CA-ConvNeXt: Coordinate Attention on ConvNeXt for Early Alzheimer's Disease Classification

Weikang Jin, Yue Yin, Jing Bai, Haowei Zhen

CA-ConvNeXt:Coordinate Attention on ConvNeXt for Early Alzheimer's disease classification*

Weikang Jin[1], Yue Yin[2], Jing Bai[1(✉)][0000−0001−5412−7793], and Haowei Zhen[1]

[1] Xidian University, Xi'an 710071, Xi'an 710071, China
baijing@mail.xidian.edu.cn(J. Bai), weikang-jin@stu.xidian.cn(W. Jin), 20061212353@stu.xidian.cn(H. Zhen)
[2] The First Affiliated Hospital of Air Force Medical University, Xi'an 710071, China
yiny110@sina.com

Abstract. Early diagnosis of Alzheimer's disease allows patients to receive early and effective treatment as a way to increase their chances of survival. We propose CA-ConvNeXt for Early Alzheimer's disease classification to solve the common MCI, AD, and NC classification problems. We employ the latest ConvNeXt network, which has a simpler topology and greater performance than ResNet and Swin Transformer. We effectively increase the model performance and reach 96% accuracy on the public ADNI dataset by adding Coordinate Attention to the ConvNeXt network.

Keywords: Early Alzheimer's disease · Coordinate Attention · ConvNeXt.

1 Introduction

Alzheimer's disease(AD), a type of dementia, is probably the most common neurological illness. Normal cognition(NC), moderate cognitive impairment(MCI), and Alzheimer's disease(AD) are the three basic stages. It is evident that AD is a progressive neurodegenerative disease [6], with results demonstrating that the transition cycle can last up to 20 years or longer. As a result, many patients are unable to detect the early indications of AD, and by the time the symptoms arise, the situation has already worsened. Furthermore, the effects of AD are permanent, and current medical treatments can only slow the progression of symptoms rather than cure the disease. AD kills brain cells over time, resulting in cognitive damage which including memory loss, inability to

* This work was supported in part by the Key Research and Development Program of Shaanxi under Grant 2022GY-062 and 2020GXLH-Y-023, in part by the National Natural Science Foundation of China under Grant 61772401, and in part by the Science and Technology on Communication Information Security Control Laboratory.(Weikang Jin and Yue Yin contribute equally to this work.)(Corresponding author:Jing Bai.)

make decisions, and trouble communicating, all of which can get a massive effect on a patient's daily life. This could have a massive influence on the patient's daily life.The related medical costs are unaffordable for most families, and the treatment procedure might take a long time. Mental illness affects more than 50 million individuals worldwide, according to Alzheimer's Disease International (ADI). This number is forecast to rise to 152 million people [1]. However, with appropriate treatment, many people can survive this disease.

Over the years, researchers have worked to develop better computer-aided systems as a way to help doctors make better early diagnoses of patients. Early prediction of AD is the task of classifying different stages of neurodegeneration, mainly NC, MCI and AD. With the rise of machine learning, there has been a lot of research on early diagnosis of AD based on machine learning; Liu et al [10] proposed a multi-template feature representation AD diagnosis method based on multi-view learning and support vector machines, and Lizarraga et al [12] provided a web platform for AD diagnosis using SVM. Earlier traditional methods required specific preprocessing steps to extract image features by manual features, which were not only time consuming but also relied on the experience and repeated attempts of the technique. Therefore, deep learning methods, which have emerged in recent years, have become a good means to extract features automatically and efficiently.

For the early diagnosis of AD, the most commonly used deep learning method over the years is convolutional neural network (CNN), Habes et al. [9] proposed the use of CNN to classify AD and NC. M. Kavitha et al [8] investigated a modified U-net-like architecture to AD, NC and MCI with remarkable results. M. Nguyenet et al. proposed an RNN network based method to diagnose AD [14]. M. Hon and N. M. Khan applied transfer learning to the diagnosis of AD [3].

Over the previous two years, with the successful application of transformer in the field of vision, a wave of transformer work has erupted. Its powerful performance once eclipsed CNN as the standard framework, which was even considered to replace CNN. Transformer was first applied to visual images with good results by Dosovitskiy et al [2]. While Vit requires large dataset pre-training to have better results and requires high computing power. Touvron et al. Furthermore, several researchers have attempted to apply transformer to medical images, but due to a shortage of medical data sets, the results are not as good as expected. Sarraf et al [15] first successfully applied transformer to the early diagnosis of AD and achieved good results, but this was due to their large amount of slice data. Matsoukas et al [13] combined self-supervised and Deit together and found that in the case of self-supervised pre-training and large dataset, the transformer outperforms CNN.

However, due to complex strcture,transformer can present various problems in practical applications. Liu et al. [11] did a lot of experimental study to figure out why it outperforms CNN, and they came up with a pure CNN-based ConvNeXt network that not only outperforms the Swin Transformer but also keeps CNN's simplicity and efficiency. As a result, we are attempting to apply ConvNeXt to the early diagnosis of AD.

The attention mechanism can emphasize the part of interest in the network by autonomously learning a set of parameters. Hu et al. [5] proposed a channel attention SE-Net that adaptively adjusts the feature response between channels by feature rescaling. Woo et al. [16] proposed CBAM, which solves the problem that SE-Net does not incorporate spatial attention. Later, Hou et al. [4] proposed a more efficient Coordinate Attention(CA) that can better utilize the position information on channel.

We propose a CA-ConvNeXt with the following main contributions:

- The Convenext, which is built on CNN and has outperformed Swin Transformer, is used in this paper as a new benchmark in CNN structure. For the first time, the ConvNeXt network was employed on the ADNI data set, and its performance was excellent, according to the experiments' results.
- Adding CA mechanism to the ConvNeXt network makes full use of the position information in the channel direction and effectively improves the network performance.

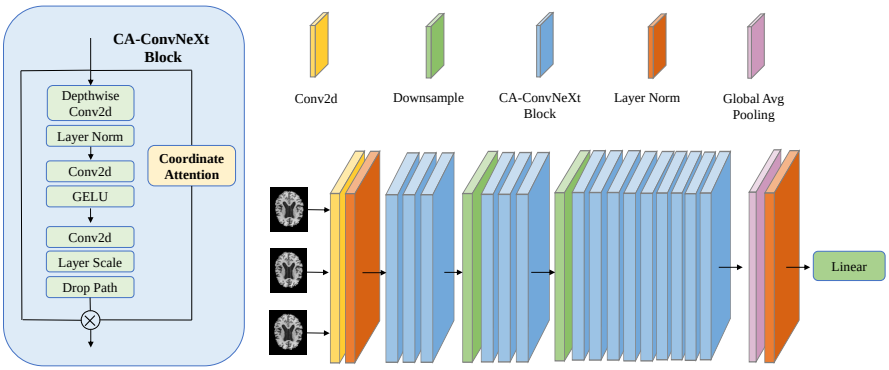

Fig. 1: Main framework of the proposed method.

2 Method

We adopt ConvNeXt as the main feature extraction framework and add CA as a module to the ConvNeXt block as a new line, and do a weighting with the normal ConvNeXt block as the output of the current block. Figure 1 shows our overall framework.

2.1 ConvNeXt

ConvNeXt's network architecture contains no structural or methodological innovations. It simply outperforms Swin Transformer in terms of performance and code complexity. This is mainly due to the effect of using Swin Transformer's strategy to train convolutional neural networks.

First, ConvNeXt takes ResNet50 as the benchmark model, Changing stage compute ratio, in the general ResNet50 network, the number of blocks stacked in four stages is (3,4,6,3), and the ratio is 1:1:2:1, but in Swin Transformer stage3 stacks a higher ratio of blocks. Therefore, ConvNeXt tries to modify the stacking number of ResNet50 from (3,4,6,3) to (3,3,9,3), and after this adjustment, the overall computation is about the same as Swin-T, and the accuracy rate is slightly improved. The original ResNet50 downsampling module is composed of a 7×7 convolutional layer with steps of 2 and a maximum pooling downsampling with a step of 2, and a width and height downsampling of four times. Then it was modified to Swin Transformer with a convolutional kernel of 4×4 with a step size of 4 to form a patchify, again with a width and height downsampling of four times, and no effect, but with a small improvement in accuracy.

Besides,ConvNeXt learns the ResNeXt group conv to increase performance. By grouping the channels of the input conv, group conv decreases computation. ConvNeXt is using each channel as a group, which becomes depthwise conv (dw conv). ConvNeXt adopts dw conv because it is quite comparable to Swin Transformer's local attention. To maintain flop consistency, the final ConvNeXt replaces the 3×3 conv in ResNet50 with a 3×3 dw conv and increases the base width of ResNet50 from 64 to 96. To emulate the transformer block's MLP module, ConvNeXt uses the inverted bottleneck from MobileNetV2, as the inverted bottleneck is very similar to the transformer block's MLP module, which also gives a small performance boost to ConvNeXt.

CNN networks after VGG use small convolutional kernels 3×3, while Swin-T uses a window size of 7×7, so to be consistent with Swin-T, ConvNeXt uses a large convolutional kernel of 7×7, which will increase Flops. to balance the computation, before this, ConvNeXt also moves the dw conv to the before this, to balance the computation, ConvNeXt also moves the dw conv to the top of the inverted bottleneck block. After this operation, the flops are about the same as before and the performance of the model unchanged.

Finally, some details were modified according to the Swin Transformer, replacing ReLU with GELU, using fewer activation functions and normalization layers, replacing the BN layer with an LN layer and adding a separate downsampling layer. Thus, based on these improvements, the final ConvNeXt is generated.

2.2 Coordinate Attention

Commonly used channel attention cannot save position information, and can only globally encode spatial information as channel information, and CA emerges to solve this problem. CA [4] is mainly divided into two parts: coordinate information embedding and CA generation. Figure 2 shows its structure.

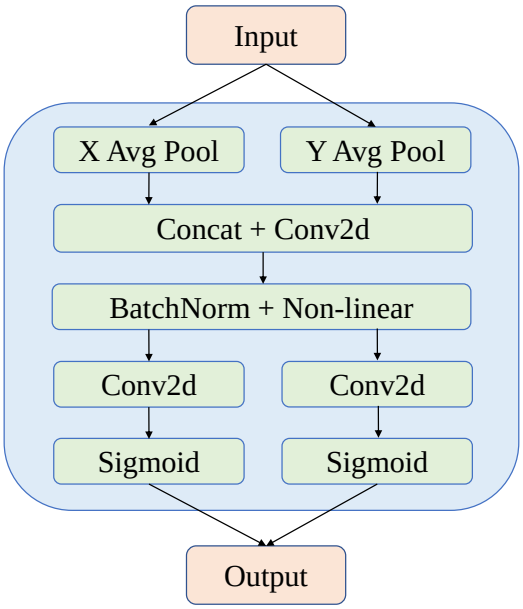

Fig. 2: Coordinate Attention

To obtain the important position information of the **x** and **y** axes, the global pooling layer is first decomposed into two one-dimensional feature codes, one for the horizontal direction and one for the vertical direction, as detailed in the following equation:

$$m_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{1}$$

$$m_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{2}$$

This allows CA to capture accurate position information in a channel direction, and in order to have a more accurate position information representation for the above operation, the CA generation is designed.

$$\mathbf{K}^h = \sigma\left(F_h\left(\mathbf{f}^h\right)\right) \tag{3}$$

$$\mathbf{K}^w = \sigma\left(F_w\left(\mathbf{f}^w\right)\right) \tag{4}$$

The two feature maps generated in the previous step are F1 transformed with a 1×1 shared convolution to generate intermediate feature maps in the

horizontal and vertical directions, and the module size is then controlled by the downsampling ratio. The intermediate feature map is split into two tensors and transform to the same number of channels as the input $\mathbf{x}$ with two 1×1 convolutions, respectively, to obtain $\mathbf{K}^h$ and $\mathbf{K}^w$, which are finally expanded to obtain the final CA module output $\mathbf{n}$.

$$n_c(i,j) = x_c(i,j) \times K_c^h(i) \times K_c^w(j) \tag{5}$$

3 Experiments and results

3.1 Data pretreatment

For the dataset we used the AD Neuroimaging Initiative (ADNI) database [7]. The downloaded 2032 samples were in NII format, which cannot be directly input into a two-dimensional network and have an obscure structure, so we preprocessed them. First, they were AC-PC corrected using the icbm152 template, and then linearly aligned and cranially separated using FSL. After preprocessing, the pathological structure of the brain map was more clearly defined. However, since it is still three-dimensional, we obtained coronal slices by fixing one of the axes. One NII subject can be sliced into 181 pieces, and we selected the middlemost one as a representative. This gave us a total of 2032 PNG images, including 1100 MCI, 321 AD and 611 NC.

3.2 Experimental details

Our experiments were conducted on a dell workstation configured with 64G of RAM, 3090 with 24g of video memory, 24cores Intel(R) Xeon(R) Gold 6248R CPU, python version 3.8, and cuda 11.0, Ubuntu 18.04.

We divided ADNI datasets into training and validation set with 4:1 ratio. Baseline control experiments use ResNet50 and Swin Transformer as a way to verify the advantages of the ConvNeXt network on the ADNI dataset. In the ablation experiments, we choose to add CA to the framework of ConvNeXt and not to add CA, respectively, to derive the performance improvement of CA on the network.

In terms of experimental parameters, experiments use the imagenet pretraining weights officially released by the respective models. Batchsize is set to 32, optimizer is selected AdamW, learning rate is $5e^{-4}$, decay weight is $5e^{-2}$, loss function is selected cross-entropy loss function. Swin Transformer was fine-tuned for 90 epochs and the rest of the experiments were fine-tuned for 50 epochs, each experiment was done 5 times and averaged, and each experiment is done 5 times to take the average. We choose accuracy as the main evaluation index.

3.3 Experimental result

According to the Table 1, the ConvNeXt network outperforms resnet50 and Swin Transformer on the ADNI dataset with only 2032 images, and its performance is good, 1.2% times better than resnet50 and 2.0 times better than Swin

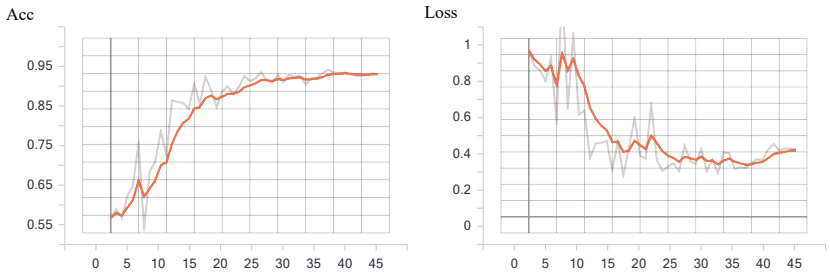

Fig. 3: Accuracy and loss in validation set of CA-ConvNeXt

Table 1: Accuracy on ADNI dataset

Dataset	Mothed	Accuracy
ADNI	ResNet50	93.8%
ADNI	Swin Transformer	94.3%
ADNI	ConvNeXt	95.3%
ADNI	CA-ConvNeXt	96.0%

Transformer. Furthermore, adding CA improves the performance of ConvNeXt by 0.7%, and the accuracy reaches 96% . We can see from the Figure 3 that the CA-ConvNeXt our proposed , final losses have all converged, and the accuracy can no longer rise. Its demonstrating the effectiveness of coordinating attention on ConvNeXt.

4 Conclusions

In this paper, we proposed CA-ConvNeXt network is experimentally proven to be effective in classifying the three categories of AD, MCI, and NC on the publicly available ADNI dataset. This is also the first application of ConvNeXt network on AD early diagnosis classification. Based on the experiments we can conclude that ConvNeXt is able to have better performance than Resnet50 and Swin Transformer on ADNI, and the experimental results meet the expectation, and the performance is further improved with the addition of CA. As can be observed, our proposed network has a high level of performance.

References

1. Association, A., et al.: 2018 alzheimer's disease facts and figures. Alzheimer's & Dementia **14**(3), 367–429 (2018)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is

worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Hon, M., Khan, N.M.: Towards alzheimer's disease classification through transfer learning. In: 2017 IEEE International conference on bioinformatics and biomedicine (BIBM). pp. 1166–1169. IEEE (2017)
4. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722 (2021)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
6. Islam, J., Zhang, Y.: Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. Brain informatics **5**(2), 1–14 (2018)
7. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
8. Kavitha, M., Yudistira, N., Kurita, T.: Multi instance learning via deep cnn for multi-class recognition of alzheimer's disease. In: 2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA). pp. 89–94. IEEE (2019)
9. Li, H., Habes, M., Wolk, D.A., Fan, Y., Initiative, A.D.N., et al.: A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's & Dementia **15**(8), 1059–1070 (2019)
10. Liu, F., Shen, C.: Learning deep convolutional features for mri based alzheimer's disease classification. arXiv preprint arXiv:1404.3366 (2014)
11. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
12. Lizarraga, G., Cabrerizo, M., Duara, R., Rojas, N., Adjouadi, M., Loewenstein, D.: A web platform for data acquisition and analysis for alzheimer's disease. In: SoutheastCon 2016. pp. 1–5. IEEE (2016)
13. Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K.: Is it time to replace cnns with transformers for medical images? arXiv preprint arXiv:2108.09038 (2021)
14. Nguyen, M., Sun, N., Alexander, D.C., Feng, J., Yeo, B.T.: Modeling alzheimer's disease progression using deep recurrent neural networks. In: 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI). pp. 1–4. IEEE (2018)
15. Sarraf, S., Sarraf, A., DeSouza, D.D., Anderson, J.A., Kabia, M., ADNI, A.D.N.I., et al.: Ovitad: Optimized vision transformer to predict various stages of alzheimer's disease using resting-state fmri and structural mri data. bioRxiv (2021)
16. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)