MR images; second, a spectral graph convolution–based recurrent network is implemented to extract both functional and spatial information; and last, our model adopts a training strategy, which utilizes demographic features as extra outputs, guiding the diagnosis network to train and focus. Our work provides not only a new and efficient way to analyze functional MR images but also a new perspective for the usage of semantic features.

However, there were still some limitations in our study. First, we used AAL atlas to demarcate the brain into 116 brain ROIs. Although there exist several brain atlases, such as automatic nonlinear imaging matching and anatomical labeling (ANIMAL) (Collins and Evans 1999), and Talariach Daemon (TT) (Lancaster et al. 2000), AAL atlas is most widely adopted in fMRI studies and also chosen as default template in SPM software. However,

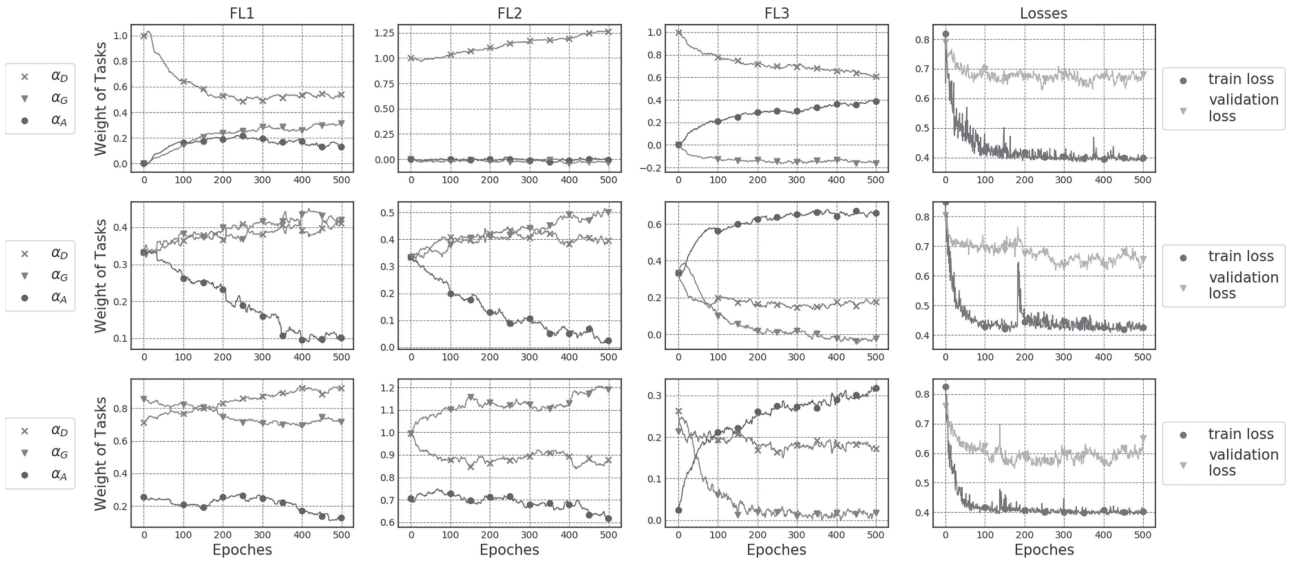

Figure 10. Alpha values of different layers under different initializations. Assistant task training shown in the first row was initialized with $\alpha_D = 1$ and$\alpha_G = \alpha_A = 0$, while in the second row was initialized with$\alpha_D = \alpha_G = \alpha_A = 0.33$. Figures of the third row show alpha values randomly initialized.

the choice of brain nodes could affect the result of functional connectivity networks. In this work, we proposed graph LSTM, which is a general method and could also be used in different brain atlas settings. Second, we only used volume as our structural node features because it is easily obtained and the volume is highly related with the onset of AD (Bartos et al. 2019; Zhao et al. 2019). However, we do not test the algorithms with more other structural features including cortex thickness and image intensity.

Data Availability

The dataset used in this work (ADNII) is an open-source dataset downloaded from http://adni.loni.usc.edu/.

Notes

This work was supported in part by the National Natural Science Foundation of China under Grant 81830056, Shanghai Health System Excellent Talent Training Program (Excellent Subject Leader) 2017BR054 and Shanghai Municipal Education Commission-Gaofeng Clinical Medicine Grant Support, 20172029. *Conflict of interest*: X.X, Q.L, H.W. are interns, and Z.X. and F.S. are employees of Shanghai United Imaging Intelligence Co., Ltd. The company has no role in designing and performing the surveillances and analyzing and interpreting the data. All other authors report no conflicts of interest relevant to this article.

References

Baldassarre F, Azizpour H. 2019. Explainability techniques for graph convolutional networks. arXiv preprint arXiv: 1905.13686.

Bartos A, Gregus D, Ibrahim I, Tintěra J. 2019. Brain volumes and their ratios in Alzheimer s disease on magnetic resonance imaging segmented using Freesurfer 6.0. *Psychiatry Res Neuroimaging*. 287:70–74.

Braak H, Braak E. 1997. Frequency of stages of Alzheimer-related lesions in different age categories. *Neurobiol Aging*. 18(4):351–357. doi: 10.1016/S0197-4580(97)00056-0.

Butwicka A, Gmitrowicz A. 2010. Symptom clusters in obsessive-compulsive disorder (OCD): influence of age and age of onset. *Eur Child Adolesc Psychiatry*. 19(4):365–370. doi: 10.1007/s00787-009-0055-2.

Caruana R, Sa VR. 1997. Promoting Poor Features to Supervisors: Some Inputs Work Better as Outputs. In: *Paper presented at the Advances in Neural Information Processing Systems*.

Collins D, Evans A. 1999. Animal: validation and applications of nonlinear registration-based segmentation. *International journal of pattern recognition and artificial intelligence*, 11(08): 1271–1294.

Defferrard M, Bresson X, Vandergheynst P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv Neural Inf Process Syst*. 29(Nips 2016): 29.

Halpern DF. 2012. *Sex differences in cognitive abilities*. 4th ed. New York: Psychology Press.

Hebert LE, Scherr PA, Beckett LA, Albert MS, Pilgrim DM, Chown MJ, Funkenstein HH, Evans DA. 1995. Age-specific incidence of Alzheimer's disease in a community population. *JAMA*. 273(17):1354.

Hunger R. 2005. *Floating point oeperations in matrix-vector calculus*. Munich University of Technology, Inst. for Circuit Theory and Signal.

Jie B, Zhang DQ, Wee CY, Shen DG. 2014. Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. *Hum Brain Mapp*. 35(7):2876–2897. doi: 10.1002/hbm.22353.

Kawahara J, Brown CJ, Miller SP, Booth BG, Chau V, Grunau RE, Zwicker JG, Hamarneh G. 2017. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. *Neuroimage*. 146:1038–1049.

Kingma DP, Ba JL. 2015. Adam: A Method for Stochastic Optimization. In: *Paper presented at the International Conference on Learning Representations*.

Kipf TN, Welling M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In: *Paper presented at the International Conference on Learning Representations*.

Ktena SI, Parisot S, Ferrante E, Rajchl M, Lee M, Glocker B, Rueckert D. 2018. Metric learning with spectral graph convolutions on brain connectivity networks. *Neuroimage*. 169:431–442. doi: 10.1016/j.neuroimage.2017.12.052.

Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT. 2000. Automated Talairach atlas labels for functional brain mapping. *Hum Brain Mapp*. 10(3):120–131.

Misra I, Shrivastava A, Gupta A, Hebert M. 2016. Cross-Stitch Networks for Multi-Task Learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (Cvpr)*, 3994–4003. doi: 10.1109/Cvpr.2016.433

Nebel RA, Aggarwal NT, Barnes LL, Gallagher A, Goldstein JM, Kantarci K, Mallampalli MP, Mormino EC, Scott L, Wai Haung Y, et al. 2018. Understanding the impact of sex and gender in Alzheimer's disease: a call to action. *Alzheimers Dement*. 14(9): S1552526018301304.

Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B, Rueckert D. 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med Image Anal*. 48:117–130. doi: 10.1016/j.media.2018.06.001.

Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, Johnson HJ, Paulsen JS, Turner JA, Calhoun VD. 2014. Deep learning for neuroimaging: a validation study. *Front Neurosci*. 8:229. doi: 10.3389/fnins.2014.00229.

Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. 2019. Explainability Methods for Graph Convolutional Neural Networks. In: *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Rajpoot K, Riaz A, Majeed W, Rajpoot N. 2015. Functional connectivity alterations in epilepsy from resting-state functional MRI. *Plos One*. 10(8):e0134944.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Paper presented at the 2017 IEEE International Conference on Computer Vision*.

Sen B, Bernstein GA, Xu TT, Mueller BA, Schreiner MW, Cullen KR, Parhi KK. 2016. Classification of Obsessive-Compulsive Disorder from Resting-State fMRI. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Embc)*, pp. 3606–3609.

Shervashidze N, Schweitzer P, van Leeuwen EJ, Mehlhorn K, Borgwardt KM. 2011. Weisfeiler-Lehman graph kernels. *J Mach Learn Res*. 12:2539–2561.

Vishwanathan SVN, Schraudolph NN, Kondor R, Borgwardt KM. 2010. Graph kernels. *J Mach Learn Res*. 11:1201–1242.

Xia M, Wang J, He Y. 2013. BrainNet viewer: a network visualization tool for human brain connectomics. *Plos One*. 8(7): e68910.

Xing X, Li Q, Wei H, Zhang M, Zhan Y, Zhou X, Xue J, Shi F. 2019. Dynamic Spectrl Graph Convolution Networks with Assistant Task Training For Early MCI Diagnosis. In: *Paper presented at the The Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Zhao W, Luo Y, Zhao L, Mok V, Su L, Yin C, Sun Y, Jie L, Shi L, Han Y. 2019. Automated brain MRI volumetry differentiates early stages of Alzheimer's disease from normal aging. *J Geriatr Psychiatry Neurol*. 32(6):354–364.