Spatiotemporal Attention Autoencoder (STAAE) for ADHD Classification

Qinglin Dong[1], Ning Qiang[2], Jinglei Lv[3], Xiang Li[1,5], Tianming Liu[4], and Quanzheng Li[1,5(✉)]

[1] Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
li.quanzheng@mgh.harvard.edu
[2] School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China
[3] School of Biomedical Engineering and Sydney Imaging, Brain and Mind Centre, The University of Sydney, Camperdown, Australia
[4] Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA, USA
[5] MGH & BWH Center for Clinical Data Science, Boston, MA, USA

Abstract. It has been of great interest in the neuroimaging community to model spatiotemporal brain function and disorders based on resting state functional magnetic resonance imaging (rfMRI). A variety of spatiotemporal methods have been proposed for rfMRI so far, including deep learning models such as convolution networks (CNN) and recurrent networks (RNN). However, the dominant models fail to capture the long-distance dependency (LDD) due to their sequential nature, which becomes critical at longer sequence lengths due to memory limit. Inspired by human brain's extraordinary ability of long-term memory and attention, the attention mechanism is designed for machine translation to draw global dependencies and achieved state-of-the-art. In this paper, we propose a spatiotemporal attention autoencoder (STAAE) to discover global features that address LDDs in rfMRI. STAAE encodes the information throughout the rfMRI sequence and reveals resting state networks (RSNs) that characterize spatial and temporal properties of the data. Considering that the rfMRI is measured without external tasks, an unsupervised classification framework is developed based on the connectome generated with STAAE. This framework has been evaluated on 281 children with ADHD and 266 normal control children from 4 sites of ADHD200 datasets. The proposed STAAE reveals the global functional interaction in the brain and achieves a state-of-the-art classification accuracy from 59.5% to 77.2% on multiple sites. It is evident that the proposed attention-based model provides a novel approach towards better understanding of human brain.

Keywords: Deep learning · rfMRI · Attention mechanism · Functional networks · ADHD

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

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12267, pp. 508–517, 2020.
https://doi.org/10.1007/978-3-030-59728-3_50

1 Introduction

The resting state networks (RSNs) from resting state functional MRI (rfMRI) provides a powerful tool to model brain functions and disorders even in the absence of an external task [1–5]. Various machine learning methods have been successful applied on rfMRI to exploit RSNs, such as independent component analysis (ICA) [6–9] and sparse dictionary learning (SDL) [10–14]. Due to the superior representation power, deep learning models have also been increasingly employed for fMRI analysis, such as Convolution Neural Network (CNN) [15–17] and Recurrent Neural Network (RNN) [18–21]. However, evidences show that the cognitive actions of multiple regions of brains is related to their earlier actions, with a potential long distance in time [22]. While modeling, the so-called long-distance dependency (LDD), is a challenging issue for CNN and RNN to address [23].

Recently, the attention mechanism has gained popularity in sequence modeling and various tasks [23]. Compared to regular CNN and RNN, the attention mechanism models every unit in the input sequence simultaneously and draws global dependencies without regard to their distance [24]. It has also been proven that pure attention mechanism has comparable representation powers than CNNs or RNNs [23]. To utilize the superior ability of attention mechanism to mining LDD, we explore the possibility to model rfMRI with attention mechanism. Considering the unsupervised nature of rfMRI data, a spatiotemporal attention autoencoder (STAAE) is proposed to model the rfMRI sequence data. With the proposed model, the relation of two volumes/frames in the sequence is captured with an attention score measuring the distance of their embeddings. We aim to improve the classification by addressing the LDD issue. To our best knowledge, this is the first study that exploits attention mechanism for fMRI modeling.

Attention Deficit Hyperactivity Disorder (ADHD) is a mental health disorder involves multiple attention related problems, but neither a comprehensive pathophysiology model nor a biomarker for clinical practice is established yet [25–30]. The proposed model has been applied on ADHD200 datasets for evaluation. First, we examined the learned representation of the input data, which encodes the global information throughout the sequence of rfMRI. Our model reveals meaningful RSNs that characterize spatial and temporal properties of the data. Secondly, based on the learned RSNs, the connectomes are generated for each subject and are used for ADHD classification with cross-validation. The experimental results indicated that the proposed framework is capable of modeling and classifying on ADHD. It's worth noting that the proposed integrated pipeline can be easily generalized for other mental disorder classification.

2 Methods

The proposed computational framework is shown in Fig. 1. In Sect. 2.1, the preprocessed rfMRI data of all subjects are registered to a standard space for group-wise learning and masked to a 2D spatiotemporal matrix. In Sect. 2.2, a STAAE model consists of a pair of encoder and decoder which takes rfMRI volumes as input. In Sect. 2.3, the intermediate representations of rfMRI data are interpreted to RSNs. The functional connectomes are built for further classification.

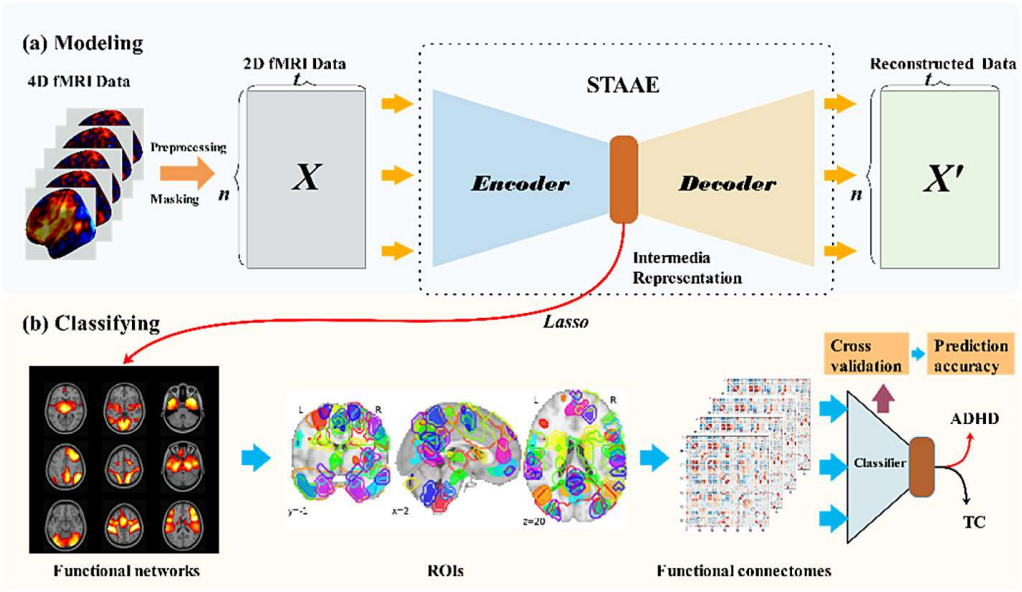

Fig. 1. Illustration of STAAE based pipeline for modeling and classifying of ADHD-200. (a) Modeling process: the outline of STAAE model, in which the input and output are raw fMRI signals and reconstructed signals respectively, and the high-level features extracted by the encoder are used to construct RSNs. (b) Classifying process: by extracting from the RSNs generated by the STAAE model, a functional connectome for each subject is calculated. A feedforward neural network is trained as classifier based on the functional connectomes and used for classification.

2.1 Dataset and Preprocessing

In this work, we used the subjects from four sites in ADHD-200 dataset: Kennedy Krieger Institute (KKI), Peking University (PU), New York University Medical Center (NYU), and NeuroImage (NI). For all our experiments, we use preprocessed data publicly available from Preprocessed Connectomes Project [27]. The Athena preprocessing pipeline was adopted, which is based on tools from the AFNI and FSL software packages, including skull striping, slice timing correction, motion correction, detrending, band filtering (0.01–0.01 Hz), normalization and masking. To perform group-wise STAAE training, all subjects' data are nonlinearly registered to the MNI152 $4 \times 4 \times 4$ mm^3 standard template space [31] (Table 1).

Table 1. Summary of fMRI dataset.

Imaging site	Total subjects	Control subjects	ADHD subjects
KKI	83	67	22
PU	194	78	116
NYU	216	98	118
NI	48	23	25

2.2 Attention Mechanism and STAAE

In respect of sequence modeling, especially high-dimension spatiotemporal sequence data like fMRI, both convolutional network and recurrent network have been used in the field. To utilize the CNN's hierarchical feature abstraction ability, a 1D temporal convolution architecture was applied on the fMRI time series [15, 16]. This approach was able to extract features from low-level to high level, however, it did not make use of the rich spatial information from fMRI. To incorporate the spatial and temporal information at the same time, recurrent network was applied on the fMRI volumes and preserving temporal features with long short-term memory (LSTM), which is a typical recurrent module [18, 20, 21, 32]. This approach established a unified spatiotemporal frame; however, it comes with three drawbacks. First, the inherently sequential nature of RNN/LSTM precludes parallelization, which causes notable time cost especially for high-dimension data like fMRI. Second, the sequential nature also leads to the notorious long-distance dependency (LDD) problem, which becomes critical at longer sequence lengths due to memory limit [24]. Third, the encoder LSTM is used to process the entire input sentence and encode it into a context vector, where the intermediate states of the encoder are ignored.

In this paper, we propose to solve the above-mentioned drawbacks by substituting the convolution and recurrent networks with the attention mechanism [22], which draws global dependencies and achieved state of the art in multiple sequence modeling tasks. The attention mechanism consists of three matrices: queries $Q \in \mathbb{R}^{n \times d_k}$, keys $K \in \mathbb{R}^{m \times d_k}$ and values $V \in \mathbb{R}^{m \times d_v}$. In the context of rfMRI, a key vector and a query vector are learned for each frame of volume, and the pairs of query-key are matched across all frame simultaneously. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. If a pair of query-key matches, it generates a high value as output. As shown in Fig. 2, we compute the matrix of outputs as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Most sequence models follow an encoder-decoder paradigm, and it also applies to our proposed STAAE. Considering the intrinsic unsupervised nature of rfMRI, i.e. no external stimulus or task is performed, an autoencoder structure is adopted, where the decoder aims to reconstruct the exact input. For STAAE, the encoder maps an input sequence of symbol representations $X = (x_1, \ldots, x_t)$ to a sequence of intermediate representations $Z = (z_1, \ldots, z_t)$. More specifically, each xi represents a volume of rfMRI and is embedded with a fully feedforward network. Given Z, the decoder tries to generate a reconstructed sequence of $X' = (x_1', \ldots, x_t')$.

Hyperbolic tangent function was chosen as the activation for the rfMRI data. To start with training, the weights and biases are initialized from a Gaussian with zero-mean and a standard deviation of 0.01. To improve the convergence, batch normalization technique was applied to each hidden layer, which explicitly forced the activations to be unit Gaussian distributed. With a learning rate of 0.0001 and batch size of 1, the models were trained for 200 epochs for convergence. All experiments were repeated 5 times to

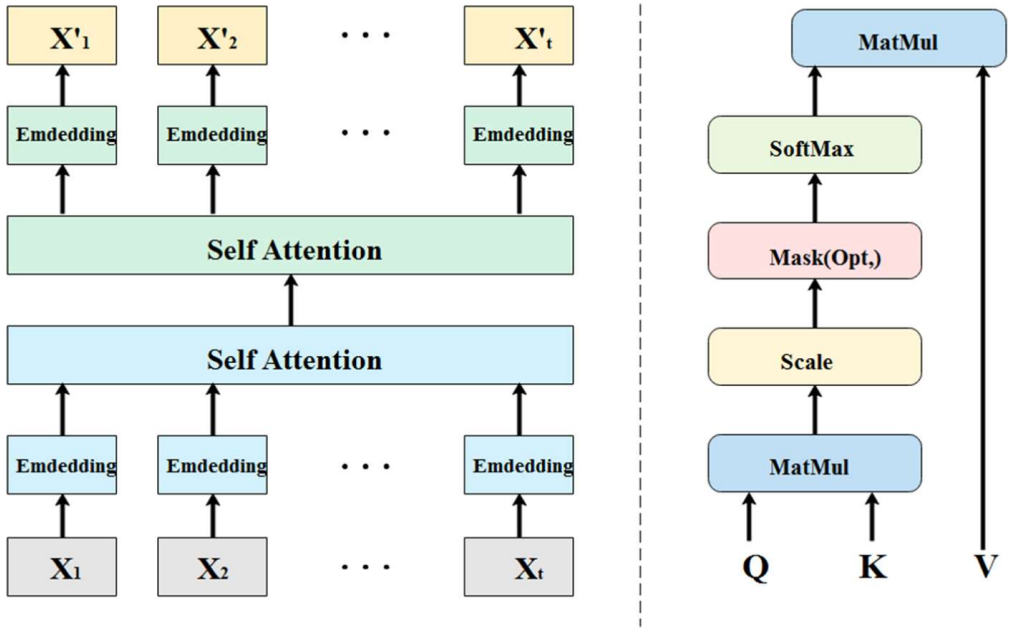

Fig. 2. Structure of STAAE and illustration of attention mechanism.

test the stability of consistency of results. The implementation of STAAE can be found at https://github.com/QinglinDong/stAAE.

2.3 Feature Interpretation and Classification

To explore the intermediate representation learned with STAAE, we apply Lasso regression to estimate the coefficient matrix which is used to build spatial maps. As shown in Fig. 1, the group-wise fMRI data X is fed into the trained encoder, yielding the intermediate representation Z from the output of encoder. Next, the RSNs W are derived from the intermediate representation and group-wise input via Lasso regression as follow:

$$W = min\|Z - XW\|_2^2 + \lambda\|W\|_1 \tag{2}$$

After the Lasso regression, W is regularized and transposed to a coefficient matrix, then each row of coefficient matrix is mapped back to the original 3D brain image space, which is the inverse operation of masking in data preprocessing. Thus, the RSNs are generated and interpreted in a neuroanatomically meaningful context. As shown in Fig. 3, after transformation into "Z-scores" across spatial volumes, all the RSNs were thresholded at $Z > 2.3$.

To exploit the spatiotemporal features including the RSNs and the intermediate representations for further classification, we follow three steps. First, as shown in Fig. 1, with the established RSNs, a union set of regions of interests (ROIs) from all the RSNs are combined to establish a comprehensive brain atlas. Second, we extract the time series from the original training data masked by the atlas. Third, the functional connectome, which reflects the level of co-activation of brain regions, is calculated based on the Pearson correlation of the extracted time series. The upper triangle matrix values were removed for less redundancy.

For classification, the functional connectome is flattened to a one-dimensional feature vector. A deep feedforward neural network classifier is used as classifier, where cross entropy loss is used for binary classification. There are 2 hidden layers in DNN, and the numbers of nodes are 1000 and 500, respectively. A 10-fold cross validation was run 10 times to measure the prediction accuracy of whole pipeline. The results of classification accuracy on all sites through the proposed STAAE based pipeline are shown in Sect. 3.2.

3 Results

3.1 RSNs from STAAE

The RSNs derived from the STAAE is shown in Fig. 3. It is observed that the RSNs are intrinsic active even when no extra tasks the subjects are doing, which provides evidence supporting the conclusion in [2–4, 33]. By visual inspection, these RSNs can be well interpreted, and they agree with domain knowledge of functional network atlases in the literature. To quantitatively evaluate the performance of STAAE in modeling RfMRI data, a comparison study between STAEE derived RSN and templates [1] is provided in this section.

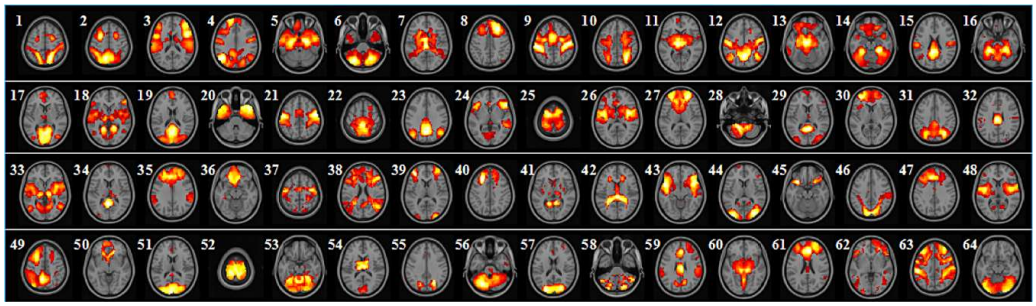

Fig. 3. Overview of the RSNs derived from STAAE.

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

$$IoU\left(N^{(1)}, N^{(2)}\right) = \frac{\sum_{i=1}^{n}\left|N_i^{(1)} \cap N_i^{(2)}\right|}{\sum_{i=1}^{n}\left|N_i^{(1)} \cup N_i^{(2)}\right|} \tag{3}$$

With the similarity measure defined above, the similarities $IoU\left(N_{ResAE}, N_{ICA}\right)$ and $IoU\left(N_{AE}, N_{ICA}\right)$ are quantitatively measured, where ICA is considered as ground truth for the intrinsic RSNs. Comparisons of pairs by these two methods are shown in Fig. 4, and the quantitative comparison are shown on the sides. This result demonstrated that STAAE can identify intrinsic RSNs very well, suggesting the effectiveness and meaningfulness of our proposed model.

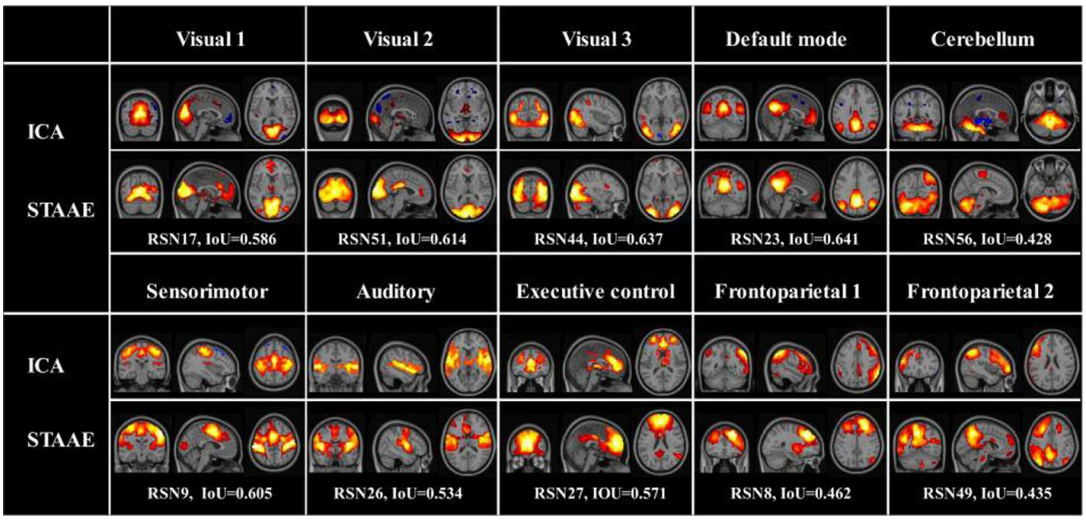

Fig. 4. Comparison of STAAE RSN with templates based on ICA [1].

3.2 Prediction Accuracy on ADHD-200

As illustrated in Sect. 2.3, we use STAAE derived RSNs to build functional connectome for each subject, which are further used for classification. We compared the prediction accuracies achieved by our STAAE based pipeline with SDL [34] based pipeline and RAE based pipeline. As shown in Fig. 5, by using the same classification framework and configurations, the STAAE based pipeline performed better than the other two method. The average prediction accuracies of STAAE based pipeline on NYU, PU, KKI and NI datasets are 59.5%, 65.2%, 77.2% and 61.0%, respectively (marked by green triangles). Besides, the variances of the STAAE based pipeline are smaller than other two methods that indicate the robust performance of our method.

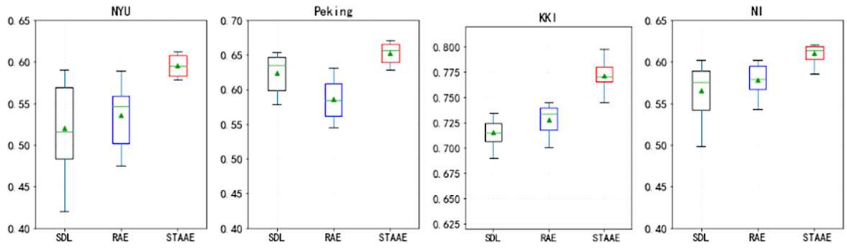

Fig. 5. Comparison of results achieved by SDL, RAE, STAAE based pipeline.

The results are also compared with other models in previous literature, including Support vector machine (SVM) and ICA. Table 2 shows the average prediction accuracies of SVM [29], ICA [30], SDL [34], RAE [18], and STAAE classification pipelines on ADHD200 dataset. It shows RAE [18] outperforms traditional shallow machine learning models such as SVM [29], ICA [30] and SDL [34] and our proposed STAAE outperforms RAE [18]. Overall, based on the RSNs derived by the STAAE model, our classification

pipeline performed excellent and competitive compared to other models and methods for ADHD classification. These results also imply the effectiveness of STAAE on RSN modeling.

Table 2. The STAAE achieves better average classification accuracy than previous state-of-the-art models on the ADHD200 dataset

Name	SVM [29]	ICA [30]	SDL [34]	RAE [18]	STAAE
NYU	–	56%	52.0%	53.5%	**59.5%**
PU	58.82%	58%	62.4%	58.7%	**65.2%**
KKI	54.55%	81%	71.6%	72.8%	**77.2%**
NI	48.00%	–	56.5%	57.8%	**61.0%**

4 Discussions

This paper is among the earliest studies that explore modeling fMRI with attention mechanism, to our best knowledge. In this paper, we proposed to adopt the encoder-decoder structure to exploit the attention mechanism for the unsupervised rfMRI sequence. With a group-wise experiment on massive rfMRI data, the proposed model shows its capability to learn RSNs. A comparison study with SDL and RAE showed that the RSNs learned by STAAE are meaningful and can be well interpreted. One limitation of our current approach is that the effects of hyperparameters is not fully explored, including the model depth, number of attention head and size of attention head. By tuning the parameters, the proposed framework can even achieve higher performance in the future.

For language modeling, it is crucial to solve the issue of ambiguity, where one word can have different meanings in different context. By learning contextualized word embedding based on attention mechanism, Bidirectional Encoder Representations from Transformers (BERT) has already achieved great success and dominated the natural language processing. [35] For brain modeling, ambiguity and context issue, not only because of the incomplete supervision nature of fMRI and lack of ground truth, but also multiple RSNs are activated simultaneously and each RSN may serve more than one function. [13, 36] It is interesting and feasible to model the multiple-demand system in brain and explore RSNs in different context with extended attention network in future work.

References

1. Smith, S.M., et al.: Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. **106**(31), 13040–13045 (2009)
2. Kanwisher, N.: Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. **107**(25), 11163–11170 (2010)

3. Harris, K.D., et al.: Cortical connectivity and sensory coding. Nature **503**(7474), 51 (2013)
4. Pessoa, L.: Understanding brain networks and brain organization. Phys. Life Rev. **11**(3), 400–435 (2014)
5. Lv, J., et al.: Task fMRI data analysis based on supervised stochastic coordinate coding. Med. Image Anal. **38**, 1–16 (2017)
6. McKeown, M.J.: Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. NeuroImage **11**(1), 24–35 (2000)
7. Calhoun, V.D., et al.: A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp. **14**(3), 140–151 (2001)
8. Beckmann, C.F., et al.: Investigations into resting-state connectivity using independent component analysis. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360**(1457), 1001–1013 (2005)
9. Calhoun, V.D., et al.: Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. IEEE Rev. Biomed. Eng. **5**, 60–73 (2012)
10. Lv, J., et al.: Sparse representation of whole-brain fMRI signals for identification of functional networks. Med. Image Anal. **20**(1), 112–134 (2015)
11. Jiang, X., et al.: Sparse representation of HCP grayordinate data reveals novel functional architecture of cerebral cortex. Hum. Brain Mapp. **36**(12), 5301–5319 (2015)
12. Ge, F., et al.: Deriving ADHD biomarkers with sparse coding based network analysis. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE (2015)
13. Li, X., et al.: Multiple-demand system identification and characterization via sparse representations of fMRI data. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE (2016)
14. Ge, F., et al.: Exploring intrinsic networks and their interactions using group wise temporal sparse coding. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
15. Huang, H., et al.: Modeling task fMRI data via mixture of deep expert networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
16. Huang, H., et al.: Modeling task fMRI data via deep convolutional autoencoder. IEEE Trans. Med. Imaging **37**(7), 1551–1561 (2018)
17. Zhao, Y., et al.: Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks. IEEE Trans. Biomed. Eng. **65**(9), 1975–1984 (2018)
18. Li, Q., et al.: Simultaneous spatial-temporal decomposition of connectome-scale brain networks by deep sparse recurrent auto-encoders. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 579–591. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_45
19. Sak, H., et al.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
20. Wang, L., et al.: Decoding dynamic auditory attention during naturalistic experience. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE (2017)
21. Wang, H., et al.: Recognizing brain states using deep sparse recurrent neural network. IEEE Trans. Med. Imaging **38**, 1058–1068 (2018)
22. Piñango, M.M., et al.: The localization of long-distance dependency components: integrating the focal-lesion and neuroimaging record. Front. Psychol. **7**, 1434 (2016)
23. Bahdanau, D., et al.: Neural machine translation by jointly learning to align and translate (2014)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
25. Riaz, A., et al.: Fusion of fMRI and non-imaging data for ADHD classification. Comput. Med. Imaging Graph. **65**, 115–128 (2018)

26. Itani, S., et al.: A multi-level classification framework for multi-site medical data: application to the ADHD-200 collection. Expert Syst. Appl. **91**, 36–45 (2018)
27. Bellec, P., et al.: The neuro bureau ADHD-200 preprocessed repository. Neuroimage **144**, 275–286 (2017)
28. dos Santos Siqueira, A., et al.: Abnormal functional resting-state networks in ADHD: graph theory and pattern recognition analysis of fMRI data. Biomed. Res. Int. **2014**, 380531 (2014)
29. Dey, S., et al.: Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. Front. Neural Circuits **8**, 64 (2014)
30. Nuñez-Garcia, M., Simpraga, S., Jurado, M.A., Garolera, M., Pueyo, R., Igual, L.: FADR: functional-anatomical discriminative regions for rest fMRI characterization. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 61–68. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24888-2_8
31. Abraham, A., et al.: Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. **8**, 14 (2014)
32. Cui, Y., et al.: Identifying brain networks of multiple time scales via deep recurrent neural network. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 284–292. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_33
33. Pessoa, L.: Beyond brain regions: network perspective of cognition–emotion interactions. Behav. Brain Sci. **35**(3), 158–159 (2012)
34. Lv, J., et al.: Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. IEEE Trans. Biomed. Eng. **62**(4), 1120–1131 (2015)
35. Devlin, J., et al.: Bert: pre-training of deep bidirectional transformers for language understanding (2018)
36. Duncan, J.: The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn. Sci. **14**(4), 172–179 (2010)