2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

Attention-based 3D Convolutional Network for Alzheimer's Disease Diagnosis and Biomarkers Exploration

*Dan Jin[1,2], Jian Xu[2,3], Kun Zhao[1,4], Fangzhou Hu[1,5], Zhengyi Yang[1,2], Bing Liu[1,2,6], Tianzi Jiang[1,2,6], Yong Liu[1,2,6]**

[1]Brainnetome Center & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [4]Shandong Normal University, Jinan, China, [5]Harbin University of Science and Technology, Harbin, China, [6]CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Modern advancements in deep learning provide a powerful framework for disease classification based on neuroimaging data. However, interpreting the classification decision of convolutional neural network remains a challenging task. It is crucial to track the attention of neural network and provide valuable information about which brain areas are particularly related to the diagnosis of disease. In this paper, we propose a novel attention-based 3D ResNet architecture to diagnose Alzheimer's disease (AD) and explore potential biological markers. Experiments are conducted on 532 subjects (227 of patients with AD and 305 of normal controls). By introducing the attention mechanism, the proposed approach further improves the classification performance and identifies important brain regions for AD classification simultaneously. The experiments also show that significant brain regions for AD diagnosis captured by our attention-based network are accompanied by significant changes in gray matter.

Index Terms— Attention mechanism, convolutional neural network, Alzheimer's disease, computer-aided diagnosis

1. INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia and leads to irreversible brain damage. The disease is accompanied by memory deficit, communication difficulties, disorientation and behavior changes with disease progression and becomes one of leading causes of death [1]. Till now, it is still a big challenge to establish robust markers for diagnosing and monitoring disease progression in the early stages of AD. In the past decades, machine learning techniques have been widely used in neuroimaging studies to accelerate automatic diagnosis and develop potential image markers. Neuroimaging data usually contain millions of voxel-wise features. Classification based on traditional machine learning methods, such as support vector machine, linear discriminant analysis or random forest, requires a complex procedure for handcrafted feature extraction and dimension reduction either using data-driven approaches, such as principal component analysis and independent component analysis, or relying on prior knowledge like brain atlas [2]. Deep learning algorithms provide great potential to overcome this problem, which automatically learn features from high-dimensional neuroimaging data and achieve more effective individualized diagnosis. There are several studies using deep learning methods for AD, mild cognitive impairment (MCI) and normal controls (NCs) classification [3-6]. Although previous studies based on deep learning model achieve great classification performance for AD diagnosis, it lacks interpretability about what makes deep learning model arrive at the conclusions and which brain regions are particularly associated with the diagnosis of disease. Meanwhile, it is crucial to provide information of regional importance to medical experts for clinical diagnosis and exploring the pathogenesis of disease.

Recently, several studies seek to explore the interpretability of the network in medical image analysis. For example, Korolev et al. generated network attention by measuring the drop of the output probability using images obstructed using a 7×7×7 box [7]. Yang et al. made extensive and detailed analysis of three different approaches for explaining 3D convolutional neural network (CNN) for AD classification [8], which are sensitivity analysis by 3D ultra-metric contour map, 3D class activation mapping (CAM) [9], and 3D gradient-weighted class activation mapping (Grad-CAM) [10]. However, classification performances based on the 3D-CAM and 3D-Grad-CAM methods drop substantially. To better explain the behavior of network and generate more discriminative feature representations, attention mechanism

* Research supported by the Natural Science Foundation of China (Nos. 81571062, 81871438, 81471120). Email: yliu@nlpr.ia.ac.cn

gradually becomes popular and attention-based networks are widely employed in natural language processing [11], image recognition [12, 13]and image synthesis [14].

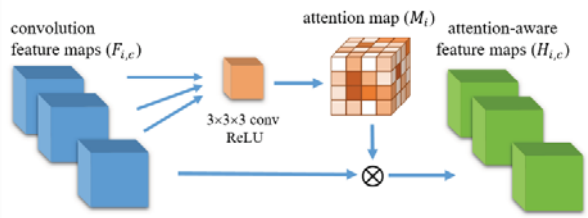

Fig. 1. Schema of attention mechanism.

Inspired by the attention mechanism, we proposed a 3D attention-based residual neural network (ResNet) for AD classification and potential biomarker exploration based on Magnetic resonance imaging (MRI) images. Without complex feature extraction and feature selection processes, our straightforward end-to-end attention-based network achieved remarkable classification performance. Based on attention mechanism (Fig.1), we further explored the importance of various brain regions for pathogenesis analysis of AD to give medical experts a better understanding of how neural network models make decisions and assist to discover potential biomarkers.

The main contributions of this paper can be summarized as follows:

(1) We proposed a simple end-to-end 3D attention-based deep learning network and achieved remarkable classification performance without handcrafted feature generation and model stacking.

(2) We introduced attention mechanism to identify important brain regions that are particularly associated with the diagnosis of disease and assist diagnosticians to explore potential biomarkers.

(3) The network combined with attention mechanism achieves better classification performance and almost does not increase the computation cost. Furthermore, the attention module is independent and able to incorporate with other deep network structures.

2. METHOD

2.1. Attention-based 3D ResNet architecture

We proposed a simple and effective attention-based 3D residual network for AD classification and important regions identification. The full architecture of attention-based 3D ResNet is depicted in Fig. 2. The ResNet [15] architecture improves image classification performance by increasing the depth of network and alleviates the problem of relatively small training dataset. Specially, we used the ResNet-18, which consists of a convolutional layer, eight basic ResNet blocks and a fully connected layer. Each basic block consists of two convolutional layers and each convolutional layer is followed by batch normalization and a nonlinearity activation function ReLU [16]. In the proposed method, we employed average-pooling function which is more suitable than max-pooling for disease classification, because average-pooling operation can reflect the information of gray matter volume of brain regions. In the output layer, we use the softmax classifier based on cross-entropy loss. The attention module is embedded into the ResNet architecture and carried out simply by a convolution layer with a set of filters of 3 × 3 × 3 kernel size (Fig. 1).

The attention module can capture significance of various voxels for classification during end-to-end training, which is instructive to explore potential imaging markers. During the forward process, the attention module serves as a feature selector. Each voxel of H×W×D-dimensional feature maps $F_{i,c}$ is weighted by the H×W×D-dimensional attention mask M_i (Fig. 1). The trainable attention mask M_i, which is independent of the channel of features and only related to spatial position, indicates the significance of each voxel i. The weighted features $H_{i,c}$ are defined as follows:

$$H_{i,c} = M_i * F_{i,c} \tag{1}$$

where, the spatial position (x, y, z) of the voxel is defined as i ($i \in \{1, \dots, H\times W\times D\}$, $x \in \{1, \dots, H\}$, $y \in \{1, \dots, W\}$, $z \in \{1, \dots, D\}$) and $c \in \{1, \dots, C\}$ is the index of the channel. The attention module can also work as a gradient update filter during the back propagation. Therefore, attention layer makes network more robust and improves the classification performance.

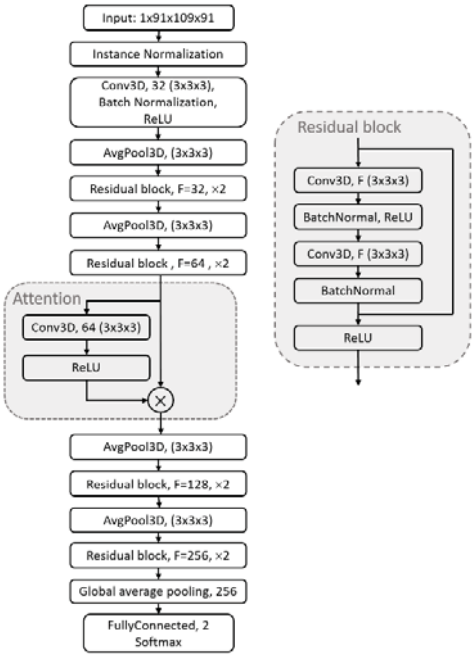

Fig. 2. Left: The architecture of attention-based 3D ResNet; Right: Basic block of residual network. F is the number of channels.

After end-to-end training, the potential biomarkers that are important for classification are enhanced by attention

mask M_i automatically. Based on the weakly-supervised classification labels (without voxel-wise significant labels), the attention-based 3D residual network not only achieves remarkable classification performance but also provides the significance of potential biomarkers that might assist the diagnosis of disease. It is worth noting that this end-to-end network has no need of prior knowledge to design handcrafted features. It leads to two benefits of this network: it may assist diagnosticians to discover potential biomarkers, and it can be easily transferred to classification of other brain diseases.

2.2. Data and preprocessing

For the experiments, we used T1 structural MRI dataset obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) to examine the performance of the proposed method. To prevent possible information leaks, we only used one MRI image for each subject with the mini-mental state examination (MMSE). The dataset includes 532 subjects: 227 patients with AD (105 females, age: 74.77±7.60, MMSE: 22.48±3.12) and 305 NC (149 females, age: 74.58±5.65, MMSE: 28.93±1.38).

To investigate valuable information about regional changes in gray matter for the training model, structural MRI images were preprocessed with the standard steps in the Cat12 toolbox (http://dbm.neuro.uni-jena.de/cat/). The images were bias-corrected, segmented into gray matter (GM), white matter (WM), and registered to the Montreal Neurological Institute (MNI) space using sequential linear (affine) and non-linear transformations (warping). The gray matter images were resliced to 2 mm cubic size resulting in a volume size of 91 × 109 × 91.

2.3. Experiment setup

To get better estimation of classification performance, we conducted stratified 10-fold cross-validation, where subjects were randomly partitioned into 10 subsets with stratification based on the subject's label. We repeated the experiments 10 times for AD/NC classification, by using 9 out of 10 subsets for training and the remaining one for testing in each cross-validation round. The accuracy, sensitivity, specificity and area under the curve (AUC) of receiver operating characteristic are used to evaluate the performance of the proposed model. In view of the limitation of GPU memory, we trained the classification models with He's initialization [17] using the optimizer Adam with initial learning rate of 10^{-5} and batch size of 8.

3. RESULTS AND DISCUSSION

3.1. Comparison of different pooling functions

We compared the effects of different pooling functions on classification performance. As shown in Table 1, when the averaging-pooling function is used, there is a substantial increase in the performance of classification. One possible reason is that max-pooling leads to the loss of volume information of gray matter. The changes of volume of gray matter in AD has been confirmed by previous quantitative volumetric MRI studies [18, 19], which is very important for disease classification. Therefore, we suggested that the averaging-pooling function in 3D deep convolutional networks for disease classification based on neuroimaging data.

Table 1. Classification performance (mean ± standard deviation) with various pooling functions.

	Accuracy	AUC	Sensitivity	Specificity
Max-pooling	0.844 (0.036)	0.868 (0.050)	0.802 (0.075)	0.875 (0.021)
Average -pooling	0.921 (0.033)	0.941 (0.035)	0.890 (0.053)	0.944 (0.051)

3.2. Comparison of methods without /with attention

To further evaluate the effectiveness of the embedded attention model, we evaluated the performance for the traditional basic 3D ResNet. Unlike previous interpretability analysis methods that lead to substantial drop in classification performance, the embedded attention module did not attenuate the classification performance of the network, and even caused a slight increase in classification performance (Table 2). The experiments demonstrated that the network combined with attention mechanism further improved the discriminative ability of network, which benefited from the important role in feature selection and gradient update filter of the attention module.

Table 2. Classification performance (mean ± standard deviation) of 3D ResNet architecture without and with attention mechanism.

Method	Accuracy	AUC	Sensitivity	Specificity
3D-ResNet	0.906 (0.031)	0.933 (0.036)	0.894 (0.064)	0.915 (0.041)
Proposed	0.921 (0.033)	0.941 (0.035)	0.890 (0.053)	0.944 (0.051)

3.4. Network's attention for important region exploration

Different with previous medical image analysis methods [7, 8], the proposed method can generate the attention map by a forward propagation in classification directly. The attention map indicates the significance of various brain regions for AD classification. The significance of regions is re-evaluated by time-consuming image occlusion strategy in [7, 8]. In the 3D-CAM and 3D-Grad-CAM methods [8], they also require additional calculations and lead to substantial drop in classification performance. By introducing the attention mechanism module, we can obtain 3D attention map for each testing sample from the trained models. The size of 3D attention map is 23 × 28 × 23. We up-sample the 3D attention maps to the size of the original images for further comparison. The mean 3D attention map of all normal

controls and AD patients from ten models trained on various fold was achieved. This 3D attention map, which indicates the significance of various brain regions related to changes in gray matter for AD classification, is normalized to a range of 0–1 for visualization.

The results showed that attention-based network highlights the brain regions mainly located in the temporal lobe, hippocampus, parahippocampal gyrus, cingulate gyrus, thalamus, precuneus, insula, amygdala, fusiform gyrus and medial frontal cortex (Fig. 3, left). The medial temporal lobes, including the entorhinal cortex, the hippocampus, the parahippocampus and the amygdala, are the earliest identified regions of histopathological changes with the hallmarks of neurofibrillary tangles and amyloid depositions in AD [20]. These regions have been proved to play an important role in encoding and retrieval of episodic and spatial memory, which are predominate deficit domains of clinical manifestation of AD [21]. In addition, regions identified by the proposed model include several important nodes of the default mode network, such as posterior cingulate cortex, precuneus, lateral temporal cortex and medial prefrontal cortex. Evidence from the fMRI studies have demonstrated that AD is associated with the disruptions of the default mode network compared to NC [22]. And these regions are also important regions of episodic memory network. It should be noted that the earliest cognitive deficits are noted in episodic memory, then all manner of cognition, motor function and personality are eventually affected in AD patients. Hence, with an increasing understanding of novel target regions, the early detection of AD is of growing importance for finding solutions to slow down the disease course.

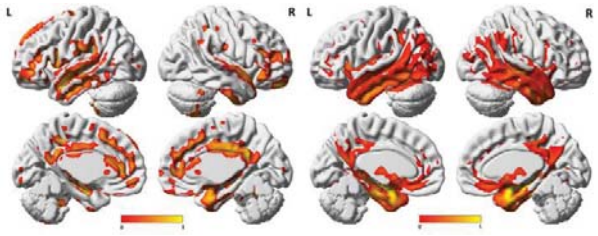

Fig. 3. Left: The mean 3D attention map. The brighter color indicates that the region is more significant for classification. Regions with normalized attention weights smaller than 0.4 are not displayed. Right: The significant group difference of gray matter between AD and NC group based on VBM analysis. The brighter color indicates that the region is more significant with changes of gray matter.

In addition, we performed voxel-based morphometric (VBM) analysis for comparison with the important regions identified by the proposed method. The group comparisons are assessed by controlling the family wise error at a threshold of $P < 0.05$ for multiple comparisons and the T statistics are also normalized to a range of $0 - 1$ for visualization. As shown in Fig. 3, the VBM analysis reveals the significant gray matter changes in the hippocampus, the parahippocampal gyrus, the medial temporal lobe and the amygdala, which is largely overlapped with those regions identified in the proposed model. In addition, our method also finds several regions particularly related to the diagnosis of disease that are not very significant in the VBM analysis, such as the medial frontal lobe and the cingulate gyrus, which is consistent with previous meta-analysis of gray matter abnormality in AD [18, 19]. It is worth noting that although the medial frontal lobe and the cingulate gyrus are not very significant in the VBM analysis, it may indicates that other changes in these regions are occurring, such as texture changes [23]. The experimental results proved the validity of our method and these significant regions may have great potential to be novel imaging biomarkers for the computer-assisted diagnosis or characterization of AD.

3.5. Comparisons with related studies

In the classification of AD and NC, Table 3 presents recent related studies that conduct classification tasks using deep learning methods based on MRI image data. The table shows the mean accuracies with cross-validation except [3]. It should be noted that we only considered the recent studies that use a single MRI scan for each subject to eliminate possible information "leaks". Table 3 shows that the performance of the proposed method is comparable to, if not better than, those previous studies [3, 24]. It indicated that the proposed attention-based 3D residual network is effective for AD classification. Beyond classification, we were able to provide valuable information about the importance of brain areas relevant to the disease.

Table 3. Performance comparison of the proposed method and reported studies on AD classification.

Methods	Sample sizes	Accuracy
Liu et al. [25]	65 AD, 77 NC	0.878
Aderghal et al. [3]	188 AD, 228 NC	0.914
Korolev et al. [7]	50 AD, 61 NC	0.800
Suk et al. [6]	186 AD, 226 NC	0.903
Li et al. [24]	288 AD, 272 NC	0.911
Liu et al. [26]	93 AD, 100 NC	0.85
The proposed	227 AD, 305 NC	0.921

4. CONCLUSION

In this paper, we proposed a simple and effective attention-based 3D residual network for AD diagnosis. Without complicated feature extraction and feature selection, our straightforward end-to-end network achieved remarkable classification performance. The major advantage of the present work is that the incorporated attention mechanism not only improved the classification performance but also captured the significant brain regions for AD classification. It should also be noted that our attention-based network can be easily transferred to classification of other brain diseases where MR imaging is available.

5. REFERENCES

[1] A. s. Association, "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 14, no. 3, pp. 367-429, 2018.

[2] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage,* vol. 155, pp. 530-548, 2017.

[3] K. Aderghal, J. Benois-Pineau, and K. Afdel, "Classification of sMRI for Alzheimer's disease Diagnosis with CNN: Single Siamese Networks with 2D+? Approach and Fusion on ADNI," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017, pp. 494-498: ACM.

[4] A. Gupta, M. Ayhan, and A. Maida, "Natural image bases to represent neuroimaging data," in *International conference on machine learning*, 2013, pp. 987-994.

[5] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.

[6] H.-I. Suk, S.-W. Lee, D. Shen, and A. s. D. N. Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Medical image analysis,* vol. 37, pp. 101-113, 2017.

[7] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, 2017, pp. 835-838: IEEE.

[8] C. Yang, A. Rangarajan, and S. Ranka, "Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification," *arXiv preprint arXiv:1803.02544,* 2018.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921-2929.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *ICCV*, 2017, pp. 618-626.

[11] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[12] F. Wang *et al.*, "Residual Attention Network for Image Classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450-6458: IEEE.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," *arXiv preprint arXiv:1805.08318,* 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807-814.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026-1034.

[18] W.-Y. Wang *et al.*, "Voxel-based meta-analysis of grey matter changes in Alzheimer's disease," *Translational neurodegeneration,* vol. 4, no. 1, p. 6, 2015.

[19] J. Yang *et al.*, "Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation," *J Neurol Sci,* vol. 316, no. 1-2, pp. 21-9, May 15 2012.

[20] H. Braak and E. Braak, "Staging of Alzheimer's disease-related neurofibrillary changes," *Neurobiology of aging,* vol. 16, no. 3, pp. 271-278, 1995.

[21] G. C. Schwindt and S. E. Black, "Functional imaging studies of episodic memory in Alzheimer's disease: a quantitative meta-analysis," *Neuroimage,* vol. 45, no. 1, pp. 181-190, 2009.

[22] Y. Liu *et al.*, "Regional homogeneity, functional connectivity and imaging markers of Alzheimer's disease: a review of resting-state fMRI studies," *Neuropsychologia,* vol. 46, no. 6, pp. 1648-1656, 2008.

[23] A. Chaddad, C. Desrosiers, and M. Toews, "Local discriminative characterization of MRI for Alzheimer's disease," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, 2016, pp. 1-5: IEEE.

[24] X. Li, Y. Li, and X. Li, "Predicting Clinical Outcomes of Alzheimer's Disease from Complex Brain Networks," in *International Conference on Advanced Data Mining and Applications*, 2017, pp. 519-525: Springer.

[25] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, 2014, pp. 1015-1018: IEEE.

[26] M. Liu, D. Cheng, K. Wang, Y. Wang, and A. s. D. N. Initiative, "Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis," *Neuroinformatics,* pp. 1-14, 2018.