ARTICLE IN PRESS

YNIMG-12213; No. of pages: 20; 4C:

NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia ☆

Junghoe Kim [a], Vince D. Calhoun [b,d], Eunsoo Shim [c], Jong-Hwan Lee [a,*]

[a] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea
[b] Department of Electrical and Computer Engineering, University of New Mexico, NM, USA
[c] Samsung Advanced Institute of Technology, Samsung Electronics, Suwon, Republic of Korea
[d] The Mind Research Network & LBERI, NM, USA

ARTICLE INFO

Article history:
Received 21 October 2014
Accepted 7 May 2015
Available online xxxx

Keywords:
Deep learning
Functional connectivity
Resting-state functional magnetic resonance imaging
Schizophrenia
Sparsity
Stacked autoencoder

ABSTRACT

Functional connectivity (FC) patterns obtained from resting-state functional magnetic resonance imaging data are commonly employed to study neuropsychiatric conditions by using pattern classifiers such as the support vector machine (SVM). Meanwhile, a deep neural network (DNN) with multiple hidden layers has shown its ability to systematically extract lower-to-higher level information of image and speech data from lower-to-higher hidden layers, markedly enhancing classification accuracy. The objective of this study was to adopt the DNN for whole-brain resting-state FC pattern classification of schizophrenia (SZ) patients vs. healthy controls (HCs) and identification of aberrant FC patterns associated with SZ. We hypothesized that the lower-to-higher level features learned via the DNN would significantly enhance the classification accuracy, and proposed an adaptive learning algorithm to explicitly control the weight sparsity in each hidden layer via L_1-norm regularization. Furthermore, the weights were initialized via stacked autoencoder based pre-training to further improve the classification performance. Classification accuracy was systematically evaluated as a function of (1) the number of hidden layers/nodes, (2) the use of L_1-norm regularization, (3) the use of the pre-training, (4) the use of framewise displacement (FD) removal, and (5) the use of anatomical/functional parcellation. Using FC patterns from anatomically parcellated regions without FD removal, an error rate of 14.2% was achieved by employing three hidden layers and 50 hidden nodes with both L_1-norm regularization and pre-training, which was substantially lower than the error rate from the SVM (22.3%). Moreover, the trained DNN weights (i.e., the learned features) were found to represent the hierarchical organization of aberrant FC patterns in SZ compared with HC. Specifically, pairs of nodes extracted from the lower hidden layer represented sparse FC patterns implicated in SZ, which was quantified by using kurtosis/modularity measures and features from the higher hidden layer showed holistic/global FC patterns differentiating SZ from HC. Our proposed schemes and reported findings attained by using the DNN classifier and whole-brain FC data suggest that such approaches show improved ability to learn hidden patterns in brain imaging data, which may be useful for developing diagnostic tools for SZ and other neuropsychiatric disorders and identifying associated aberrant FC patterns.

Abbreviations: AAL, automated anatomical labeling; AE, autoencoder; ANOVA, analysis of variance; AUC, area under the receiver operating characteristic curve; BOLD, blood-oxygenation-level-dependent; CC, correlation coefficient; CSF, cerebrospinal fluid; CV, cross-validation; DARTEL, Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra; DMN, default-mode network; DNN, deep neural network; DSM, Diagnostic and Statistical Manual of Mental Disorders; EPI, echo-planar imaging; FC, functional connectivity; FNC, functional network connectivity; GICA, group independent component analysis; GM, gray matter; HC, healthy control; ICA, independent component analysis; IC, independent component; ICC, intra-class correlation coefficient; LASSO, least absolute shrinkage and selection operator; MNI, Montreal Neurological Institute; MNIST, Mixed National Institute of Standards and Technology; MSE, mean squared error; NITRC, Neuroimaging Informatics Tools and Resources Clearinghouse; PANSS, Positive and Negative Syndrome Scale; PCA, principal component analysis; PCC, posterior cingulate cortex; RBF, radial basis function; ROIs, regions of interest; rsfMRI, resting-state functional magnetic resonance imaging; SAE, stacked autoencoder; SCID, Structured Clinical Interview for DSM Disorders; SD, standard deviation; SP, spatial pattern; SPM, statistical parametric mapping; TC, time course; SVM, support vector machine; SZ, schizophrenia; TS, time series; WM, white matter.

☆ Sources of support: This work was supported by the Global Research Network (GRN) program funded by the National Research Foundation (NRF) of Korea (NRF-2013S1A2A2035364), in part by the BK21 plus program of the NRF of Korea, in part by a grant from the Korean Health Technology R&D Project, Ministry of Health & Welfare, Korea (HI12C1847), and in part by National Institutes of Health (NIH) grant R01EB005846. These sponsors had no involvement in the study design, data collection, analysis or interpretation of data, manuscript preparation, or the decision to submit for publication.

* Corresponding author at: Department of Brain and Cognitive Engineering, Korea University, Anam-dong 5ga, Seongbuk-gu, Seoul 136-713, Republic of Korea. Fax: +82 2 3290 3667.
E-mail address: jonghwan_lee@korea.ac.kr (J.-H. Lee).

http://dx.doi.org/10.1016/j.neuroimage.2015.05.018
1053-8119/© 2015 Published by Elsevier Inc.

Introduction

Resting-state functional MRI (rsfMRI) without a task paradigm has been successfully employed to exploit neuronal underpinnings implicated in neuropsychiatric disorders (Anand et al., 2005; Castellanos et al., 2008; Li et al., 2002), including schizophrenia (SZ) (Greicius, 2008; Jafri et al., 2008; Liang et al., 2006; Liu et al., 2008; Mingoia et al., 2012; Yu et al., 2012; Zhou et al., 2007). For example, Liu et al. (2008) presented evidence of significantly altered functional connectivity (FC) pairs (i.e., locally connected networks), or disrupted “small-world FC networks,” in the prefrontal, parietal, and temporal areas of the brain in patients with SZ. In that study, the hypothesis on dysfunctional network integration in SZ was supported by the lower strength of the FC in the pairs of nodes and decreased synchronization of functionally connected brain regions as well as the longer absolute path to reach global functional networks (Bullmore et al., 1997, 1998; Calhoun et al., 2009; Friston and Frith, 1995; Liu et al., 2008). In addition, Liang et al. (2006) reported that aberrant SZ-associated FC patterns were widely distributed throughout the entire brain (i.e., the FC levels of approximately 89% of the observed pairs of nodes were decreased), as opposed to showing a restricted pattern within only a few specific brain regions.

Machine-learning algorithms have been successfully deployed in the automated classification of altered FC patterns related to SZ (Arbabshirani et al., 2013; Du et al., 2012; Shen et al., 2010; Tang et al., 2012; Watanabe et al., 2014). In this regard, Du et al. (2012) developed a method combining kernel principal component analysis (PCA) and group independent component analysis (ICA) aimed at the computer-aided diagnosis of SZ, achieving 98% accuracy by using fMRI data acquired from an auditory oddball task paradigm. In addition, Shen et al. (2010) introduced an unsupervised learning-based classifier to discriminate SZ patients from HC subjects by applying a combination of nonlinear dimensionality reduction and self-organized clustering algorithms to rsfMRI data. The results of this analysis demonstrated the highest discriminating power for FC patterns between the cerebellum and the frontal cortex, with a classification accuracy of 92.3%. In addition, the altered resting-state functional network connectivity (FNC) among auditory, frontal-parietal, default-mode, visual, and motor networks were gainfully adopted for classification of SZ patients and 96% accuracy was achieved using k-nearest neighbors classifier (Arbabshirani et al., 2013). A recent schizophrenia classification challenge demonstrated clearly, across a broad range of classification approaches, the value of rsfMRI data in capturing useful information about this disease (Silva et al., 2014).

Of late, a strategy applying sparsity constraint to spatial patterns has favorably been deployed in various scenarios of fMRI data analysis directed toward extracting information from whole-brain FC patterns (Grosenick et al., 2013; Kim et al., 2012; J.H. Lee et al., 2008; Watanabe et al., 2014). This explicit control of sparsity to analyze fMRI data also includes certain widely used ICA algorithms, such as the popular default algorithms of Infomax and FastICA, which jointly maximize sparsity and independence (Calhoun et al., 2013). This sparsity control has also been beneficial for brain decoding via fMRI data classification (Ng and Abugharbieh, 2011). The sparsity constraint strategy is particularly well-suited to fMRI data given the inherent high dimensionality and intra-subject variability. Moreover, sparsity constraint using total variation penalization (Michel et al., 2012) or anatomically-informed spatiotemporally smooth sparse constraint (Ng et al., 2012) for decoding of fMRI data can explicitly model intra/inter-subject variability, thus resulting in superior performance compared with the least absolute shrinkage and selection operator (LASSO)-based classifier (Michel et al., 2012; Ng and Abugharbieh, 2011; Ng et al., 2012).

The sparsity constraint strategy was recently put into play with rsfMRI data acquired from SZ patients and other neuropsychiatric patients, facilitating the identification of aberrant FC-based attributes, the extraction of distinct and sparse SZ-associated FC networks, and the subsequent application of these attributes and networks to automated classification and diagnosis (Cao et al., 2014; Watanabe et al., 2014). For instance, Watanabe et al. (2014) discovered clinically informative feature sets by using the same data set employed in the current study (see Methods section) via a sparsity constraint with a fused LASSO scheme for the conventional support vector machine (SVM) classifier. Altered FC patterns were prominent in the fronto-parietal networks, the default-mode networks (DMNs), and the cerebellar areas, and the corresponding accuracy was 71.9% (Watanabe et al., 2014).

A deep neural network (DNN) with multiple hidden layers has achieved unprecedented classification performance relative to the SVM and other conventional models (e.g., the hidden Markov model) in various data sets such as image and speech data (Graves et al., 2013; Krizhevsky et al., 2012). This technical breakthrough was accomplished by overcoming the limitations of traditional multilayer neural networks that are based on standard back-propagation algorithms and prone to over-fitting to the training data (Schmidhuber, 2014). More specifically, the distinct characteristics of DNN training encompass (1) unsupervised layer-wise pre-training followed by fine-tuning (Bengio et al., 2007), and (2) stochastic corruption of the input pattern or weight parameters via random zeroing, for example, a denoising autoencoder (Hinton et al., 2012; Vincent et al., 2010). Despite accumulating evidence showing the superiority of the DNN, previous applications of DNN to neuroimaging data are limited to only a few studies (Brosch and Tam, 2013; Hjelm et al., 2014; Plis et al., 2014; Suk et al., 2013). Among the limited attempts to apply the DNN to neuroimaging data, the restricted Boltzmann machine as a building block for the DNN network model has demonstrated its improved capacity to extract spatial and temporal information of fMRI data compared with conventional matrix factorization schemes, such as ICA and PCA algorithms (Hjelm et al., 2014). In addition, Suk et al. (2013) investigated the DNN training strategy by employing a stacked autoencoder (SAE) to discriminate Alzheimer's disease patients from mild cognitive impairment patients. This was done by using volumetric information derived from structural MRI data combined with cerebral glucose metabolism data obtained by positron emission tomography. More recently, Plis et al. (2014) provided a validation study of DNN applied to several types of neuroimaging data, providing evidence that DNN can learn important features such as disease severity (Plis et al., 2014).

Whole-brain FC patterns from fMRI data have not yet been utilized as input patterns to demonstrate the efficacy of the DNN for classification of SZ or other neuropsychiatric disorders. Therefore, the objective of the present investigation was to enhance the classification accuracy of SZ patients vs. HC subjects by using the DNN classifier and whole-brain FC patterns estimated from rsfMRI data. The DNN has been applied to various data sets, such as image and speech data as well as neuroimaging data, with less than 1000 input dimensions (i.e., number of nodes in the input layer) (Graves et al., 2013; Krizhevsky et al., 2012). Compared with these data sets, a dimension of the whole-brain FC patterns can easily reach approximately 5000 when the whole brain is divided into 100 sub-regions. This high dimensionality would be confounded by a lack of straightforward interpretations of whole-brain FC patterns compared with those of speech, image data, and other neuroimaging modalities such as raw fMRI volumes and structural MRI data. Thus, training the DNN using complex and high-dimensional whole-brain FC patterns is inherently challenging. To this end, we evaluated our supposition that classification accuracy can be enhanced by (1) deploying sparsity control of DNN weight parameters and (2) systematically initializing the weight parameters via a pre-training scheme.

We defined a sparsity level of DNN weights as the ratio between a number of non-zero values of DNN weights and a total number of DNN weights (i.e., non-zero ratio). Then, to explicitly control the sparsity of the DNN weights, we developed an adaptive scheme to control the non-zero ratios of the weights between two connected layers to target levels. We then hypothesized that the DNN using the proposed scheme would improve the classification accuracy of SZ patients and

Please cite this article as: Kim, J., et al., Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence ..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.05.018