Structured Deep Generative Model of fMRI Signals for Mental Disorder Diagnosis

Takashi Matsubara(✉), Tetsuo Tashiro, and Kuniaki Uehara

The Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada, Kobe, Hyogo 657-8601, Japan
matsubara@phoenix.kobe-u.ac.jp
http://www.ai.cs.kobe-u.ac.jp/

Abstract. Machine learning-based accurate diagnosis of psychiatric disorders is expected to find their biomarkers and to evaluate the treatments. For this purpose, neuroimaging datasets have required special procedures including feature-selections and dimensional-reductions since they are still composed of a limited number of high-dimensional samples. Recent studies reported a certain success by applying generative models to fMRI data. Generative models can classify small datasets more accurately than discriminative models as long as their assumptions are appropriate. Leveraging our prior knowledge of fMRI signal and the flexibility of deep neural networks, we propose a structured deep generative model, which takes into account fMRI images, disorder, and individual variability. The proposed model estimates the subjects' conditions more accurately than existing diagnostic procedures, general discriminative models, and recently-proposed generative models. Also, it identifies brain regions related to the disorders.

Keywords: Deep learning · Generative model
Functional magnetic resonance imaging · Mental-disorder diagnosis
Schizophrenia · Bipolar disorder

1 Introduction

With continuously collecting neuroimaging datasets such as functional magnetic resonance imaging (fMRI) [1], many studies have been conducted on machine learning techniques to find specific biomarkers of neurological and psychiatric disorders [2] such as schizophrenia [3,4]. They also provide an opportunity for appropriate treatments and potentially evaluate the effectiveness of the treatments. Since each neuroimaging dataset is still limited in size compared to datasets for other machine-learning tasks, it requires special analysis procedures including hand-crafted features, feature-selections, and dimensional-reductions [3–5].

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11072, pp. 258–266, 2018.
https://doi.org/10.1007/978-3-030-00931-1_30

Recent studies reported a certain success by applying generative models to fMRI images [6–8]. Generative models classify a small-sized dataset better than discriminative models when their assumptions are appropriate [9]. We can leverage our prior knowledge and auxiliary information by constructing the model structure. Suk et al. [6] used hidden Markov models (HMMs) to model the temporal dynamics underlying fMRI signals. Yahata et al. [5] used the sparse canonical correlation analysis (SCCA) to remove features related to known attributes of no interest (e.g., age and sex). Chen et al. [7] employed a linear model composed of a subset shared by all subjects and the remaining adjusted for expressing the functional topography of each subject. These models take into account the individual variability but cannot generalize to an unknown attribute or subject; the generalization is a fundamental problem for diagnosing disorders [10].

On the other hand, *deep neural networks* (DNNs) are attracting attention as flexible machine-learning frameworks (see [11] for a review). DNNs learn high-level features of a given dataset automatically. DNNs have been used as a supervised classifier (a multilayer perceptron; MLP) [4,12] and an unsupervised feature-extractor (an autoencoder; AE) [4,6,12]. Not limited to them, DNNs called *deep neural generative models* (DGMs) build generative models describing relationships between multiple factors in their network structures [13,14]. Tashiro et al. [8] implemented relationships between fMRI images, class label, and scan-wise variability (signals of no interest, such as something in mind) on a DGM and achieved a better diagnostic accuracy than comparative models.

Given the above, we propose a deep generative model dedicatedly structured for fMRI data analysis called *subject-wise DGM* (sw-DGM). The proposed sw-DGM takes into account individual variability (i.e., a subject-wise feature), which is shared by and inferred from all fMRI images obtained from a subject. Thanks to this inference, the proposed sw-DGM generalizes to an unknown subject unlike the study by Chen et al. [7] and potentially deals with unknown attributes unlike the study by Yahata et al. [5].

We evaluate the proposed sw-DGM using resting state fMRI (rs-fMRI) datasets of schizophrenia and bipolar disorders. Our experimental results demonstrate that the proposed sw-DGM provides a more accurate diagnosis than the conventional methods based on the functional connectivity extracted using the Pearson correlation coefficients (PCC) [3,5] and comparative discriminative and generative models; support vector machine (SVM) [15], long short-term memory (LSTM) [16], DGM [8], and AE+HMM [6]. In addition, the proposed sw-DGM identifies brain regions related to the disorders.

2 Subject-Wise Deep Neural Generative Model

2.1 Subject-Wise Generative Model of FMRI Images

We first propose a structured generative model of a dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ composed of fMRI images $\boldsymbol{x}_i$ and class labels y_i of N subjects indexed by i. Each subject i is a control subject ($y_i = 0$) or has the disorder ($y_i = 1$), and provides T_i fMRI images $\boldsymbol{x}_i = \{x_{i,t}\}_{t=1}^{T_i}$. We assume that each subject i has its own

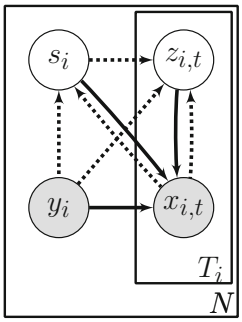

Fig. 1. Our proposed generative model composed of fMRI images $x_{i,t}$, a diagnosis y_i, a subject-wise feature s_i, and scan-wise variabilities $z_{i,t}$.

feature s_i following a prior distribution $p(s)$. The subject-wise feature represents individual variability, which could be brain shape and baseline signal intensity not removed successfully by preprocessing. We also assume that each fMRI image $x_{i,t}$ is associated with the subject's class y_i, the subject-wise feature s_i, and a latent variable $z_{i,t}$. The latent variable $z_{i,t}$ follows a prior distribution $p(z)$ and represents a scan-wise variability, e.g., brain activity related to something in the subject's mind at that moment, body motion, and so on [8]. Then, we build a generative model p_θ of fMRI images $\boldsymbol{x}_i$ conditioned by the class label y_i and parameterized by θ. This is depicted in Fig. 1 and expressed as

$$p_\theta(\boldsymbol{x}_i|y_i) = \prod_{t=1}^{T_i} p_\theta(x_{i,t}|y_i) = \prod_{t=1}^{T_i} \int_{s_i} \int_{z_{i,t}} p_\theta(x_{i,t}|z_{i,t}, y_i, s_i) p(z_{i,t}) p(s_i).$$

According to the variational method [13], the model evidence $\log p_\theta(\boldsymbol{x}_i|y_i)$ is bounded using an inference model q_ϕ parameterized by ϕ as

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}_i|y_i) &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_i, s_i|\boldsymbol{x}_i, y_i)} \left[\log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i, s_i|y_i)}{q_\phi(\boldsymbol{z}_i, s_i|\boldsymbol{x}_i, y_i)} \right] \\
&= -D_{KL}(q_\phi(s_i|\boldsymbol{x}_i, y_i)||p(s_i)) \\
&\quad - \textstyle\sum_{t=1}^{T_i} \mathbb{E}_{q_\phi(s_i|\boldsymbol{x}_i, y_i)} \left[D_{KL}(q_\phi(z_{i,t}|x_{i,t}, y_i, s_i)||p(z_{i,t})) \right] \\
&\quad + \textstyle\sum_{t=1}^{T_i} \mathbb{E}_{q_\phi(s_i|\boldsymbol{x}_i, y_i)} \left[\mathbb{E}_{q_\phi(z_{i,t}|x_{i,t}, y_i, s_i)} \left[\log p_\theta(x_{i,t}|y_i, z_{i,t}, s_i) \right] \right] \\
&=: \mathcal{L}_g(\boldsymbol{x}_i, y_i),
\end{aligned} \tag{1}$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence and $\mathcal{L}_g(\boldsymbol{x}_i; y_i)$ is the evidence lower bound (ELBO); the ELBO is the ordinary objective function of the conditional generative model p_θ and the inference model q_ϕ to be maximized.

The ELBO $\mathcal{L}_g(\boldsymbol{x}_i; y)$ is considered to converge to the model evidence $\log p_\theta(\boldsymbol{x}_i|y)$. We estimate the posterior probability $p(y|\boldsymbol{x}_i)$ of the class y of a subject i based on Bayes' rule:

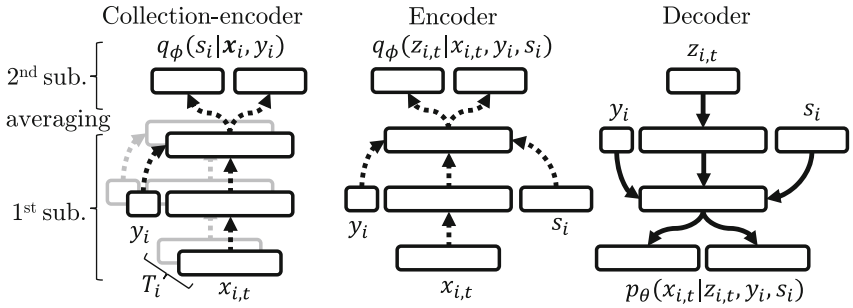

Fig. 2. Implementation of the proposed generative model on the deep neural networks.

$$p_\theta(y|\boldsymbol{x}_i) = \frac{p(y)p_\theta(\boldsymbol{x}_i|y)}{\sum_{y'\in\{0,1\}} p(y')p_\theta(\boldsymbol{x}_i|y')} \approx \frac{p(y)\exp\mathcal{L}_g(\boldsymbol{x}_i, y)}{\sum_{y'\in\{0,1\}} p(y')\exp\mathcal{L}_g(\boldsymbol{x}_i, y')} =: \exp\mathcal{L}_d(\boldsymbol{x}_i, y). \tag{2}$$

We assume the prior probability $p(y)$ of class y to be $p(y=0) = p(y=1) = 0.5$. Hence, if the ELBO $\mathcal{L}_g(\boldsymbol{x}_i, y=1)$ has a large value, the subject i is more likely to have the disorder.

In addition, the approximation of the log-likelihood of the class label, i.e., $\mathcal{L}_d(\boldsymbol{x}_i, y_i)$, can be an alternative objective function to be maximized, progressing discrimination between the classes [9]. We balanced the two objective functions using the coefficient $\omega \in [0, 1]$ as

$$\mathcal{L}(\boldsymbol{x}_i, y_i) = \omega\mathcal{L}_g(\boldsymbol{x}_i, y_i) + (1-\omega)\mathcal{L}_d(\boldsymbol{x}_i, y_i). \tag{3}$$

2.2 Implementation on Deep Neural Networks

We implement the generative model p_θ and inference model q_ϕ described above on deep neural networks, and thereby, propose a subject-wise deep generative model (sw-DGM). We assume a preprocessed fMRI signal $x_{i,t}$, a subject-wise feature s_i, and a scan-wise variability $z_{i,t}$ as vectors of n_x, n_s, and n_z-dimensions, respectively. The inference model $q_\phi(z_{i,t}|x_{i,t}, y_i, s_i)$ and generative model $p_\theta(x_{i,t}|y_i, s_i, z_{i,t})$ are expressed by multivariate Gaussian distributions with diagonal covariance matrices; their parameters are the outputs of the corresponding DNNs called encoder and decoder (see the right two panels in Fig. 2 and the previous studies [8,13,14] for more detail). The implementation of the inference model $q_\phi(s_i|\boldsymbol{x}_i, y_i)$ requires modification because it accepts a variable-length sequence of fMRI images $\boldsymbol{x}_i = \{x_{i,t}\}_{t=1}^{T_i}$ obtained from a subject i. We propose a neural network architecture called *collection-encoder*, which is composed of stacked two sub-networks as depicted in the leftmost panel in Fig. 2. The first sub-network accepts a preprocessed fMRI signal $x_{i,t}$ and the class label y_i, and then outputs a hidden activation $h_{i,t}$. The second sub-network accepts

the averaged hidden activation $\bar{h}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}[h_{i,t}]$ and outputs the variational posterior $q_\phi(s_i|\boldsymbol{x}_i, y_i)$ of the subject-wise feature s_i.

Note that the proposed sw-DGM is not equivalent to other structured DGMs such as Skip Deep Generative Model [14]. They assumed that each sample is generated with more than two latent variables. In contrast, the proposed sw-DGM assumes that the samples $x_{i,t}$ obtained from the same subject i share the subject-wise feature s_i as a latent variable. This assumption potentially gives a good constraint based on a prior knowledge of the fMRI images.

We used three-layered neural networks as the encoder and decoder. We used a two-layered and a single-layered neural networks as the first and the second sub-networks of the collection-encoder, respectively. Each hidden layer of all the DNNs has u_h hidden units followed by the layer normalization [17] and the ReLU activation function [18]. For approximating the expectations in Eq. (1), the subject-wise feature s_i and the scan-wise variability $z_{i,t}$ were sampled from the variational posteriors $q_\phi(s_i|\boldsymbol{x}_i, y_i)$ and $q_\phi(z_{i,t}|x_{i,t}, y_i, s_i)$ once per sample $x_{i,t}$ in the training phase and were substituted with the MAP estimations in the test phase following the previous work [13]. The preprocessed fMRI signals $x_{i,t}$ were augmented using the dropout [19] of a ratio p. All the DNNs were jointly trained using the Adam optimization algorithm [20] with parameters $\alpha = 10^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We selected hyper-parameters from $p \in \{0.0, 0.5\}$, $n_h \in \{50, 100, 200, 400\}$, $n_z = n_s \in \{5, 10, 20, 50, 100\}$ for $n_h > n_z = n_s$, and $\omega \in \{0.0, 0.9, 0.99\}$. We adjusted the imbalance in the classes via oversampling.

3 Experiments and Results

3.1 Data Acquisition and Comparative Models

We used datasets obtained from the OpenfMRI database. Its accession number is ds000030 (https://openfmri.org/dataset/ds000030/). We performed a preprocessing procedure for rs-fMRI using the SPM12 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). We discarded the first 10 scans of each subject to ensure magnetization equilibrium. We performed time-slice adjustment, realignment of brain positions via a rigid body rotation, and spatial normalization using the MNI space with a voxel thickness of 2.0 mm. We parcellated each fMRI image into 116 regions of interest (ROIs) using the automated anatomical labeling (AAL) template [21] and averaged intensities of voxels in each ROI region, obtaining a 116-dimensional vector as a preprocessed fMRI signal $x_{i,t}$. As scrubbing, we discarded frames with frame displacements (FD) of more than 1.5 mm or angular rotations of more than 1.5° in any direction as well as the following frames. We also discarded subjects who had less than 100 remaining frames and subjects whose fMRI images did not match the MNI template after the spatial normalization. As a result, we obtained 113 control subjects, 44 patients with the schizophrenia, and 45 patients with the bipolar disorder.

As baselines, we evaluated two conventional procedures, which use Pearson's correlation coefficients (PCCs) between the ROIs as the functional connectivities

Table 1. Diagnostic accuracies.

	Schizophrenia			Bipolar		
Model	BACC	SPEC	SEN	BACC	SPEC	SEN
PCC+Kendall+PCA+c-means [3]	0.640	0.635	0.645	0.602	0.565	0.640
PCC+SCCA+SLR [5]	0.639	0.779	0.500	0.607	0.735	0.480
SVM [15]	0.505	0.788	0.223	0.512	0.855	0.169
LSTM [16]	0.661	0.854	0.467	0.571	0.802	0.340
DGM [8]	0.722	0.920	0.524	0.619	0.650	0.587
AE+HMM [6]	0.618	0.554	0.682	0.616	0.490	0.742
sw-DGM (proposed)	**0.767**	0.812	0.722	**0.622**	0.844	0.401

(FCs) [3,5]. Following Shen et al. [3], we selected m features in the FCs using the Kendall τ correlation coefficient, compressed the feature vector into a d-dimensional space using the locally linear embedding (LLE) with a parameter of k, and clustered them into two classes using the c-means algorithm. This procedure was confirmed to outperform direct classification of the PCCs by the SVM and MLP. Following Yahata et al. [5], we selected m features in the FCs using the SCCA and classified the features using the sparse logistic regression (SLR) with a sparsity determined by automatic relevance determination (ARD). We selected the hyper-parameters from $m \in \{50, 100, 200, 400, 600\}$, $k \in \{5, 8, 10, 12, 15\}$, and $d \in \{2, 5, 10, 20, 50\}$ following the original study [3].

For comparison, we evaluated classifiers; support vector machine (SVM) [15] and long short-term memory (LSTM) [16]. The SVM accepted a single image $x_{i,t}$ and outputted a binary value representing the estimated class using linear kernels. The diagnosis of a subject i is determined by majority voting of T_i estimations, consistent with other comparative models. We selected the hyper-parameter C adjusting the trade-off between classification accuracy and margin maximization from $C \in \{\ldots, 0.1, 0.2, 0.5, 1, 2, 5, 10, \ldots\}$. The LSTM is a recurrently connected neural network, accepting fMRI signals $\boldsymbol{x}_i = \{x_{i,t}\}_{t=1}^{T_i}$ sequentially and outputting the posterior probability $p(y|\boldsymbol{x}_i)$ using the logistic function. The other conditions were the same as those for the proposed sw-DGM.

We also evaluated a simpler DGM proposed in the previous study [8] and hidden Markov model (HMM) with autoencoder (AE) [6]. The DGM modeled relationships between the fMRI signals $\boldsymbol{x}_i$, the class label t_i, and the scan-wise variability $z_{i,t}$ using an encoder $q(z_{i,t}|x_{i,t}, y_i)$ and a decoder $p(x_{i,t}|y_i, z_{i,t})$ but does not take into account the subject-wise feature s_i [8]. Following Suk et al. [6], we compressed each fMRI image into a d-dimensional space using an AE. Then, we trained a pair of HMMs; $p_\theta(x_{i,t}|y = 1)$ for patients and $p_\theta(x_{i,t}|y = 0)$ for control subjects. Each HMM had Gaussian distributions with full covariance matrices and was trained using Expectation-Maximization (EM) algorithm. We calculated the posterior probability $p(y|\boldsymbol{x}_i)$ using Bayes' rule. We selected the number n_z of units in the bottleneck layer from $n_z \in \{2, 3\}$, the number n

Table 2. Top 5 contribution weights for diagnosis.

Schizophrenia		Bipolar	
ROI	Weight	ROI	Weight
Cerebelum_6_L	0.0555	Cingulum_Ant_R	0.0132
Postcentral_L	0.0532	Frontal_Inf_Orb_L	0.0121
Cingulum_Mid_L	0.0531	Cerebelum_7b_R	0.0116
Lingual_R	0.0529	ParaHippocampal_L	0.0114
Lingual_L	0.0526	Temproal_Mid_L	0.0106

of mixture components of the HMM from $n \in \{2, 3, 4, 5, 6, 7\}$, and the hyper-parameters of the AE in the same ranges as the proposed sw-DGM.

3.2 Results of Diagnosis and Contribution Weights of ROIs

Since the datasets are imbalanced, we used the following measures; sensitivity $\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, specificity $\mathrm{SPEC} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, and balanced accuracy $\mathrm{BACC} = 0.5 \times (\mathrm{SEN} + \mathrm{SPEC})$, where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. We performed 5 trials of 10-fold cross-validation (CV) and summarized the results in Table 1. The proposed sw-DGM achieved the best balanced accuracies among the competitive approaches in the both datasets. Especially, the proposed sw-DGM outperformed or at least performed no worse than the existing DGM [8], implying that the introduction of the subject-wise feature s_i (i.e., individual variability) worked as an appropriate constraint.

As shown in Eq. 2, the diagnosis of a subject i is based on the difference in the conditional log-likelihood $\log p_\theta(\boldsymbol{x}_i|y)$ between the class labels $y = 0$ and $y = 1$. Since each element $x_{i,t,r}$ of an fMRI signal $x_{i,t}$ corresponds to an ROI r, we can calculate the ROI-wise average marginal log-likelihoods $\mathbb{E}_{i,t}[\log p_\theta(x_{i,t,r}|y_i)|x_{i,t}]$. An ROI with a large difference in the log-likelihoods between correct and incorrect labels has a large effect on the accurate diagnosis. Hence, we defined

$$W_r = \mathbb{E}_{i,t}\left[\log p_\theta(x_{i,t,r}|y_j) - \log p_\theta(x_{i,t,r}|1 - y_j)|x_{i,t}\right]$$

as the contribution weight W_r of the ROI r and summarized the ROIs with the top 5 contribution wights in Table 2. Previous studies (e.g., the review paper [22]) have discussed the relationships of some of the listed ROIs to the disorders. The results suggest that the proposed sw-DGM identified the ROIs related to the disorders.

4 Conclusion

This study proposed a subject-wise deep generative model (sw-DGM) of fMRI images dedicatedly structured for diagnosing psychiatric disorders. The sw-DGM

modeled the joint distribution of rs-fMRI images, class label, individual variability, and scan-wise variability. The individual variability worked as an appropriate constraint, and the sw-DGM achieved a diagnostic accuracy higher than other conventional and comparative approaches. Also, the sw-DGM identified brain regions related to the disorders.

Acknowledgments. The authors would like to acknowledge Dr. Ben Seymour, Dr. Kenji Leibnitz, Dr. Hiroaki Mano, and Dr. Ferdinand Peper at CiNet for valuable discussions. This study was supported by the JSPS KAKENHI (16K12487), SEI Group CSR Foundation, and the MIC/SCOPE #172107101.

References

1. Sejnowski, T.J., et al.: Putting big data to good use in neuroscience. Nat. Neurosci. **17**(11), 1440–1441 (2014)
2. Group, B.D.W.: Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinic. Pharmacol. Ther. **69**(3), 89–95 (2001)
3. Shen, H., et al.: Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. NeuroImage **49**(4), 3110–3121 (2010)
4. Castro, E., et al.: Deep independence network analysis of structural brain imaging: application to schizophrenia. IEEE Trans. Med. Imaging **35**(7), 1729–1740 (2016)
5. Yahata, N., et al.: A small number of abnormal brain connections predicts adult autism spectrum disorder. Nat. Commun. **7**(7), 11254 (2016)
6. Suk, H.I., et al.: State-space model with deep learning for functional dynamics estimation in resting-state fMRI. NeuroImage **129**, 292–307 (2016)
7. Chen, P.H., et al.: A Reduced-Dimension fMRI Shared Response Model. In: NIPS. (2015) 460–468
8. Tashiro, T., et al.: Deep neural generative model for fMRI image based diagnosis of mental disorder. In: NOLTA (2017)
9. Lasserre, J., et al.: Principled hybrids of generative and discriminative models. In: CVPR, pp. 87–94 (2006)
10. Abraham, A., et al.: Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. NeuroImage **147**, 736–745 (2017)
11. Schmidhuber, J.: Deep learning in neural networks: an overview. Neur. Netw. **61**, 85–117 (2015)
12. Liu, S., et al.: Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. **62**(4), 1132–1140 (2015)
13. Kingma, D.P., et al.: Semi-supervised learning with deep generative models. In: NIPS, pp. 3581–3589 (2014)
14. Maaløe, L., et al.: Auxiliary deep generative models. In: ICML, vol. 48, pp. 1445–1453 (2015)
15. Pereira, F., et al.: Machine learning classifiers and fMRI: a tutorial overview. NeuroImage **45**, S199–S209 (2009)
16. Dvornek, N.C., et al.: Identifying autism from resting-state fMRI using long short-term memory networks. MLM **I**, 362–370 (2017)
17. Ba, J.L., et al.: Layer normalization, pp. 1–14. arXiv (2016)
18. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: ICML, pp. 807–814 (2010)

19. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**, 1929–1958 (2014)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR, pp. 1–15 (2015)
21. Tzourio-Mazoyer, N., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage **15**(1), 273–289 (2002)
22. Andreasen, N.C., Pierson, R.: The role of the cerebellum in schizophrenia. Biol. Psychiatry **64**(2), 81–88 (2008)